TRITON 2021

TRanslation and Interpreting Technology ONline

Proceedings of the Conference

05-07 July, 2021

Editors: Ruslan Mitkov, Vilelmini Sosoni, Julie Christine Giguère,
Elena Murgolo and Elizabeth Deysel

ISBN 978-954-452-071-7
This document is downloadable from: `https://triton-conference.org/proceedings`

Preface

Technologies for translation and interpreting have benefited from promising recent advances including the employment of Natural Language Processing (NLP) and deep learning technology amongst other latest solutions. The current pandemic has prevented the in-person exchange of ideas and networking of researchers and practitioners working on translation technology, interpreting technology, machine translation and NLP in general, but virtual communication opportunities have enabled continued collaboration and provided alternative communication channels. While eagerly awaiting the return of normality, the international conference TRITON (TRanslation and Interpreting Technology ONline) is offering an opportunity to exchange ideas, learn from each other and interact virtually.

TRITON is an online event which features two days of conference presentations and one day of tutorials. A number of leading scholars and industry stakeholders act as keynote speakers; oral presentations and posters were accepted with every submission evaluated by 3 reviewers. The programme is complemented by TRITON tutorials on hot topics from the fields of translation technology, interpreting technology or NLP.

We are pleased to offer the e-proceedings of the conference which feature the regular and poster papers to be presented at the TRITON conference as well as a selection of the keynote speeches. This e-proceedings volume comes with ISBN and DOI numbers assigned to every contribution.

We would like to thank all colleagues who made this truly international event possible. In the first place, we would like to acknowledge the keynote speakers of the conference: Fabio Alves, Lynne Bowker, Sabine Braun, Gloria Corpas Pastor, Elena Davitti, Stephen Doherty, Florian Faes, Marcello Federico, Stephanie Labroue, Adam LaMontagne, William D. Lewis, André Martins, Konstantin Savenkov, Rico Sennrich, Elsa Sklavounou, Josef van Genabith and Rosanna Villani. Words of gratitude go to our invited tutorial speakers too: Frédéric Blain, Esther Bond, Félix do Carmo, Clara Ginovart Cid, David Orrego-Carmona, Bianca Prandi, Tharindu Ranasinghe Hettiarachchige and Moritz Schaeffer. We are grateful to all members of the Programme Committee and the additional reviewers for carefully examining all submissions and providing substantial feedback on all papers, helping the authors of accepted papers to improve and polish the final versions of their papers.

Last but not least, we would like to use this paragraph to acknowledge the members of the Organising Committee, who worked very hard during the last few months and whose dedication and efforts made the organisation of this event possible. We would like to mention (in alphabetical order) the following colleagues who carried out numerous organisational tasks and were eager to step in and support the organisation of the conference whenever needed: Lucía Bellés-Calvera, Maria Carmela Cariello, Rocío Caro Quintana, Ana Isabel Cespedosa Vázquez, Parthena Charalampidou, Marie Escribe, Darya Filippova, René Alberto García Taboada, Dinara Gimadi, Ali Hatami, Valentini Kalfadopoulou, Sotirios G. Keramidas, Lydia Körber, Maria Kunilovskaya, Ljubica Leone, Ana Isabel Martínez-Hernández, Nikolai Nikolov, Martha Maria Papadopoulou, Kateryna Poltorak, Nikola Spasovski and Marina Tonkopeeva. Finally, our big ‘thank you’ goes out to the team leader of the Organising Committee Rocío Caro Quintana who was available 24/7 to coordinate the tasks of the Organising Committee and to assist us in a competent and unreserved manner.

Conference Chairs:
Ruslan Mitkov, Vilelmini Sosoni, Julie Christine Giguère, Elena Murgolo and Elizabeth Deysel

7 July 2021
Wolverhampton, Athens, London, Milan, Cape Town

Conference Chairs:

Ruslan Mitkov (University of Wolverhampton), Vilelmini Sosoni (Ionian University), Julie Christine Giguère (Andovar), Elena Murgolo (Aglatech14) and Elizabeth Deysel (Parliament of South Africa) are the TRITON conference chairs.

Programme Committee:

Frédéric Blain (University of Wolverhampton)
Lindsay Bywood (University of Westminster)
Parthena Charalampidou (Aristotle University of Thessaloniki)
Kevin Cohen (University of Colorado)
Noa Cruz Díaz (CaixaBank)
Félix do Carmo (University of Surrey)
Anastasia Dudkina (Saint Petersburg University)
João Esteves-Ferreira (Asling, Switzerland)
Richard Evans (University of Wolverhampton)
Claudio Fantinuoli (University of Mainz and KUDO Inc.)
Yota Georgakopoulou (Athena Consultancy)
Clara Ginovart Cid (Pompeu Fabra University and Datawords)
Mina Ilieva (Mitra Translations)
Valentini Kalfadopoulou (Oracle)
Alina Karakanta (Fondazione Bruno Kessler)
Katia-Lida Kermanidis (Ionian University)
Payal Khullar (International Institute of Information Technology Hyderabad)
Maria Kunilovskaya (University of Wolverhampton)
Todor Lazarov (New Bulgarian University)
Lieve Macken (Ghent University)
Despoina Mouratidis (Ionian University)
Constantin Orasan (University of Surrey)
Jun Pan (Hong Kong Baptist University)
Bianca Prandi (University of Mainz)
Tharindu D. Ranasinghe Hettiarachchige (University of Wolverhampton)
Rozane Rebechi (Federal University of Rio Grande do Sul)
María Recort Ruiz (International Labour Organization, Switzerland)
Maria Stasimioti (Ionian University)
Olaf-Michael Stefanov (JIAMCATT, Austria)
Rossana Villani (European Central Bank)
Anna Władyka-Leittretter (AMWL-Sprachen)
Mai Zaki (American University of Sharjah)
Marcos Zampieri (Rochester Institute of Technology)
Anna Zaretskaya (TransPerfect)

Tutorial Leaders:

Esther Bond (Slator)
Frédéric Blain (University of Wolverhampton)
Félix do Carmo (University of Surrey)
Clara Ginovart Cid (Datawords)
David Orrego-Carmona (Aston University)
Bianca Prandi (University of Mainz)
Tharindu Ranasinghe Hettiarachchige (University of Wolverhampton)
Moritz Schaeffer (University of Mainz)

Keynote Speakers:

Fabio Alves (Federal University of Minas Gerais, Belo Horizonte)
Lynne Bowker (University of Ottawa)
Sabine Braun (University of Surrey)
Gloria Corpas Pastor (University of Malaga)
Elena Davitti (University of Surrey)
Stephen Doherty (University of New South Wales, Sydney)
Florian Faes (Slator)
Marcello Federico (Amazon)
Stephanie Labroue (Systran)
Adam LaMontagne (RWS)
William D. Lewis (University of Washington and Microsoft Translator)
André Martins (Unbabel)
Konstantin Savenkov (Intento)
Rico Sennrich (University of Zürich)
Elsa Sklavounou (RWS)
Josef van Genabith (German Research Centre for Artificial Intelligence)
Rosanna Villani (European Central Bank)

Additional Reviewers:

Maria Carmela Cariello (University of Pisa)
Rocío Caro Quintana (University of Wolverhampton)
Anna Beatriz Dimas Furtado (University of Malaga)
Viveta Gene (Ionian University)
Lilit Kharatian (University of Wolverhampton)
Shaifali Khulbe (University of Wolverhampton)
Aida Kostikova (New Bulgarian University)
Plamena Krasteva (Mitra Translations)
Ljubica Leone (Lancaster University)
Elpida Loupaki (Aristotle University of Thessaloniki)
Ana-Isabel Martínez-Hernández (Universitat Jaume I)
Paola Ruffo (Heriot-Watt University)
Nikola Spasovski (University of Wolverhampton)
Natalia Sugrobova (University of Gent)
Marina Tonkopeeva (University of Wolverhampton)

Organising Committee:

Lucía Bellés-Calvera
Maria Carmela Cariello
Rocío Caro Quintana
Ana Isabel Cespedosa Vázquez
Parthena Charalampidou
Marie Escribe
Darya Filippova
René Alberto García Taboada
Dinara Gimadi
Ali Hatami
Valentini Kalfadopoulou
Sotirios G. Keramidas
Lydia Körber
Maria Kunilovskaya
Ljubica Leone
Ana-Isabel Martínez-Hernández
Nikolai Nikolov
Martha Maria Papadopoulou
Kateryna Poltorak
Nikola Spasovski
Marina Tonkopeeva

Table of Contents

Keynote Speakers

Evaluation Metrics for Machine Translation

Cognitive Aspects of Interpreting

Named Entity Recognition and Resources for Machine Translation

Interpreting Technology/Text-to-Sign Translation/Translation Memories

MT for Low Resourced Languages/MT for Translating of Subtitles

NLP projects

Post-Editing Efforts/Automatic Post-Editing

Multiword Expressions in MT/MT Training

Post-Editing and Post-Editing Applications

Online Translation Courses/Localisation

What a future talent would say about translation automation?

Elsa Sklavounou [1]

[1] RWS Senior Director, International Partnerships
esklavounou@rws.com

1 Introduction

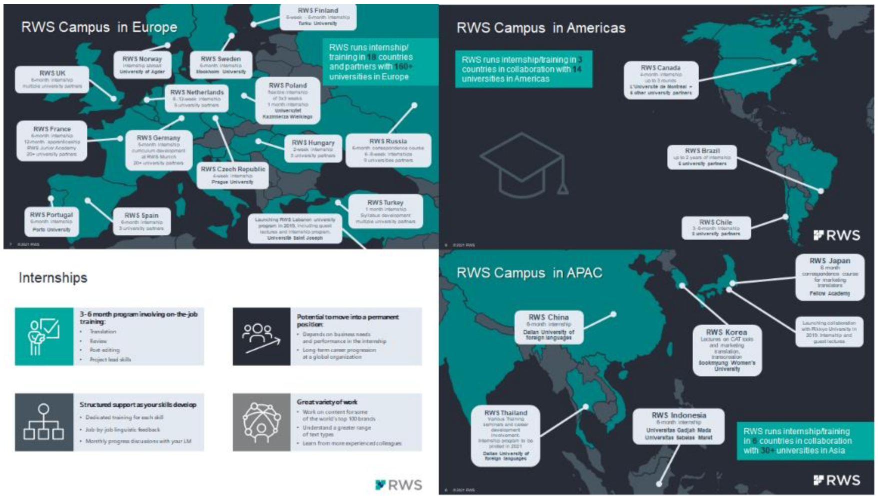
This keynote script aims to present future talent's vision on the role translation automation plays in building a continuous intelligent content supply chain under subject matter expert supervision highlighting the importance of the localization experts and professionals.

Artificial intelligence and automation in general are defining corporate digital content strategy. When the automation is enabled thanks to AI-based language technology within localization project workflows, RWS, through the RWS Campus, aims to inspire great futures in localization and to be recognized in our industry for developing localization talent and markets worldwide: more than 200 universities in 37 countries are figuring amongst the academic institutions RWS supports through

- University events (55 universities across 22 countries)
- Advisory programs (8 universities)
- Presentations and Workshops (38 universities in 17 countries)
- Internship Programs (80+ universities in 20+ countries)

Highly skilled interns are getting hired

- for permanent roles at RWS in 17 countries
- for freelance roles at RWS in 8 countries


RWS Campus in Europe
RWS Finland
6-week - 6-month internship
Turku University
RWS runs internship/ training in 18 countries and partners with 160+ universities in Europe
RWS Norway
Internship abroad
University of Agder
RWS Sweden
6-month internship
Stockholm University
RWS UK
6-month internship
multiple university partners
RWS Netherlands
8-12-week internship
6 university partners
RWS Poland
flexible internship of 3x3 weeks
1 month internship
Uniwersytet Kazimierza Wielkiego
RWS France
6-month internship
12-month apprenticeship
RWS Junior Academy
20+ university partners
RWS Germany
5-month internship
curriculum development at RWS Munich
20+ university partners
RWS Hungary
2-week internship
3 university partners
RWS Russia
6-month correspondence course
6-8-week internships
9 universities partners
RWS Czech Republic
4-week internship
Prague University
RWS Portugal
6-month internship
Porto University
RWS Spain
6-month internship
3 university partners
Launching RWS Lebanon university program in 2015, including guest lectures and internship program.
Université Saint Joseph
RWS Turkey
1 month internship
Syllabus development
multiple university partners
RWS Campus in Americas
RWS runs internship/training in 3 countries in collaboration with 14 universities in Americas
RWS Canada
4-month internship
up to 3 rounds
L'Université de Montréal + 6 other university partners
RWS Brazil
up to 2 years of internship
6 university partners
RWS Chile
3-6-month internship
2 university partners
RWS
RWS Campus in APAC
RWS Japan
6 month correspondence course for marketing translators
Fellow Academy
Launching collaboration with Rikkyo University in 2019. Internship and guest lectures
RWS China
6-month internship
Dalian University of foreign languages
RWS Korea
Lectures on CAT tools and marketing translation, transcreation
Sookmyung Women's University
RWS Thailand
Various training seminars and career development involvement. Internship program to be piloted in 2021
Dalian University of foreign languages
RWS Indonesia
6-month internship
Universitas Gadjah Mada
Universitas Sebelas Maret
RWS runs internship/training in 6 countries in collaboration with 30+ universities in Asia
RWS

Internships

3-6 month program involving on-the-job training:
- Translation
- Review
- Post-editing
- Project lead skills

Potential to move into a permanent position:
- Depends on business needs and performance in the internship
- Long-term career progression at a global organization

Structured support as your skills develop
- Dedicated training for each skill
- Job-by-job linguistic feedback
- Monthly progress discussions with your LM

Great variety of work
- Work on content for some of the world's top 100 brands
- Understand a greater range of text types
- Learn from more experienced colleagues

RWS

Fig. 1. RWS Campus Internships

Translation and Interpreting Technology Online, pages 1–6,
July 05-07, 2021.
https://doi.org/10.26615/978-954-452-071-7_001

2 Building an intelligent content supply chain

The priority for any business must be its reputation – without it, it is doomed to fail. Compliance with local laws and regulations is essential, and businesses should have a strategy in place for how they will ensure that their content assets continue to meet ever-changing rules around the world. Automatic translation has a vital role to play in helping to produce Multilingual Content – as does weaving a Corporate Content Fabric – where Artificial Intelligence, based on subject-matter expertise, helps businesses at every stage of their content development – from creation, through translation, to complex delivery.

Building an intelligent content supply chain under an optimized operating model from content creation to multi-channel content distribution requires companies to adopt the right Content Management Solution for their business-critical information so as to enable

- Full Digital Transformation
- Successful AI Implementation

A component-based content infrastructure provides optimal content reuse and life cycle management across any device and any format. By ensuring it integrates seamlessly with existing systems you can achieve unified collaboration for authoring and reviewing, and improved governance at scale in 250+ languages, fully secured and cost-optimized.

Content management in a multilanguage, multichannel world is fiendishly difficult and makes it hard to maximize the value of your data because of these complex content deliveries.

However, there is a way…

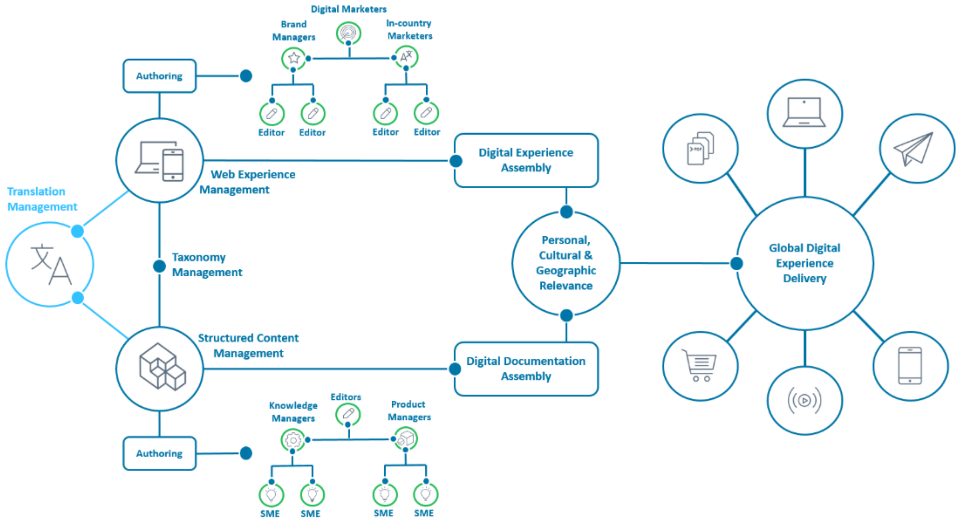


Fig. 2. Intelligent Content Supply Chain

Intelligent Content platforms allow companies to optimize and automate content processes at scale, so they can cost-efficiently import, create, manage, localize and distribute a wide spectrum of content types – from highly structured regulatory content to engaging marketing content manageable throughout the content life cycle.

3 Streamlining intelligence from multi-target content

Single-sourcing of content reduces cost of content development (30%) and localization (30%-50%). Shifting translation automation from Localization to Authoring and Management flows allows the immersion of locals in the taxonomy model based on the delivery target destinations not only from delivery channel and domain perspective but also it handles local language variants. The advantages in moving the machine translation to the authoring stage of a well-defined content supply chain are multiple:

- Delivering the right post-sale experience

Enhance customer experience by delivering the technical post-sale content experience that matches the personalized, dynamic and rich pre-sale content experience.

- Shorter turnaround times

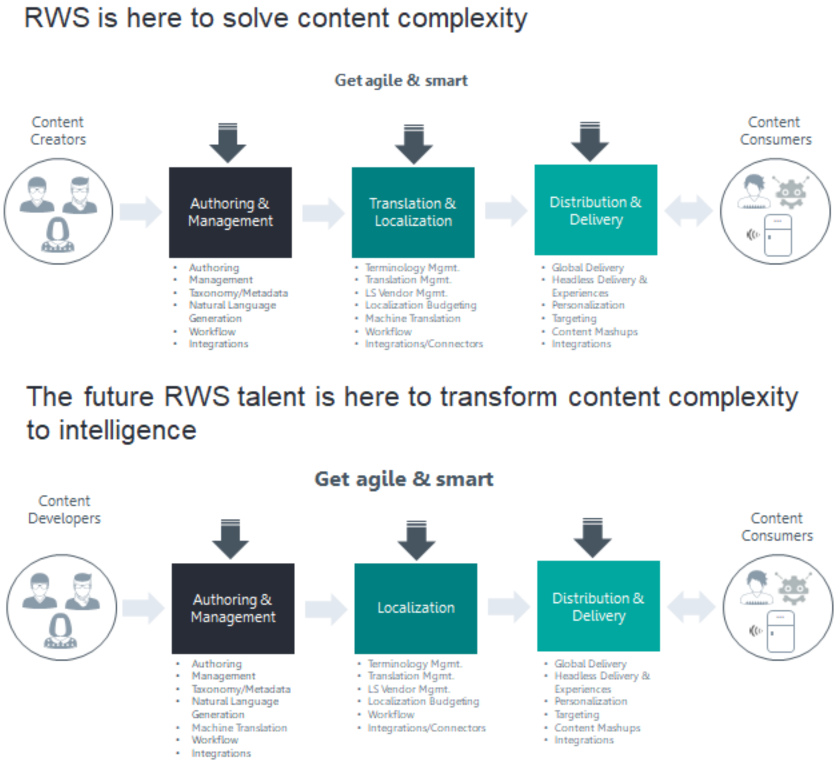
The advanced version management for the creation, management, review, edit and delivery steps are operated within a single workflow process and the streamlined collaboration contribute to eliminate manual, error-prone reviews and stop being the middleman when reviewers disagree. Empower SMEs to contribute feedback for cross-departmental sharing.

- Better fit of content to the form-factor of the delivery channel

Automated rendering into multiple formats based on style sheets reduces - or completely eliminates Desktop Publishing costs. Eliminate manual post-processing for translated content layout fixes as content is developed at database level and pulled based on the delivery channel depending on the digital experience.

- Ability to train MT based on the delivery channel

By enabling content reuse, breaking publications down into smaller, reusable pieces, managed on a single platform and available to all stakeholders accelerating the digital experience is reducing costs, allows to feed back as many localizations as digital experiences tagged for machine learning purposes, via the respective metadata by the subject matter experts and content developers within a circular workflow for long content life cycle.



RWS is here to solve content complexity

Get agile & smart

Content Creators

Authoring & Management

- Authoring
- Management
- Taxonomy/Metadata
- Natural Language Generation
- Workflow
- Integrations

Translation & Localization

- Terminology Mgmt.
- Translation Mgmt.
- LS Vendor Mgmt.
- Localization Budgeting
- Machine Translation
- Workflow
- Integrations/Connectors

Distribution & Delivery

- Global Delivery
- Headless Delivery & Experiences
- Personalization
- Targeting
- Content Mashups
- Integrations

Content Consumers

The future RWS talent is here to transform content complexity to intelligence

Get agile & smart

Content Developers

Authoring & Management

- Authoring
- Management
- Taxonomy/Metadata
- Natural Language Generation
- Machine Translation
- Workflow
- Integrations

Localization

- Terminology Mgmt.
- Translation Mgmt.
- LS Vendor Mgmt.
- Localization Budgeting
- Workflow
- Integrations/Connectors

Distribution & Delivery

- Global Delivery
- Headless Delivery & Experiences
- Personalization
- Targeting
- Content Mashups
- Integrations

Content Consumers

Fig. 3. Translation Automation shifting to Authoring and Management workflow

4 From Content Management Systems to Content Intelligence Management Systems

The content intelligence as seen in the previous chapter is derived during intelligent content development using metadata within localization processes supporting digital experience acceleration and multi-channel deliveries. Digital-first distribution being continuous, linear is no more good enough. Authoring, content management and localization can be agile and iterative. Commercial flexibility, such as monthly or quarterly utility billing, is paired with overlapping handoffs, smooth management of updates, and economical handling of bite-sized content. Besides new commercial models, this demands new translation technologies to unfragment traditional tech in favour of content lifecycle. The next-generation tech is agile. An agile content intelligence management system is designed from the ground up to support continuous localization, building an intelligent content supply chain with:

- High levels of automation and visibility
- Deep integration with new hybrid CAT and linguistic AI tools (including neural machine translation)
- Content connectors and APIs that make it easy to embed translation into business workflows to connect a content repository to a localization management solution. Connector functionality ranges from the simple (e.g. automated logins to cloud file-sharing services) to the complex (e.g. embedding translation modules into an on-premises content management system).

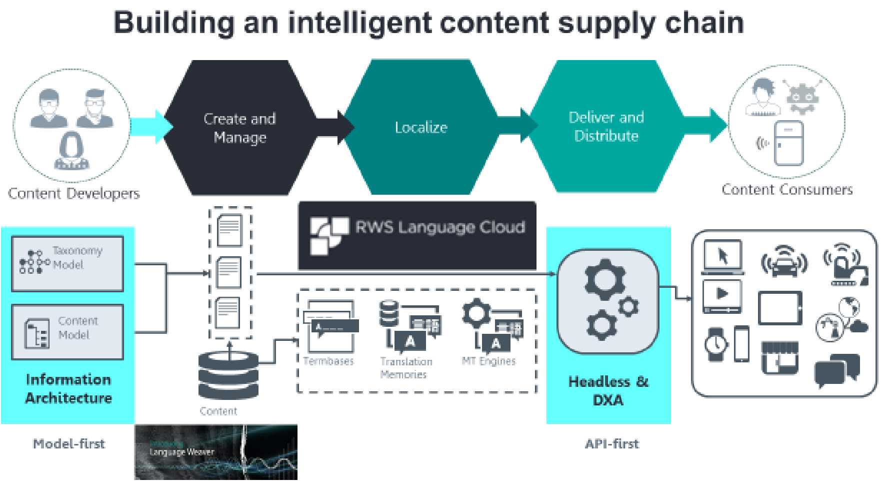


Fig. 4. Intelligent Content Supply chain

The key benefits are to go to market sooner, content connectors jump-start the content translation process so it is more agile, streamlined and efficient without the overhead of manual project standardized tasks less subject to error.

As projects get smaller and faster, technology must be used to eliminate overhead as much as possible by automating handoffs and repetitive tasks for each project. Continuous localization is the best model for managing large volumes of content and more iterations. To keep up with market demand for more personalized content, it's essential to invest in a robust technology framework built for continuous localization.

Continuous content creation requires dynamic, real-time localization, which is why continuous localization helps you achieve:

- Faster time to market: Translate content as needed.
- Higher quality translations: Minimize errors by reducing manual effort.
- Cost reduction: Eliminate repetitive project tasks.
- Process optimization: Analyze data for productivity gains.

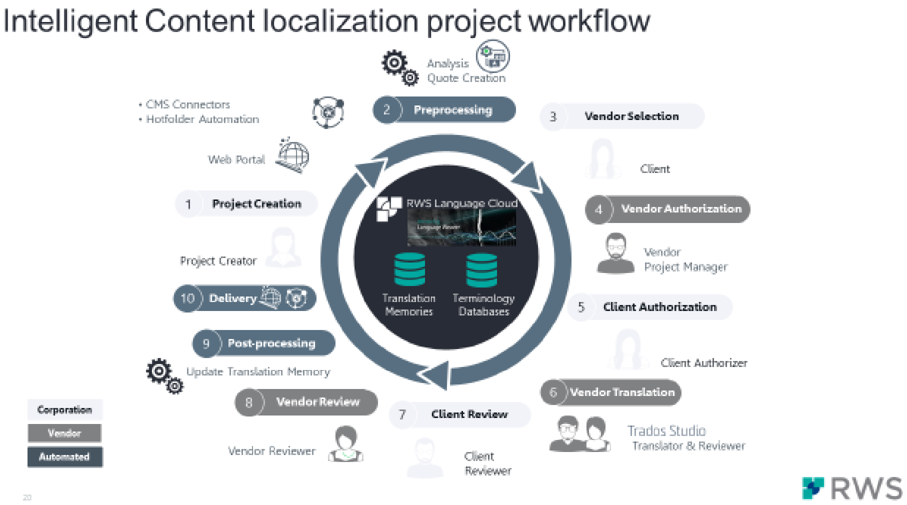


Fig. 5. Agile Intelligent Content localization management project workflow

Agile localization management is the use of specific technologies to manage translation as a continuous iterative process, rather than as a linear project. Content is translated as it is being developed.

An agile localization management system supports the following capabilities:

- Incremental flexibility
- Modern UX
- Easy file hand-off
- Side-by-side review

The key benefits include: Faster time to market, shorter turnaround time for greater automation and efficiency in end-to-end processes.

The solution constantly updates and evolves. Enabling staff around the world to detect and understand relevant information in other languages by sharing knowledge with one another. In a world without language barriers, this helps ensure local laws and regulations are being followed correctly worldwide.

Subject matter linguistic expertise is brought in to ensure the highest quality localizations, the longevity of the content through industry specific compliance standards, and helps Linguistic AI get smarter and smarter.

The changing profile of the translator profession at the European Central Bank

Rossana Villani[1]

[1] Head of the Language Services Division, European Central Bank
Rossana.Villani@ecb.europa.eu

Abstract. The translator profile at the European Central Bank (ECB) has evolved increasingly rapidly over the past years. Staff of the linguistic services moved from their traditional set of tasks - translation, editing and proofreading, often done on print-ups or in plain Word documents - to dealing with a multitude of tasks, such as project management, procurement, recruitment and training, and lately cultural mediation. But the real game changer in this transformation has been the quantum leap that technology applied to translation (computer-assisted translation (CAT) and machine translation (MT)) has enabled in complex organisational structures. This brought about significant challenges in terms of people and change management, but also presented huge opportunities and the chance to broadcast the ECB's language services as a cutting-edge professional unit within the organisation. The transformation is still underway, but we are not in uncharted waters anymore. In fact, we are broadening our interest in new technologies to make our communication more accessible, even beyond the pure translation function.

1 Introduction

The European Central Bank is one of the institutions of the European Union. Founded in 1998, it is the central bank of a group of currently 19 countries in the EU which have adopted the euro as their common currency, the equivalent of the Federal Reserve System in the United States or the Bank of England in the United Kingdom.
The ECB’s key tasks include keeping inflation under control, keeping the banking system sound and the issuance of euro banknotes and coins. The bank’s headquarters are in Frankfurt am Main, Germany’s financial capital.

The bank’s linguistic area is called the Language Services Division. Somewhat paradoxically, it is the largest division in the bank in terms of staff; but at the same time, it is a relatively small operation compared to the linguistic services of other EU institutions.

The Language Services Division is part of the Communications department, which reflects the fact that for an international organisation such as the ECB, multilingual communication forms an integral part of its communication. The Division was incorporated into the Communications department as recently as 2013, which is indicative of some of the trends that will be addressed.

Translation and Interpreting Technology Online, pages 7–14,
July 05-07, 2021.
https://doi.org/10.26615/978-954-452-071-7_002

2 Twenty years of change

So how has our translator's world transformed over the past 20 years?

Starting with target audiences, we have observed a strong shift of focus from expert to non-expert audiences.

This was also reflected in the broadening range of text types. New text types were also added with the emergence of new channels of communication, and the digital revolution in communications also broadened the range of file formats that translators are exposed to. On the resources front, the focus clearly shifted from investing in internal staff to outsourcing, which meant a higher coordination overhead for the function as a whole and more quality control for the translator. However, for the translator, the biggest change of all was to experience how technology slowly crept into and gradually took over their day-to-day work. In addition to being expected to use computer-assisted translation tools and (with the quantum leap in artificial intelligence) machine translation, they were suddenly exposed to a broad range of tools and technical environments to perform their translations (e.g. web editors, social media platforms), task management and administration systems, workload measurement systems etc. With management style becoming less top-down and more participative, translators were empowered to take on small coordination tasks or manage small projects (in dedicated task forces), which required training in a number of soft skills. The overall impact of all these factors was either a change in focus of translators' existing activities (e.g. translating new text types, post-editing instead of classic editing) or the need to learn to perform completely new activities like transcreation, cultural mediation and project management.

The factors driving this complex set of trends can be grouped and analysed in different ways, and what you see here is my personal take.

The first major factor is the digital revolution that has reshaped the communication landscape over the last few years. A series of technological inventions and breakthroughs have created new communication platforms as well as new ways of communicating – also multilingually. The speed of communication has increased exponentially, the tone has become lighter, the language simpler, and there is a flood of information at our fingertips, which requires new ways of processing.

Second, there is a set of factors which are specific to the ECB. The financial crisis of 2007-09, followed by the sovereign debt crisis in the euro area led to fundamental institutional innovations in the EU: as of November 2014, the ECB (together with the national authorities) was given the responsibility to supervise large banks operating in the euro area. As a result, in a matter of years, the organisation almost doubled in size, and support areas such as the language services were now supposed to serve a much larger organisation. The impact on the translation function was huge: new counterparts, new subject matters, new text types, a spike in demand (some of it cyclical, some of it completely unpredictable) etc. We were granted some additional resources, but this was not commensurate with the increase in workload.

Another part of the ECB's response was its decision to beef up the communication area both in terms of its policy communication as well as the need to keep up with the latest communication trends. For the translation function this meant the organisational (and functional) integration into the communication area. A new communications approach was adopted: communication was to become pro-active, focus more on non-expert audiences and satisfy the increased demand for transparency and accountability. All of this had a direct impact on our day-to-day work in terms of text types, registers, tools, deadlines, volumes and visibility.

Finally, the translation industry has also undergone sweeping changes. Here are just a few key aspects: (i) again, technology plays a key role (CAT tools, advances in MT), (ii) a broader range of client needs need to be met (shorter deadlines, broad variety of text types and file formats), and (iii) there is a huge pressure to drive down costs (through the use of technology, scaling up operations and heavy outsourcing).

3 A new vision

How did we go about transforming ourselves into an agile, cutting-edge service that is better placed to weather the challenges in years to come?

First of all, we needed to invest in job coordination. For example, we reorganised our coordination assistant teams and created several smaller teams with specialised tasks (contract management, translation task coordination, project teams). In addition, we created "team lead" positions, filled mostly by former translators with a coordination profile who would bridge the functional gap between translators and management, relieving managers of work-coordination tasks and adding an organisational layer with a strong operational background benefitting the translators.

Second, we switched to a more inclusive style of management and involved translators in certain projects and tasks. In particular, we created a set of task forces which deal with specific operational and organisational issues.

Third, we switched to a more pro-active stakeholder management to make sure that we can influence our requesters early on in the production process.

We increased our outsourcing activity especially in the languages heavily involved in banking supervision work. Over the years, we have introduced various measures to make a sufficient capacity buffer available to us at any point in time.

These measures made perfect sense and have increased our efficiency enormously, however the real game changers were the ones which were related to technology - simply because the common thread in the transformation that we were seeing was technology. Of the technology-related measures, three are key: investment in language technology, investment in business process development and adopting data-driven decision making.

4 Language technology tools

At the ECB, we have been using computer-assisted translation tools since 1998, namely Trados Studio and MultiTerm. In addition, we have a set of additional custom-made tools which automate certain processes to render them more efficient.

For example, special custom-made editors were developed by our webmasters which optimise the translation workflow of our website. These allow the translators to directly make changes on their language version of webpages before they are published. Given that our webmaster (naturally) do not speak all official EU languages, it makes this process much less error-prone.

Before the creation of these special text editors for website translation, the webmasters had to copy the translatable text from the html files into Microsoft Word which the translators translated and returned to the webmasters. The translations were then manually copied into the HTML files by the webmasters. Thereafter the translators proof-read the pages and had to report the linguistic changes to the webmasters in Microsoft Word, who then in turn tried to implement these linguistic amendments on the individual language web pages. This was an extremely inefficient and very frustrating process for both, the translators and the webmasters, as it involved a lot back and forward, esp. for languages the webmasters didn't speak or let alone read (e.g. Greek). It was very time-consuming. The new workflow involves the translation of html files in Trados Studio, with the proof-reading and finalisation of the webpage being done in special editors. Thanks to the special editors which are now used by the translators the whole process has become much leaner and streamlined and allows everybody to focus on what they can do best: the webmasters focus on the design while the translators can focus on the linguistic content.

We have developed a tool which automates the creation of the Trados Studio projects and automatically applies all language specific settings to the individual projects. This saves considerable amounts of time and allows the translators to start working sooner.

We also use a set of tools and macros to deal with repetitive and time-consuming tasks. These are developed by a small team of language technologists (6 people) who have a background in computational linguistics and translation. They assist the linguists and their freelance collaborators in case of any technical issues with the tools and workflows and perform a business-critical help desk function for the translators. The language technologists also offer advice on the workflow of particular file formats in order to make the most efficient use of the tools.

Over the past years there have been many changes also in the language technology field which have had a huge impact on the translators' work.

We started with the Translator's Workbench in 1998 translating only Word files. Today, we can translate a large variety of file formats for different purposes, using many different tools. It is a little like changing from using a horse and cart to now driving a Formula 1 car: it is much faster and the output that can be achieved is much higher.

In addition to the obvious reasons for introducing computer-assistant translation tools in 1998, we had an additional one, namely that we only had one translator per language, translating out of English. We therefore saw the need to record all our translations in a database to have this available for consistency when the translator went on holiday and was replaced by a freelance on site. As of 2005, we used an additional Trados application (TagEditor) which allowed the processing of more file formats.

As of 2008, we also had to translate more complex products, such as educational games which were translated in xml format, implying a big change for the translators.

In addition to the regular translation of Microsoft Office documents and html files, over the years we were increasingly asked to translate different products from different customers throughout the Bank. We needed to be agile and remain open to deal with different file formats.

For example, we were asked to translate websites that were hosted externally, such as the temporary external website that was set up in 2013 for the introduction of the new euro (Europa series) banknotes. The translators had to learn how to use a new content management system and work in that environment.

In the context of the new euro banknotes, we also translated educational games such as €uro Cash Academy, which was created to help people familiarise themselves with the banknotes' new security features. It still exists as an App.

The ECB has a very interesting Visitor Centre which unfortunately is temporarily closed due to the current restrictions but also exists as a virtual tour which you can access from the ECB's website. There are many digital interactive tools which had to be translated in different formats. For this project, the translators had to carefully consider the target audience and to adapt the style, register and terminology accordingly. You might like to have a look at it or come to see the Visitor Centre if you happen to be in Frankfurt once the pandemic restrictions are lifted.

5 Machine Translation

We cooperate with other EU institutions that run language services. For instance, we financially contribute to the running of the linguistic tools, including eTranslation (the machine translation tool developed by the European Commission).

Currently, eTranslation is used by two sets of users at the ECB – banking supervision experts and linguists. The banking supervision function produces a very high number of pages which must be translated in several language combinations. Given that the ECB's translation department is not staffed to handle these volumes of pages, it was agreed that the supervision experts would use eTranslation for gisting purposes and identify the parts of text for which they needed accurate human translation. Between 2014 and 2019, the supervisors were using the statistical machine translation engines of eTranslation (aka MT@EC). Since the rollout of the new neural engines (eTranslation), they have reported a considerable improvement in the translation quality.

Our linguists use machine translation in conjunction with the computer-assisted translation tools, as an additional resource which offers a translation if there is none in the translation memories. We consider this the most efficient way of using it, as it allows the best of both worlds given that you also receive the high-quality translations from the translation memories which were populated by professional linguists. The more technology is used, especially machine translation, the more it proves how critical the human component and added value is in the translation process. To arrive at the perfect translation with regard to style, message, register and terminology, the translators must always have the final 'word'. Together with the European Commission, we are currently creating a domain-specific engine with ECB documents as well as documents from the national central banks, which should yield even higher quality output. These engines will be ready by the end of August.

Machine translation is here to stay. We must embrace it, face it head-on and use it to our advantage.

6 Analytics

The Analytics team in the Language Services Division has developed a reporting component which monitors the usage of the machine translation tool, combining two metrics: the number of pages and the number of requests. The data shows that there is a steadily increasing trend with a high degree of daily volatility ranging from 0 to over 30,000 (pages) and from 1 to over 1,000 (requests), depending on the submission of documents from commercial banks of the euro area which cannot be the same every month; this increase in machine translation usage is especially noticeable since the rollout of the neural engine.

Data is compiled by language pairs and also on the annual translation output, in terms of pages and requests. In most cases the monthly eTranslation output is higher than annual human translation output.

Other data is being compiled to provide us with more granular information on the use of the machine in order for us to prioritise the training of the engines.

7 Translation workflow system

For a long time, translation jobs were managed via emails: requests came in emails, they were assigned to the translators concerned in emails and the final product was delivered in emails. As workload increased and processes became more complex, this system became untenable but for many years no viable solution was found. In 2012 we finally launched a project to set up such a workflow, based on a well-established IT system in the bank, to provide a platform with which translators, assistants and managers interact and without which we simply could not function these days.

The main challenge for the translators was that they had to learn to use yet another digital tool and completely alter the way they were managing their day-to-day tasks. The key change management challenge was to raise translators' awareness of the need to invest time in learning how to use the tool and to use it correctly, in order to reap long-term benefits.

In any institution, if you want to persuade your decision-makers of the validity of your argument or business case, you have to learn to speak their language, which in our case is the language of numbers. In addition, investing in data analytics was a logical consequence of creating a task management tool. This tool is a rich depository of operational data waiting to be explored and used for resource-related decisions.

The key change management challenge here was to gain staff acceptance that data is being collected on what they were doing ("Big Brother is watching you"-effect). We had to make sure staff understands that numbers are not (ab)used out of context and need to be interpreted carefully. We managed to get this staff buy-in primarily through information sessions and by highlighting success stories (getting new resources, conflict management, better planning).

8 Conclusion

In summary, there were several challenges which the translators have had to face and deal with over the past few years, notably: (i) the number of file formats involving different or specialised workflows; (ii) the increasing workload and handling different types of jobs simultaneously; (iii) tighter deadlines; (iv) an increased number of tasks (beyond pure translation), e.g. having to learn how to use new tools; (v) the different target audiences and platforms, e.g. experts vs general public; and of course (vi) the reluctance to change, which is only human.

Most recently, the coronavirus pandemic and the forced working from home since March 2020 entailed significant additional changes for the translators. They were forced to deal with technical problems by themselves since they could not rely on an on-site help desk function. It was a very challenging time both professionally and personally and they also had to learn how to work on remote platforms, with new communication tools (WebEx, Microsoft Teams, Jabber, etc), with only virtual access to support teams.

Over the years we have had to address these challenges and fears that the translators were facing. The changes are happening faster and faster. Furthermore, today's technology is no longer just an option which we can decide to use or not, but it is an absolute must. The change management process had to be planned very carefully, especially with the arrival of machine translation which could have been perceived as a professional threat to the translators' jobs.

To support the technological changes, since 1998 we have been offering dedicated expert language technology support by our language technologists who have become a critical player in the entire translation process given the huge role that technology plays in our translation environment today.

In order to arrive at more acceptance of changes, we involved the translators when planning translation workflows or when developing our new tools. This has proven very helpful and effective as the tools are being devised 'around' the translators' needs as much as possible. This also helped in the acceptance of the workflows and tools by the translators as they felt involved and saw the benefits of the new tools as they were designed to better suit their needs.

To mitigate the translators' scepticisms of machine translation, we invited speakers from academia, European institutions, and the private sector to talk about the topic and reinforce the message that machine translation is 'everywhere' and is here to stay. The speakers outlined the limitations but also the opportunities which machine translation can offer them. How to use machine translation to their advantage and at the same time to use it as a possibility to affirm the professional value of the human translator. Moreover, we involved the translators in the assessment projects of machine translation output, thereby bringing them closer to realising that it might be of use to them. We did not want to leave anybody behind but encourage all translators to move forward and keep abreast of technology. Over the past few years, we have noticed that a change process is going on in the mindset of our translators with regard to machine translation. We perceive an increasing openness, acceptance, and willingness to embrace it, as they are exploring how they can use it to their advantage.

To be able to handle the increased workflow, we have also been making increased use of outsourcing.

In addition to the training offered in the context of the language technology tools, and in order to prepare themselves better for all the different demands we are facing, translators have been attending training courses such as ‘Writing for the Web’ to better reach the dedicated target audience, ‘Post-editing training’ to better integrate machine translation output in the translation work and ‘Subtitling training’.

To conclude, the translator’s profile has changed dramatically and will continue to do so. The translator of the future certainly needs to be IT-savvy, detail-oriented and able to follow instructions, but at the same time be agile enough to switch between text types, be able to deal with multiple tasks quickly and be eager to learn new skills, both on the soft skills side and on the IT side. They will also need the ability to adapt quickly to unexpected situations and jump between micro and macro level tasks.

At the European Central Bank, we have come a long way and been successful in embracing the changes that these turbulent times have presented to us. However, this is a journey that has not yet ended and if there is one thing that we have learned: public organisations of the future cannot afford to lag behind when it comes to the fast pace of change in the world around them.

Interpreting and Technology:
Is the Sky Really the Limit?

Gloria Corpas Pastor[1-2]

[1] University of Malaga, Malaga 29071, Spain
[2] University of Wolverhampton, Wolverhampton WV1 1LY, United Kingdom
gcorpas@uma.es

Abstract. Nowadays there is a pressing need to develop interpreting-related technologies, with practitioners and other end-users increasingly calling for tools tailored to their needs and their new interpreting scenarios. But, at the same time, interpreting as a human activity has resisted complete automation for various reasons, such as fear, unawareness, communication complexities, lack of dedicated tools, etc.

Several computer-assisted interpreting tools and resources for interpreters have been developed, although they are rather modest in terms of the support they provide. In the same vein, and despite the pressing need to aiding in multilingual mediation, machine interpreting is still under development, with the exception of a few success stories.

This paper will present the results of VIP, a R&D project on language technologies applied to interpreting. It is the 'seed' of a family of projects on interpreting technologies which are currently being developed or have just been completed at the Research Institute of Multilingual Language Technologies (IUITLM), University of Malaga.

Keywords: Automation, Computer-assisted Tools, Interpreting Technologies

1 Introduction

Despite the actual interdependence between technology and the human skills in interpreting (Jekat, 2015), interpreting technologies are reportedly scarce (Costa, Corpas Pastor and Durán Muñoz, 2014) and have entered the profession only in recent years (Fantinuoli, 2018). Some studies suggest that interpreters are still largely unaware of them or even reluctant to use them (Corpas Pastor and Fern, 2016). Major concerns are the loss of quality and the dehumanisation of interpreting that allegedly tend to accompany technological developments (Jourdenais and Mikkelson, 2015).

However, there is a growing interest for language technologies and digital resources in the field of interpreting. See, for instance, the number of related papers presented in relevant conferences and workshops (e.g., the 8th AIIC Interpreters for In-

Translation and Interpreting Technology Online, pages 15–24,
July 05-07, 2021.
https://doi.org/10.26615/978-954-452-071-7_003

terpreters Workshop, 2017[1]; the two editions of HiT-IT[2], 2017, 2019; and all editions of Translating and the Computer since 2017[3]). Interpreters' attitude has also evolved in recent years. New generations of interpreters seem to be ready to embrace technology (cf. Corpas Pastor, 2018), although levels of interpreting technology uptake among interpreters remain quite low (Kerremans et al., 2019; Gaber and Corpas Pastor, 2020).

While technology changes and developments have paved the way for profound transformations in the discipline, the academic debate is just starting to address these changes, their implications and the challenges that lie ahead. Suffice to mention seminal contributions by Fantinuoli (2018a, 2018b), Mellinger and Hanson (2018), Braun (2019) or Drechsel (2019), and the papers in the volume edited by Rodríguez Melchor, Horváth and Ferguson (2020).

This paper will present a R&D project aimed at providing technology solutions for the pressing needs of both practitioners and trainees. The first part of this paper will describe the project components, and report preliminary results about users' perceptions and system performance. This first project has given rise to several complementary projects that will be also briefly covered in the second part. We will conclude with a discussion of possible new avenues of research that could impact this emerging field.

2 VIP - Voice-text Integrated system for interPreters

The VIP project seeks to provide technology solutions for the pressing needs of both practitioners and trainees. VIP integrates a suite of tools to assist interpretation at all phases, plus an open catalogue of interpreting-related technologies (tools and resources). The system architecture is described below.

2.1 Portal

The VIP portal contains a catalogue of interpreting-related tools and resources. This collection of semi-structure data is surveyed by means of a relational database management system (RDBMS). This RDBMS uses the SQL (Structured Query Language) to access the catalogue database.

The catalogue can be searched by individual categories and subcategories. For more refined searches, keywords can also be entered to locate tools and resources with specific features that appear in their description field. They are first classified by their main category and subcategories, and then, further characterised through a general feature-based template. General advanced searches query the database through the categories and subcategories and the basic feature template (platform, languages, license available) and keywords. Specific advanced searchers can be performed for Terminology management systems and Cloud-based interpreting systems.

The VIP portal is an open and collaborative database that includes language technology intended to enhance interpreter's performance, technology aimed at delivering

[1] https://aiic.de/event/8-dolmetscher-fuer-dolmetscher-workshop/. (Last accessed: 05-04-2021)

[2] http://rgcl.wlv.ac.uk/hit-it2019/. (Last accessed: 30-06-2021)

[3] https://www.asling.org/. (Last accessed: 13-04-2021)

an interpretation remotely, technology intended to replace human interpreters, computer-assisted interpreting training tools, resources and e-learning platforms, and miscellaneous resources.

Computer-assisted interpreting (CAI) tools. The first category of CAI tools includes terminology management tools designed for and intended to be used by interpreters (Intragloss, InterpretBank, Interplex UE, Interpreters' Help, Flashterm, etc.). Besides the basic features mentioned above, specific information is also provided for each tool (documentation, available, export and import formats, author, further information). Note-taking applications have not been specifically designed with interpreters in mind, but they are increasingly being used in digital consecutive interpreting and in hybrid interpreting modalities (SimConsec and SightConsec). They are further divided into standalone software and smart pens (e.g., Evernote, Livescribe).

Speech-to-text applications (also termed S2T and ASR, automatic speech recognition) are currently being used as a central component of CAI tools, either bundled or standalone (cf. Gaber, Corpas Pastor and Omer, 2020). Only S2T standalone applications are included in VIP (e.g., Voice Dictation, Dragon NaturallySpeaking, etc.).

Remote interpreting (RI). Unlike telephone- and video-mediated interpreting, cloud-based RI usually involves two main components: (a) the Interpretation Management System, designed to schedule and manage interpreting assignments, and (b) the Interpretation Delivery Platform, designed to support the delivery of the interpretation. Some examples are Kudo and Interprefy.

Machine interpretation (MI). Although it cannot be considered interpreting-related technology proper, MI is expected to have a significant impact on professional interpreters' work environment (akin to MT in translation). They usually involve several complex language technologies in a three-phase cascade: (a) speech-to-text conversion, (b) machine translation, and (c) text-to-speech synthesis. Some examples are SpeechTrans and VoiceTra4U.

Computer-assisted interpreting training (CAIT). This category encompasses various training materials (oral resources, digitised interpretations, videos, transcribed speeches, portals, research projects, institutional multimedia repositories, etc.), and virtual training platforms (e.g., IVY, Virtual Interpreting Environment or Melissi VS).

Miscellanea. In addition, a broad category of miscellaneous resources that can aid interpreters is also included. They encompass terminology management tools used by translators that could be useful for interpreters (e.g., SDL Multiterm, TermSuite), units converters (e.g., Converto, Units), and other relevant speech technologies, like voice recording (e.g., Voice Pro, Audacity). Corpus tools appear as a separate category in the catalogue due to their increasing importance for the preparation phase of an interpreting assignment and most specifically for glossary building. They encompass tools for building, tagging, parsing and managing comparable corpora (Bootcat, SketchEngine) and multi-lingual parallel corpora (ParaConc, ParaVoz).

2.2 Modules

VIP also comprises a modular system that includes various functionalities and technologies. Initially, components were allocated modules according to their main purpose: to help interpreters prepare for an interpreting job, to assist interpreters during an interpretation, or to train interpreters. In the second phase of this project (VIP II), components have been grouped according to their complementary nature or combined functionalities (cf. Section 3).

Module I. Designed to be used in the preparation for an interpreting assignment, this module comprises four main functionalities: (i) corpus management, (ii) glossary management, (iii) named entity recognition (NER), and (iv) automatic text summarisation. Corpus management offers different functionalities related to corpora: automatic and user-assisted corpus compilation (webcrawling), uploading of corpora, and corpus query (concordances, right/left sorted KWIC, n-grams, patterns, candidate terms). Glossaries can be created from corpora or manually compiled. Dictionary and glossary management allows users to create, upload and delete glossaries, perform external searches to locate translation equivalents or, else, translate terms automatically by using machine translation and post-editing. Automatic bilingual glossary creation of multiword terms and postediting through external searches is also possible.

Named entities (NEs) can be extracted automatically (NER) by pasting a text or uploading a corpus. NEs are then retrieved according to a set of predefined categories: location (LOC), person (PER), organisation (ORG), etc., and highlighted within the text or, else, as tables. They can be also added automatically to a given glossary.

Text summarisation allows users to produce a domain survey on any topic automatically, either by uploading texts or by selecting several documents. The domain survey can be also downloaded as a corpus and managed as such. This option is particularly relevant to extract key terms for a specific topic. Key terms can be then used recursively as seed terms for automatic corpus compilation or added automatically to glossaries, etc.

Module II. Intended to be used when delivering an interpreting job, it includes: (i) automatic note-taking, (ii) machine translation and (iii) glossary query. Automatic note taking incorporates speech recognition and automatic transcription. The system detects NEs and numbers, including physical magnitudes (e.g., *25 tons*) and common nouns (e.g. *2 rockets*). Glossaries created in Module I can be searched with Glossary query. This functionality provides instant access to terms, NEs and multiword expressions (MWEs), and to their translation equivalents (either by typing the first three letters or orally through the microphone). Machine translation is provided in case an instant draft equivalent is needed on the spot. VIP includes Translate Shell and VIP translator, a neural experimental system. While automatic note-taking could be easily used in sight translation and consecutive interpretation, in simultaneous modalities it would be more convenient for the interpreter’s booth mate.

Module III. Primarily designed for training student interpreters or for life-long learning purposes, it includes (i) a training module with exercises that are automatically generated, and (b) symbols for practising note-taking (experimental). Anticipation exercises enable users to practice terminology and phraseology from selected corpora.

Exercises with numbers are also customisable as regards range (e.g., from 1 to 1000), decimals, and language (Spanish, American English or British English). Based on ASR, this type of exercises enables users to practice listening/reading random numbers and then typing or saying the answers. The system indicates mistakes and provides the correct answers.

Sight-translation exercises also make use of ASR techniques, combined with parallel corpora management. The user is presented with a fragment of the source language subcorpus and produces the sight translation orally in the target language variety selected. Then, the system recognises the speech automatically, transcribes the user's spoken utterances automatically and checks the accuracy of his/her output against the aligned subcorpus 2 in the target language. The accuracy rate is approximate as it compares user's output with the actual bitext in the target language. Synonyms, term variations, syntactic transformations or paraphrases are considered errors by the system, similarly to standard translation memory systems. In addition, some errors could be due to the ASR system in place.

The fourth type of exercises in this category are intended to practice terms, multiword terms, and multiword expressions. Glossary exercises allow selection of the glossaries, language and diatopic variety configurations, and number of exercises desired. Users can practice with one or several bilingual glossaries of their choice, either orally or by typing. The system checks and provides correct answers.

Finally, note-taking training exercises combine speech technologies and artificial intelligence for image recognition. The tool displays terms/concepts randomly (spoken, but also written if the option is selected) for users to draw the corresponding symbol. The image is then processed automatically. A checking bar indicates the percentage of accuracy of the symbol with regard to the displayed concept.

2.3 Access to VIP and users' perceptions

The VIP system is freely accessible for research purposes. User licences can be requested from the Research Results Transfer Office (OTRI) of the University of Malaga (alinares@uma.es). A beta version of the VIP system has been tested on a number of occasions by various user groups. Lack of space prevents us from providing a full account. For this reason, only two user cases have been selected. The first one studied interpreters' perceptions after using VIP, while the second example replicates the experiment with translators.

A workshop with interpreters (professionals and trainees) was organised as part of TC42 (London, 2020). In this hands-on session participants could use the tool at ease to prepare for a blind interpretation and then provide feedback. In general, the VIP systems was rated either useful or very useful as a tool to prepare for an interpretation in a 0-5 Lickert scale (3=33%; 4=33%; and 5=22%), as well as quite intuitive and user-friendly (3=22%; 4=56%, 5=22%). And to the question, "which exercises do you like most?", participants selected Glossaries (42%), followed by Numbers (33%) and Anticipation exercises (25%).

Then, individual modules and functionalities were also evaluated as regards their usefulness in the preparation phase. The average rating was 4 out of 5. The results obtained are as follows: (a) corpus compilation: 3=20%; 4=40%; 5=40%; (b) corpus management: 4=55%; 5=45%; (c) text summarisation: 3=25%; 4=75%; (d) glossary creation: 3=8%; 4=58%;5= 33%.

As part of this workshop, participants were asked to provide feedback in order to improve VIP. The most repeated suggestions mentioned were adding more languages (the present version of VIP only supports English and Spanish), quick glossary search, inclusion of abbreviations, plus some technical issues, like browser compatibility, increased website capacity and future site maintenance.

A similar survey was used to test the system among translators (professional and students) that attended a postgraduate seminar at the University of Valladolid, Spain (March 2021). Participants were asked to use VIP in a mock translation project. The results are in line with the previous survey on interpreters. Most participants rated VIP useful (4=16.67%) or very useful (83,33%) for translation, and quite intuitive and user-friendly (4=66.67%). Among the main assets of the system participants highlighted the integration of various functionalities in one single platform, its simplicity and user-friendliness, and its fast performance. As to the relevance of various functionalities as an aid to translation, their preferences were, in descending order: (a) glossary creation (3=16.67%; 4=16.67%; 5=66.67%), (b) corpus compilation and corpus query (3= 16.67%; 4=33.33%; 5=50%); and (c) text summarisation (2= 16.67%; 4=33.33%; 5=50%). When asked what functionalities they found most useful, participants unanimously mentioned automatic corpus compilation and automatic glossary creation.

As to possible ways to improve VIP, most respondents mentioned adding more languages (French, German, Italian, Russian), and more functionalities specifically designed for translation purposes (for instance, filtering by language variety and specialised domain).

Our findings suggest that VIP could be equally useful to prepare an interpretation or translation assignment, although the specific needs of both user groups might be rather different. Further studies need to be conducted on the perception and usability of our system for different tasks and mediation modalities.

3 The R&D cluster on interpreting technologies

VIP was the first research project on interpreting technologies we were granted (ref. no. FFI2016-75831-P, 2016-2020). It opened several paths to explore the impact of technology in interpreting training and research. For instance, *INTERPRETA 2.0: application of ICT tools for the teaching-learning process of interpreting* provided teaching tools and resources for undergraduate students in order to foster autonomous learning and tech-savviness (ref. no. PIE 17-015, 2017-2019, University of Malaga). Other related initiatives worth mentioning were *Training Network on Language Technologies for Interpreters* (ref. no. EUIN2017-87746, 2017-2020, Spanish Ministry of Economy and Competitiveness), and *Application of Advanced NLP Techniques to the Field of Translation and Interpreting Technologies* (Ref. PIE 17-015, 2018-2020), already completed; as well as the *European Masters in Technology for Translation and Interpreting* (Ref. 599287-EPP-1-2018-1-UK-EPPKA1-JMD-MOB, European Commission) and the *Research network INTEC: Interpreting and New Technologies* (University of Malaga, 2021-2022), both of them still on-going.

The former are just training initiatives or networks of various kinds. In this section we will provide a brief overview of some of the research and/or transfer projects on interpreting technologies conducted within our research group (PI: Prof. G Corpas).

The first two investigate the possibilities of Natural Language Processing (NLP) and neural networks for the automation of interpretation, while the other two are extensions of VIP.

3.1 Multilingual dialogue systems using neural networks for apps in the healthcare domain: the triage (Spanish-English/Arabic)

The *triage* project (ref. no. UMA18-FEDERJA-067, ERDF, 2019-2021) falls within interpreting for the public services. It follows the digitalisation and technologicalisation trends in the sector. Specifically in the field of public services, so-called remote interpreting is beginning to be introduced, which allows the service to be offered by telephone or video conference software. This modality reduces the cost of interpreting, although it still requires a high investment on human interpreters. In a healthcare context, effective communication with the patient is essential for adequate and quality care. However, since interpreting services are expensive, not all hospitals and health centres can afford to treat foreign patients in their own language. This situation is particularly complex in the case of the Andalusian health system, due to migratory movements and the influx of tourists.

The ultimate goal of this project is the development of a multilingual system to automate triage. The term 'triage' refers to the process by which people are selected based on their need for immediate medical treatment when available resources are limited. Our focus is emergency triage, as it is the scenario that requires the fastest reaction time and has the least time to resort to external interpretation services. Central to the project is the design and implementation of an app for smartphones and other mobile devices, such as iPad, to enable effective communication between the healthcare professional and the patient which will allow patients to be assessed and ordered according to the severity and urgency of the case. The system is based on multilingual dialogical models and multimodal neural automatic translation (speech-text-speech). The system allows automatic translation/interpretation for the language pairs Spanish-English and Spanish-Arabic. In addition, patient' medical records can be generated in the three languages and stored in/retrieved from the hospital database. The Agencia Sanitaria Costa del Sol and the Hospital Costal del Sol collaborate in this project.

3.2 MI4ALL - Machine Interpretation for All Through a Deep Learning API

The advances of recent years in Artificial Intelligence (AI) are making it possible to develop applications that improve people´s lives at different levels. Within AI, the most widely used technique is Deep Learning (DL). Compared to other learning methods, this technique stands out for its high performance when solving problems and addressing tasks related to language and communication. Some examples with excellent results are found in machine translation (MT), automatic speech recognition (ASR), text generation, question-answering systems, and many other areas of NLP.

The MI4ALL project (UMA-CEIATECH-04, 2020-2022) aims at providing an automatic interpretation software platform that will combine DL and corpus linguistics to facilitate public services for foreigners and immigrants, allowing them to communicate in a language in which they are fluent. The platform consists of a REST API that offers automatic translation, speech recognition and voice synthesis functionalities that can be accessed through an app or integrated into other systems and can be retrained for other fields and languages. Multilingual and multimodal corpora are being compiled for Spanish, English, and Arabic. The data will be used to train the system and to assess its performance. A technology company (Intelligenia S.A.) is also involved in the development of the system.

3.3 Voice-text integrated system for InterPreters: Proof of Concept

Nowadays, technology permeates the translation industry The VIP project aims at improving interpreters' performance and work conditions by taking advantage of current developments and technology in a similar way. The resulting system is fit to purpose: it brings to the current state of the art a novel technological solution that can benefit interpreters and their work environment before, during and after an interpreting assignment.

VIP technology is mature enough to meet both functional and non-functional requirements and to be considered as a functional system. However, given that the system has been developed within the framework of a research project, it is not orientated towards obtaining an industrial and marketable tool. For this reason, the objective of this proof-of-concept project (ref. no. E3/04/21, 2021-2022) is to evaluate the system in terms of usability, effectiveness, security and robustness in order to establish clearly and precisely what changes and improvements the system needs to become a marketable product (end-up product). To this end, two companies have entered the project: Kudo and EL Translations.

3.4 VIP II - Multi-lingual and Multi-domain Adaptation for the Optimisation of the VIP system

The VIP II project II (ref. no. PID2020-112818GB-I00, 2021-2025) seeks to continue and extend the pioneering work carried out on the previously funded project (see Section 2 and subsections). The VIP system (version 1.0, henceforth VIP1) represents a new-generation of interpreting-related technologies that is based in interdisciplinary cutting-edge research. VIP1 has filled several major research gaps that were identified at the time of its submission, also during the project timeframe. To the best of our knowledge, it is the first open-source purpose-built integrated system designed to fulfill interpreters needs and requirement, that is intended to provide support to both interpreters and trainees.

VIP II intends to improve the system further. While internal and external evaluations of the tool have shown very good results in general, the findings also point out to the existence of areas in need of further research (e.g., ASR, multimodal corpora), functionalities that should be improved (e.g., note-taking, machine translation), new desired features, more languages, new integrations and improved functionalities, etc. Besides, it is also necessary to adapt to new interpretation scenarios in terms of technology uptake, real needs, and degree of automation in a rapidly changing world.

Our main aim is two-fold: to improve VIP1 to better accommodate the needs and requirements of interpreters (professional and trainees), and to establish the feasibility and impact of achieving automation in real interpreting scenarios. To this end, five specific objectives have been set up: to survey interpreters needs and technology uptake to enhance their performance, in general and for specific domains; to study the technical configurations of VIP1 with a view to developing an improved system (VIP2), taking into account the findings reported and in light of the latest research; to reuse and compile multilingual data (written, multimedia, oral) to increase the number of resources integrated into the system and to enhance adaptation (multilingual and multi-domain); to perform intrinsic and extrinsic evaluation of VIP2, the latter being indicative of the impact on users; and, finally, to design user cases to establish usefulness and benefits of VIP2 according to different interpreting modes, modalities, language-pairs, scenarios, and purposes.

4 Conclusion

By and large, interpreters seem to view technology as a key asset. Their attitude towards interpreter-related technology has undergone a positive development in recent years and they are willing to use technology in their daily practice. The question now is whether academia and developers are prepared to tap into interpreters' needs and provide them with the appropriate tools and resources.

At this stage, a handful of thought-provoking questions could be a good starting point. Below there is tentative list, including, but not limited to, the following: (i) Is interpreting-related technology considered friend or foe?; (ii) Will interpreting technologies replace interpreters?; (iii) Are interpreting technologies a key asset? If so, for whom?; (iv) What type of technology do interpreters need?; (v) How do interpreting stakeholders interact with technology?; (vi) Are interpreting technologies the "new normality" in the sector?; (vii) What is the role of artificial intelligence (AI), automatic speech recognition (ASR), speech-to-text (S2T), big data, neural networks, etc. in the future of the profession?

Too many questions and almost no answers to set off on a journey into the unknown. Perhaps the sky will be the limit… or maybe not.

Funding: This work was supported by the Spanish Ministry of Research (Ref. no. FFI2016-75831-P, 2017-2020; PID2020-112818GB-I00, 2021-2025) and the European Regional Development Fund (ERDF) (Ref. no. UMA18-FEDERJA-067, ref. no. E3/04/21, 2021-2022).

References

1. Braun, S.: Technology and interpreting. In O'Hagan, M. (Ed.). The Routledge Handbook of Translation and Technology. Routledge (2019).
2. Costa, H., G. Corpas Pastor, Durán Muñoz, I. Technology-assisted Interpreting. MultiLingual Computing 143, 25(3), 27-32 (2014).
3. Corpas Pastor, G.: Tools for Interpreters: the Challenges that Lie Ahead. Trends in Translation Teaching and Learning E, 5, 157-182 (2018).
4. Corpas Pastor, G., Gaber, M: Remote interpreting in public service settings: technology, perceptions and practice. Skase Journal of Translation and Interpretation 13 (2), 58-78 (2020).

5. Drechsel, A: Technology literacy for the interpreter. In: Sawyer, D. B., Austermühl, F., Enríquez Raído, V. (eds.) The Evolving Curriculum in Interpreter and Translator Education: Stakeholder perspectives and voices, pp. 259-268. John Benjamins, Amsterdam (2019).
6. Gaber, M., Corpas Pastor, G., Omer, A.: Speech-to-Text technology as a documentation tool for interpreters: A new approach to compiling an ad hoc corpus and extracting terminology from video-recorded speeches. TRANS: Revista de Traductología, 24, 1-18 (2020).
7. Fantinuoli, C.:Computer-assisted Interpreting: Challenges and Future Perspectives. In: Corpas Pastor, G.,Durán Muñoz, I. (eds.) Trends in e-tools and resources for translators and interpreters, pp 153-174. Brill, Leiden. (2018a).
8. Fantinuoli, C.. Interpreting and technology: the upcoming technological turn. In: Fantinuoli, C. (ed.). Interpreting and Technology, pp. 1-12. Language Science Press, Berlin. (2018b)
9. Jekat, S: Machine Interpreting. In: Pöchhacker, F. (ed.) Routledge Encyclopedia of Interpreting Studies, pp. 239-242. Routledge, London and New York. (2015).
10. Jourdenais, R., Mikkelson, H.: Conclusion. In: Mikkelson, H., Jourdenais, R. (eds.) The Routledge Handbook of Interpreting, pp 447-450. Routledge, London/New York. (2015).
11. Kerremans, K., Lázaro Gutiérrez, R., Stengers, H., Cox, A., Rillof, P: Technology Use by Public Service Interpreters and Translators: The Link Between Frequency of Use and Forms of Prior Training. FITISPos International Journal 6(1), 107-122 (2019).
12. Mellinger, C. D., Hanson, T.A: Interpreter traits and the relationship with technology and visibility. Translation and Interpreting Studies 13 (3), 366-392 (2018).
13. Rodríguez Melchor, M. D., Horváth, I., Ferguson, K. (Eds.): The Role of Technology in Conference Interpreter Training. (New Trends in Translation Studies 31). Peter Lang, Berlin, Bern, Brussells, New York, Oxford, Warszawa, Wien. (2020).

Machine translation use outside the language industries: a comparison of five delivery formats for machine translation literacy instruction

Lynne Bowker [1] [0000-0002-0848-1035]

[1] School of Translation and Interpretation, University of Ottawa, Canada
lbowker@uottawa.ca

1 Introduction

Since the launch of the free online tool Google Translate in 2006, which has been followed by the release of a host of similar tools (e.g. Microsoft Bing Translator, DeepL Translator, SYSTRAN Translate, Baidu Translate, Yandex.Translate, Naver Papago), machine translation (MT) has been easily accessible to anyone with an internet connection. Not only are machine translation tools easy to access, they are also easy to use. In many cases, users need only choose their language pair, copy and paste a text, and click "Translate". In other cases, a machine translation widget may be embedded in a web browser or social media platform, meaning that translation is just a click away. It is very easy to see the appeal of a tool that is free, fast, and easy to use! Therefore, it comes as no surprise that these tools are indeed being used widely. While language professionals certainly constitute an important user group, they are by no means the only one. Indeed, various groups outside the language professions use machine translation actively:

- Anazawa et al. [1] describe how practicing nurses in Japan use machine translation to stay on top of the latest developments in the international nursing literature;
- Bowker and Buitrago Ciro [2] explore the use of machine translation by researchers seeking to publish in other languages;
- Nurminen [3] recounts how patent professionals use machine translation to search for international patents;
- O'Brien and Ehrensberger-Dow [4] note that machine translation is sometimes used to support communication in a crisis situation.

In all of these cases, the authors emphasize that some kind of training can help tool users to make better decisions about employing machine translation and to optimize its use. What's more, authors such as Mundt and Groves [5] and Lee [6], among others, have identified university students as a very active group of machine translation users, and what better place to offer machine translation literacy instruction than at a university?

However, even at a university, instruction can take many forms, and over the past couple of years, we have had the opportunity to try out five different formats for delivering machine translation literacy instruction. In this paper, we first introduce the basic

Translation and Interpreting Technology Online, pages 25–36,
July 05-07, 2021.
https://doi.org/10.26615/978-954-452-071-7_004

notion of machine translation literacy and share some general content that we believe could usefully be included in a machine translation literacy module for non-translation students. Next, we briefly present the five different machine translation literacy instruction formats that we have pilot tested, as well as some general feedback received from the participants. This is followed by a comparative summary of some strengths and weaknesses of each format, along with some general conclusions.

2 Machine translation literacy

As noted previously, machine translation tools are easy to access and straightforward to use, but this does not mean that people without a translation background instinctively know how to use these tools critically. Machine translation literacy is less about knowing which buttons to press and more about deciding whether, when or why to use this technology [2]. In this way, it has a strong cognitive or conceptual element that focuses more on critical thinking tasks, such as evaluating the suitability of a text for translation by machine, or weighing the benefits and risks of using machine translation against other translation solutions. Owing to space limitations, it is not possible to provide a comprehensive description of the contents of a machine translation literacy module; however, key elements that can be usefully covered as part of such a module are briefly summarized below. It is also important to recognize that machine translation literacy does not take one single form; rather, it is a customizable concept that can (and should) be adapted to meet the needs of the target audience. The summary below focuses on the needs of university students who are not studying to become language professionals, but machine translation literacy instruction for other groups (e.g. primary or secondary school students or teachers, translator trainees, journalists, health care workers, workers in NGOs) may incorporate different elements or explore them to a different depth.

For non-translation undergraduate students, the basic machine translation literacy module that we designed had four main components, which were covered in more or less depth, depending on the format and time available:

1. Understanding data-driven approaches to machine translation
2. Transparency and machine translation use
3. Risk assessment and machine translation
4. Interacting with machine translation

2.1 Understanding data-driven approaches to machine translation

Having a basic understanding of how data-driven approaches to machine translation (including neural machine translation [7]) work will enable students to better understand the strengths and limitations of these tools. For instance, understanding the notion of sensitivity to training data can help users to realize why these tools can be more or less useful for different language pairs, domains or text types (e.g. low vs high resource situations).

Students who understand how data-driven machine translation systems work will also recognize that different tools (e.g. Google Translate, DeepL Translator, etc.) are likely to produce different results. Many of the non-translation students that we worked with had previously believed that while the interfaces of these tools may differ slightly, all machine translation tools were driven by the same engine and would produce the same results. It had not occurred to many of these students that they would get different results by trying different systems (which had been trained using different corpora), and many had never looked beyond Google Translate. Similarly, they had not realized that the systems were constantly “learning” and so the results may improve from one trial to the next, and they should not write off a tool as being unhelpful based on one experience.

Finally, learning about sensitivity to training data also makes students aware of the potential for algorithmic bias, including problems such as inappropriate selection of pronouns in languages that are marked for gender [8].

2.2 Transparency and machine translation use

The concept of transparency is relevant in several ways for student users of machine translation tools. Firstly, it may be important to point out that the use of machine translation for course work may be more or less appropriate depending on the learning objectives of the course and the preferences of the instructor. It is also important to emphasize academic integrity with regard to the need to properly cite and reference material that has been translated from another language; the wording may change, but the original author should still be cited as the source of the ideas. Another reason that transparency is important is that it allows the readers of the text to take the fact that it has been machine translated into account as part of their own decision-making when deciding how much to trust the content. For all these reasons, students are encouraged to be transparent about their use of machine translation.

2.3 Risk assessment and machine translation

The notion of transparency has links to the idea of risk assessment. For students without a background in translation, the idea that translations can have different purposes and take place in different contexts may not be immediately apparent. Students need to learn to evaluate different types of translation tasks and recognize them as being low-stakes or high-stakes tasks where the use of a machine-translated text may carry a lower or a higher risk. This could include educating students about the differences between using machine translation for information assimilation (e.g. understanding a friend’s social media post) versus text dissemination (e.g. submitting an essay to a professor), or the difference between using machine translation to compose an email to a friend versus composing an email to a prospective employer.

Another type of risk assessment that is relevant to students is determining whether the material that they want to translate is sensitive or confidential. Most of the students that we worked with had not given much thought to what happens to the text that is entered into a free online system, and many were surprised to learn that this text does not simply disappear once they exit the tool. Making students aware that they should not enter sensitive information (e.g. banking details, health information, proprietary research) into a free online machine translation tool is an important component of machine translation literacy for this group.

2.4 Interacting with machine translation

Finally, the students that we worked with were eager to learn about how they can interact with a machine translation tool in order to improve the quality of the results. While the vast majority of students are aware of the likely need to make some adjustments to the output, few have given any consideration to the idea that changing the input to reduce ambiguity can result in higher quality output. Therefore, if the goal is to use machine translation as a writing aid to help produce a text in a second language, it may be easier for students to make adjustments to their input text, which is likely written in their dominant language. While this idea of "garbage in, garbage out" seems very obvious to translators, it is not necessarily something that occurs to people outside the language professions.

Depending on the language combinations of the participants, it may be easier or more difficult to work together on practical exercises on pre- or post-editing. The tips that are relevant for one language, or the errors that are made by a given system, may not be the same as those that are typical for another. Nonetheless, students consistently expressed an interest in gaining hands-on experience with pre- and post-editing, so if it is not possible to include this in the instruction session, it could be worth preparing a resource sheet and some exercises that participants can work on later.

3 Machine translation literacy instruction formats

In the 20-month period between October 2019 and May 2021, we had the opportunity to test five different formats for delivering machine translation literacy instruction to undergraduate students who are *not* studying to become language professionals. These five formats include: 1) a library workshop; 2) an English-as-a-second-language course; 3) a translation for non-translators course; 4) an information literacy course; and 5) a digital humanities summer institute course. Below, we will present a summary of these experiences, including some highlights from student feedback.

3.1 Library workshop

The first format that we tested for delivering machine translation literacy instruction to university students took the form of an optional one-hour workshop offered in the autumn semester of 2019 through the university library at two different institutions in Canada: the University of Ottawa and Concordia University. At each institution, this workshop was promoted through the library, the international students office, and the student success centre. The short length of the workshop meant that it was necessarily high-level and was mostly a lecture-style format with some time for questions and answers. Students were given a resource sheet and some ideas for practical exercises (e.g.

tips on pre- and post-editing, comparing different machine translation systems) that they could take away and try at home. Between 25 and 30 students attended at each institution. The vast majority were undergraduate students, although a handful of graduate students participated. The students came from a wide range of disciplines in both the humanities and sciences, and they spoke a diverse range of languages (although none were Anglophone).

At the end of the workshop, participants were asked to provide a short evaluation of the workshop. Most claimed to find it valuable, noting that they had learned new things. Suggestions for improvement included using fewer specialized terms and giving more language-specific tips for pre- and post-editing. More time for practical exercises was also identified as something to strive for in future iterations. More details about this experience can be found in Bowker et al. [9].

3.2 English-as-a-second language course

The second format that we tested consisted of integrating a module on machine translation literacy into a course on English-as-a-second language (ESL). Once again, we tried this experiment at both the University of Ottawa and at Concordia University. At the University of Ottawa, all 22 students in the class were native speakers of Chinese, but they were studying a wide range of subjects. In contrast, at Concordia University, the 23 students spoke a range of languages, but all were studying business and the course therefore focused on business English.

In addition, at Concordia University, we were also invited to offer the workshop to a group of 24 ESL instructors prior to delivering it to the students. This was organized as a type of professional development opportunity for the instructors, and it was very informative because it allowed us to better understand the concerns of the instructors and to adapt the workshop content accordingly.

At both institutions, we essentially participated as a guest speaker to deliver the one-hour workshop in a single class, and again, the short timeframe for delivering the workshop limited the amount and depth of material that could be shared or the number of practical exercises that could be undertaken. Once again, students claimed to find the content interesting and relevant overall, but they expressed a desire for more time to be dedicated to practical exercises.

Additional details about the experience of integrating machine translation into an ESL course at the University of Ottawa, and in a “train the trainers” format and ESL course at Concordia University can be found in Bowker [10] and Bowker [11] respectively.

3.3 Translation for non-translators course

Next, we attempted to move away from the standalone workshop format or guest lecture delivery and to better integrate machine translation literacy into a broader course. In the Fall 2020 semester, we designed a full-semester (12-week) course on translation for non-translators at the University of Ottawa. The course was offered at an introductory (first-year) level and was open to students across the whole campus, regardless of their major area of study. The course proved to be popular and attracted 50 students, so it was offered again in the Winter 2021 and Summer 2021 semesters, with a similar result. Over the three iterations of the course, students came from more than 20 disciplines and more than 20 different native languages were represented (including English).

Because the course was an introduction to translation more broadly and not only to machine translation, it meant that students had an opportunity to learn some key translation concepts first, and therefore to have a slightly firmer footing in some of the basics before undertaking the machine translation literacy module in the fifth week of the course. In addition, because the course unfolded over a longer period, we were able to increase the amount of time spent on machine translation literacy to three in-class hours, and to increase the level of practical activity, both in class, but also in the form of homework and assignments to be conducted before and after class.

Another benefit of offering machine translation literacy in this format is that we reached many students who have English as a native language – a group that had not been represented at all in our prior efforts to deliver this training through the library or in an ESL class. The Anglophone students confirmed that they too are active users of machine translation, though more for their leisure activities than for their studies. Nonetheless, according to the results of the feedback survey, they claimed to find the machine translation literacy training to be informative, and overall, the majority of students in the translation for non-translators course recommended that this type of training be offered regularly and on a wide-scale across campus.

Additional information about the course on translation for non-translators and how it incorporated machine translation literacy will be shared in an upcoming publication [12].

3.4 Information literacy course

Up to this point, our efforts had focused mainly on targeting venues where participation was elective, so as a next step, we investigated the possibility of embedding machine translation literacy instruction into a compulsory course that focused not on language or translation but rather on digital and information literacy. One reason for doing this is that we wanted to see how machine translation literacy instruction would be received by those who were not expressly looking for it. At the University of Ottawa, all first-year students in the Faculty of Arts are required to take a minimum of four compulsory courses that focus on developing critical thinking and academic writing skills. These courses must be selected from a pool of courses offered by the Department of Philosophy, the Department of English, and the Interdisciplinary Studies program. Each year, the Faculty of Arts invites professors from different departments within the Faculty to team up and pitch a theme for an interdisciplinary course that meets the following requirements:

> **AHL 1100 Introduction to Interdisciplinary Study in the Arts (3 units)**
> Exploration of at least two disciplines in the Faculty of Arts whose conjunction illuminates contemporary situations and debates. Development of critical reading and academic writing.

> This course has variable topics. Students may take this course twice with different topics.

For the Winter 2021 semester, we successfully pitched the topic “New Literacies for the Digital Age” to be co-taught by a professor from the School of Information Studies and a professor from the School of Translation and Interpretation (i.e., the present author). A total of 80 students registered for the course, and they came from 13 different programs in the Faculty of Arts. While the majority were native English speakers, there were also speakers of eight other languages, including 14 Chinese speakers, 8 French speakers, and a smaller number of speakers of Arabic, Hindi, Persian, Spanish, Ukrainian and Vietnamese.

Along with modules on more traditional aspects of information literacy (e.g. effective searching in library catalogues and on the internet, referencing and citation), the course also contained instruction on media literacy (e.g. fake news) and scholarly communication literacy (e.g. predatory publishing), as well as a module on machine translation literacy. Similar to the case of the module that had been integrated into the course on translation for non-translators, it consisted of three in-class hours, along with some homework and assignments to be done outside of class. Once again, as reported on the feedback survey, the vast majority of the students claimed to find the machine translation literacy module to be useful and they recommended that it should continue to be offered as part of a compulsory course on information literacy.

To learn more about the experience of integrating machine translation literacy instruction into a broader course on digital and information literacy, consult Bowker [13].

3.5 Summer institute course

The final format in which we piloted the delivery of machine translation literacy instruction was as part of the 2021 Digital Humanities Summer Institute: Technologies East (DHSITE) [14], which is open to students from all disciplines and levels who are interested in exploring aspects of the Digital Humanities (DH). The summer institute took place during the last two weeks of May 2021, and it consisted of an offering of six 18-hour mini-courses on different subjects in DH (e.g. Python programming, text analysis, linked open data), from which students could choose up to two. We offered a course on machine translation in which there were eight participants, including both undergraduate and graduate students. Of these, three had a background in translation, while the other five came from disciplines that included computer science, business, music, psychology, and public administration. Four different native languages (English, French, Chinese and Polish) were represented. The diversity of backgrounds, languages and levels brought richness to the discussions but also posed challenges with regard to pitching the material appropriately.

As this format had 18 hours of in-class time, as well as additional time for homework outside of class, it was possible to explore the subject of machine translation much more deeply than in the previous formats. This meant that, in addition to the key elements of machine translation literacy content described previously, there was also time to consider the history of machine translation, methods of evaluating machine translation systems as well as their output, and a broader range of ethical issues surrounding tool use. At the time of writing, the formal course evaluations have not yet been received, but anecdotally, we can report that the participants were active and engaged throughout the course, and even students with a translation background appeared to be learning new things.

4 Comparison of different formats for delivering machine translation literacy instruction

Having experimented with five different formats for delivering some kind of machine translation literacy instruction, we can observe that the various formats have different strengths and weaknesses. We have summarized some of the main pros and cons in Table 1.

	Strengths	**Weaknesses**
Library workshop	• Low level of commitment required (for both participants and instructors) • Open to anyone on campus • Potential for immediate feedback • Potential to gauge interest in a more advanced follow-up workshop	• Challenging (and time-consuming) to promote • People don't recognize that they need it and so may not register • Participants have no background in translation so it's a steep curve • Very short, resulting in superficial treatment and limited practice • May only be offered once or twice per year (will take a long time to reach a critical mass of people) • No opportunity for longitudinal observation (e.g. to see if the information is put into practice or if behaviour changes over time) • "Train the trainer" required beforehand if delivered by a non-translator
Integrated into a compulsory English-as-a-second language course	• No need for marketing • Can reach a wide range and large group of international students • Can work with authentic texts in the context of course requirements (e.g. texts students need to produce for assignments)	• May meet resistance from language teachers who fear that MT use may be contrary to language learning objectives • May be misinterpreted by students who could perceive MT use as a shortcut to alleviate the need to learn a language

	• MT has potential to act as an aid for language learning and reinforcement (presents MT in a positive light, rather than as a taboo or shameful practice)	• Participants' knowledge of translation is limited and often restricted to a language learning context • "Train the trainer" required beforehand if delivered by a non-translator • Will not reach Anglophones or those with a high level of English
Integrated into an optional translation course for non-translators	• Participants are interested and motivated to learn about translation • Participants learn some basic translation concepts first and can build on these in the MT literacy module of the course • Can spend more time on it, and incorporate more practical work (e.g. homework, exercises) • Opportunity for more longitudinal observation (e.g. to see if knowledge is put into practice or leads to a change in behaviour) • Can reach both English speakers and speakers of English as an additional language • Course taught by a translation professor already up to speed (or able to get up to speed quickly) on MT	• High level of commitment required by participants (must take a whole course on translation, not just a module on MT) • As an optional course, it will only reach those who are actively seeking this knowledge
Integrated into a compulsory information literacy course	• Reaches a wider range of students (including those who may not realize they need it, and English speakers, who may not think MT is as relevant to them) • Can spend more time on it (3-6 hours), and incorporate more practical work (e.g. homework, exercises)	• Participants have no background in translation • "Train the trainer" required beforehand if it is to be delivered by a non-translator
Digital Humanities Summer Institute course on MT	• Participants are highly motivated to learn about MT • Can explore concepts thoroughly and incorporate more practical work (e.g. homework, exercises) • Course taught by a translation professor already up to speed on MT	• Reaches relatively few students • High level of commitment required by participants • Challenging to manage different backgrounds and levels of prior knowledge

Table 1. Comparative summary of some strengths and weaknesses of different formats for delivering machine translation literacy instruction.

5 Concluding remarks

Free online machine translation systems are very attractive because they are easily accessible and easy to use. However, this does not mean that users – especially those without a background in translation – instinctively know how to use them in a critical way. Therefore, there is an emerging need for machine translation literacy, and correspondingly, a need for machine translation literacy instruction. Having said that, there is no single right way to help users develop machine translation literacy. Rather, as noted previously, this is a highly customizable concept, and the content can (and should) be adapted to meet the specific needs of the target audience. In this article, we have focused on sharing the results of our efforts to teach machine translation literacy to (primarily) undergraduate students who are *not* training to become translators or other language professionals. We have tested five different delivery formats, each of which have strengths and weaknesses that may make them more or less suitable for different contexts. While no approach is perfect, we believe that all have some value and indeed the feedback received in each case was largely positive (though there was always room for improvement).

A general take-away from this experience is that it confirms that students of many backgrounds are eager to learn how to make better use of machine translation, and it is very important to recognize that things which are obvious to language professionals are *not* obvious to those without a translation background (e.g. GIGO, different machine translation systems generate different results). In other words, there is no reason why students would instinctively know how to be informed and critical users of machine translation tools, so there is scope for and benefit to offering some form of machine translation literacy instruction to this group.

Beyond undergraduate students, machine translation literacy instruction is relevant for other groups too, including those in the language professions, but also graduate students and more established scholars, as well as secondary or even primary school students. Indeed, the next delivery format that we will be piloting during Canada's Science Literacy Week in September 2021 is a machine translation literacy workshop for teens that will be delivered in collaboration with the University of Ottawa's Faculty of Engineering Outreach team.

Moving forward, a key question to consider when planning machine translation literacy instruction is *who* will do the training. Will it always be done by a machine translation expert or even a language professional? If machine translation literacy is to become embedded in other contexts (e.g. information literacy instruction, digital literacy instruction, ESL teaching, high school or primary school), then it will be necessary for people from other backgrounds (e.g. librarians, teachers) to become involved in delivering machine translation literacy instruction. In cases where the training will be delivered by a non-language professional, some type of "train the trainer" preparation will likely be necessary, as was done for the ESL instructors at Concordia University [10]. With a view to helping to "train the trainers", we are in the process of developing a range of resources for machine translation literacy, which can be found on the Machine Translation Literacy Project website [15]. Some additional resources are also available through the European Union Erasmus+ project "MultiTraiNMT — Machine Translation training for multilingual citizens" [16, 17].

Acknowledgements

The Machine Translation Literacy Project is funded (2020-2026) through an Insight Grant (#435-2020-0089) from the Social Sciences and Humanities Research Council of Canada. Seed funding was also provided by Concordia University Library through the 2019 Researcher-in-Residence program.

References

1. Anazawa, R., Ishikawa, H., & Kiuchi, T.: Use of online machine translation for nursing literature: A questionnaire- based survey. *Open Nursing Journal, 7*(1), 22–28 (2013).
2. Bowker, L., & Buitrago Ciro, J.: *Machine translation and global research*. Bingley: Emerald (2019).
3. Nurminen, M.: Raw Machine Translation Use by Patent Professionals. A case of distributed cognition, *Translation, Cognition & Behavior, 3*(1), pp. 100–121 (2020).
4. O'Brien, S., & Ehrensberger-Dow, M.: MT literacy: A cognitive view, *Translation, Cognition & Behavior 3*(2), pp. 145–164 (2020).
5. Mundt, K., & Groves, M.: A double-edged sword: the merits and the policy implications of Google Translate in higher education, *European Journal of Higher Education 6*(4), pp. 387–401 (2016).
6. Lee, S.: The impact of using machine translation on EFL students' writing, *Computer Assisted Language Learning, 33*(3), pp. 157–175 (2020).
7. Koehn, P.: *Neural machine translation*. Cambridge: Cambridge University Press (2020).
8. Monti, J.: Gender issues in machine translation: An unsolved problem? In: L. von Flotow & H. Kamal (Eds.), *Routledge Handbook of Translation, Feminism and Gender*, 457–468. London: Routledge (2020).
9. Bowker, L., Kalsatos, M., Ruskin, A., & Buitrago Ciro, J.: Artificial Intelligence, Machine Translation, and Academic Libraries: Improving Machine Translation Literacy on Campus, *The Rise of AI: Implications and Applications of Artificial Intelligence in Academic Libraries*. (Eds. S. Hervieux and A. Wheatley). Chicago: *Association of College & Research Libraries (ACRL)* (2021).
10. Bowker, L.: Chinese speakers' use of machine translation as an aid for scholarly writing in English: A review of the literature and a report on a pilot workshop on machine translation literacy, *Asia Pacific Translation and Intercultural Studies*, 7(3): pp. 288–298 (2020a).
11. Bowker, L.: Machine translation literacy instruction for international business students and Business English instructors, *Journal of Business and Finance Librarianship*, 25(1-2): pp. 25–43 (2020b).
12. Bowker, L.: *Introducing translation to non-translators*. London: Routledge (forthcoming).
13. Bowker, L.: Promoting linguistic diversity and inclusion: Incorporating machine translation literacy into information literacy instruction for undergraduate students, *The International*

Journal of Information, Diversity and Inclusion 5(3) (2021). https://jps.library.utoronto.ca/index.php/ijidi/issue/archive

14. Digital Humanities Summer Institute: Technology East (DHSITE): https://dhsite.org/dhsite-2021/
15. Machine Translation Literacy Project: https://sites.google.com/view/machinetranslationliteracy/
16. MultiTraiNM—Machine Translation training for multilingual citizens: https://www.multitrainmt.eu/index.php/en/
17. Ramírez-Sánchez, G., Pérez-Ortiz, J.-A., Sánchez-Martínez, F., Rossi, C., Kenny, D., Superbo, R., Sánchez-Gijón, P., & Torres-Hostench, O.: MultiTraiNMT: Training materials to approach neural machine translation from scratch. *TRITON 2021: Proceedings of the Conference* (2021).

NoDeeLe: A Novel Deep Learning Schema for Evaluating Neural Machine Translation Systems

Despoina Mouratidis[1][0000−0002−2844−5488], Maria Stasimioti[2][0000−0001−9541−4676], Vilelmini Sosoni[2][0000−0002−9583−4651], and Katia Lida Kermanidis[1][0000−0002−3270−5078]

[1] Department of Informatics, Ionian University, 491 00 Corfu, Greece
{c12mour,kerman}@ionio.gr

[2] Department of Foreign Languages, Translation and Interpreting, Ionian University, 491 00 Corfu, Greece {stasimioti,sosoni}@ionio.gr

Abstract. Due to the wide-spread development of Machine Translation (MT) systems—especially Neural Machine Translation (NMT) systems—MT evaluation, both automatic and human, has become more and more important as it helps us establish how MT systems perform. Yet, automatic evaluation metrics have lagged behind, as the most popular choices (e.g., BLEU, METEOR and ROUGE) may correlate poorly with human judgments. This paper seeks to put to the test an evaluation model based on a novel deep learning schema (NoDeeLe) used to compare two NMT systems on four different text genres, i.e. medical, legal, marketing and literary in the English-Greek language pair. The model utilizes information from the source segments, the MT outputs and the reference translation, as well as the automatic metrics BLEU, METEOR and WER. The proposed schema achieves a strong correlation with human judgment (78% average accuracy for the four texts with the highest accuracy, i.e. 85%, observed in the case of the marketing text), while it outperforms classic machine learning algorithms and automatic metrics.

Keywords: Machine Learning · Deep Learning Schema · Neural Machine Translation · Pairwise Evaluation.

1 Introduction

Recently, studies in Natural Language Proccessing (NLP) have been using neural networks [31,1]. Neural networks have made significant progress in several NLP tasks including MT [20], summarization [7], dialogue generation [21] and image captioning [11]. The evaluation of MT systems is a crucial field of research, as has been highlighted by a number of researchers [34,15,16,3], given that it is used to compare different systems but also to identify a system's weaknesses and help improve it. Various methods have been suggested for the evaluation of MT—both automatic and human [4]. Although, human evaluation is considered to be the best indicator of a system's quality, it is an expensive and time-consuming process, so it cannot be readily used for the development of the

Translation and Interpreting Technology Online, pages 37–47,
July 05-07, 2021.
https://doi.org/10.26615/978-954-452-071-7_005

MT system. As a result, MT researchers and developers mostly use automatic evaluation metrics which constitute an acceptable estimation quality and they are easy and cheap to compute. Some of them rely on score-based metrics, such as Bilingual Evaluation Understudy (BLEU) [26], National Institute of Standards and Technology (NIST) [12] and Word Error Rate (WER) [30], metrics using external resources, like METEOR [10], and neural metrics such as ReVal [17] and Regressor Using Sentence Embeddings (RUSE) [28], while some others use machine learning schemata [13,32,23]. Automatic evaluation methods must be evaluated with specific criteria. According to Banerjee and Lavie [2], a satisfactory automated evaluation system should meet the following conditions: high correlation with human judgments quantified in relation to translation quality, sensitivity to nuances in quality among systems or outputs of the same system in different stages of its development, result consistency, reliability, a great range of fields and speed and usability. The most important condition is considered to be correlation with human judgment [29]. Yet, the automatic evaluation metrics mentioned above have lagged behind, as they do not correlate well with human judgments [27].

This paper seeks to put to the test an evaluation model based on a novel deep learning schema developed by Mouratidis et al. [25] used to compare two NMT systems on four different text genres, i.e. medical, legal, marketing and literary in the English-Greek language pair. The model, NoDeeLe, utilizes information from the source segments, the MT outputs and the reference translation, as well as the automatic metrics BLEU, METEOR and WER.

2 Related Work

Deep Learning (DL) is one of the fastest-growing fields of Information Technology (IT) today being used among others for MT evaluation. Duh [13] decomposes rankings into parallel decisions, with the best translation for each candidate pair predicted, using a ranking-specific feature set, BLEU score and the Support Vector Machine (SVM) classifier. A similar pairwise approach was proposed by Mouratidis and Kermanidis [22], using a random forest (RF) classifier. Cho et al. [6] proposed a score-based schema to learn the translation probability of a source phrase to a target phrase (MT output) with a Recurrent Neural Network (RNN) encoder-decoder. They showed that this learning schema has improved the translation performance. The schema proposed by Sutskever et al. [31] is similar to the work by Cho et al. [6], but Sutskever et al. chose the top 1000 best candidate translations produced by a Statistical Machine Translation (SMT) system with a Long Short-Term Memory (LSTM) sequence-to-sequence model. Wu et al. [32] also trained a deep LSTM network to optimize BLEU scores focusing on German-English and German-French language pairs, but they found that the improvement in BLEU scores did not reflect the human evaluation of translation quality. Mouratidis et al. [24] used LSTM layers in a learning framework for evaluating pairwise MT outputs using vector representations, in order to show that the linguistic features of the source text (ST) can affect MT evaluation.

Gehring et al. [14] proposed an architecture for sequence to sequence modeling based on a Convolutional Neural Network (CNN). The model is equipped with linear units [9] and residual connections [18].

3 Materials and Methods

3.1 Dataset

The STs used in this study are four texts of comparable complexity, i.e. with a Lexile score between 1210 and 1400, belonging to different genres: medical (*T1*), legal (*T2*), literary (*T3*) and marketing (*T4*). All texts were originally written in English. The medical text is a 382-word excerpt from a clinical trial retrieved from the National Center for Biotechnology Information, the legal text is a 367-word excerpt from a purchase agreement, the literary text is a 365-word excerpt from the book *The English* by Jeremy Paxman, while the marketing text is a 410-word excerpt about the Venice Simplon-Orient-Express holidays retrieved from the website of luxury travel tour operator The Luxury Holiday Company. The Lexile score was calculated on the basis of the Lexile Analyzer[1] which relies on an algorithm to evaluate the reading demand –or text complexity– of books, articles, and other materials. In particular, it measures the complexity of the text by breaking down the entire piece and studying its characteristics, such as sentence length and word frequency, which represent the syntactic and semantic challenges that the text presents to a reader.

The STs were machine-translated without any pre-editing and the NMT systems used to produce the raw MT output were DeepL[2] and Google Translate[3] (output obtained June 2, 2021). Google Translate and DeepL are both generic NMT systems that use state-of-the-art AI to translate texts from one language into another. However, these systems differ in the technology they use and the language data they are trained on. More specifically, DeepL uses CNNs and is trained on the Linguee bilingual corpora database, while Google Translate, uses RNNs and is trained on various digital resources in many languages [35].

The reference translations, i.e. the gold-standard human translations, were produced by highly experienced professional translators. In particular, the medical text was translated by a professional translator specialising in the Life Sciences with over 15 years of experience, the legal text was translated by a professional translator and Law graduate with over 10 years of translation experience, while the literary and marketing texts were translated by a professional translator specialising in creative genres and having more than 20 years of experience.

3.2 The Feature Set Used

Two different features categories were employed from source segments, MT outputs and reference translation.

[1] https://lexile.com/
[2] https://www.deepl.com/translator
[3] https://translate.google.com/

The first one derives from string-based linguistic features and the second one from MT evaluation automatic metrics. The first category contains i. string-based similarity features (such as length in words and characters, the longest word length, some ratios e.g. the ratio between lengths in words in the source segments and the two MT outputs, the ratio between longest words from source segments and the two MT outputs and reference translation, etc., the percentage of segments similarity, suffix similarity etc.) and ii. noise features (such as repeated words or special characters). All the features were calculated for the two MT outputs, the source segments and the reference translation. More details on the feature set used can be found in Mouratidis et al. [24]. The second category contains the BLEU score, METEOR and WER.

3.3 Word Embeddings

Word embeddings helped us to model the relations between the two MT outputs and the reference translation. In this paper, the embedding layer, the one provided by Keras [19], is used for the two MT outputs and the reference translation. The encoding function applied is the one-hot function. The embedding layer size, in number of nodes, is 16.

3.4 The DL Architecture

The deep learning schema in Figure 1 is used for classification purposes.

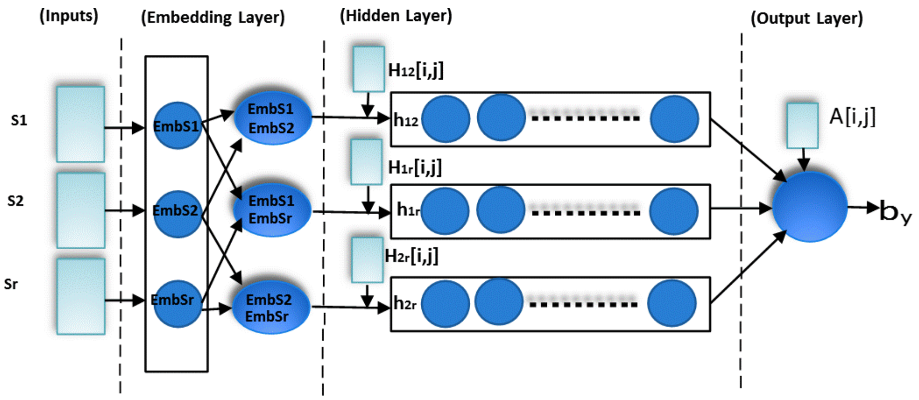


Fig. 1. Deep learning architecture

The input segments in the learning schema are the two MT outputs *S1*, *S2* and the reference translation *Sr*. These segments are converted into numerical vectors (*EmbS1*, *EmbS2*, *EmbSr*) by passing the embedding layer and then they are merged by pair in order to become the input to the hidden layers. In this step, the architecture takes as an extra input the matrices H containing linguistic features from the source segments and the automatic metrics BLEU, METEOR and WER. The architecture takes an extra input to the output layer, the matrix A containing the linguistic features from the MT outputs and the reference translation. Finally, we used as ground truth the ranking information produced by two linguists —both native speakers of Greek, both translators with over 10 years of professional experience each and with a specialisation in MT evaluation/annotation in the English-Greek language pair. The linguists ranked the MT outputs of the four texts at sentence level as follows: 1 if the DeepL MT output is better than the Google MT output, and 0 if the Google MT output is better than the DeepL MT output. The inter-annotator agreement was calculated using Cohen's kappa coefficient (κ) which measures the inter-annotators' reliability; this can take a value between 0 and 1 where 1 indicates perfect agreement and 0 indicates no agreement [8]. In the cases of disagreement between the two annotators, a third, mediating annotator was introduced to resolve the disagreement [33]. The mediating annotator was a professional translator with 15 years of translation experience in the English-Greek language pair and extensive experience in MT evaluation/annotation. The network model architecture for the experiments is a classic architecture of LSTM and feedforward layers.

To avoid over-fitting, a dropout rate of 0.05 is applied, using the binary cross entropy as a loss function and 10-fold Cross Validation. More details about the model's parameters can be found in [25].

4 Results

According to the annotators, and as it emerges from Figure 2, DeepL performed better than Google Translate for all texts. It should be noted that an almost perfect agreement between the annotators for all four texts (κ=0.83 for *T1*, κ=1.0 for *T2*, κ=0.92 for *T3* and κ=0.85 for *T4*) was observed. In the few cases of disagreement, the mediating annotator's decision was used.

Unequal values between the classes were observed with the class belonging to Google Translate being the minority class. The SMOTE supervised filter [5] was applied to the minority class. Figure 3 presents the classification results (classification accuracy) for the two MT outputs over the four different datasets. It emerges that the classification accuracy level is higher in the case of the marketing text followed by the legal and the literary text. The lowest accuracy level is observed in the case of the medical text, most probably due to its rich and highly-specialised terminology. We also applied a SMOTE filter with a view to improving the model accuracy. Indeed, an increase of 2% of the classification accuracy for the medical text, 5% for the legal and the marketing text, and 3% for the literary text is observed. The above accuracy results are in accordance with the annotators' results (see Figure 2). Better accuracy results (*F1 score*) are observed for DeepL (*S1*) compared to Google Translate (*S2*) for all texts (Figure 4).

The BLEU and METEOR scores for the MT outputs of the four texts are given in Figure 5. In particular, the medical text received the highest score

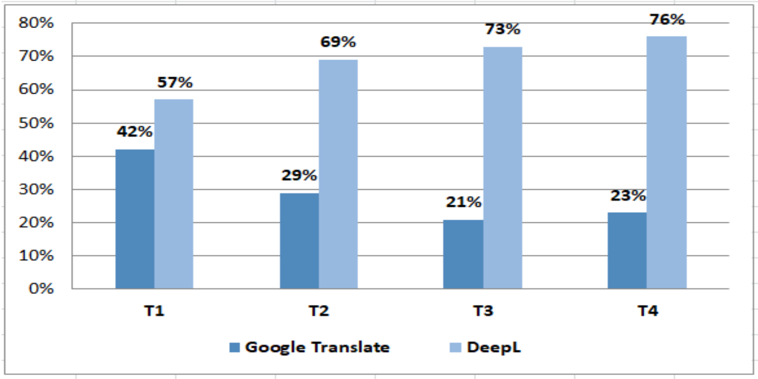


Fig. 2. Ranking information

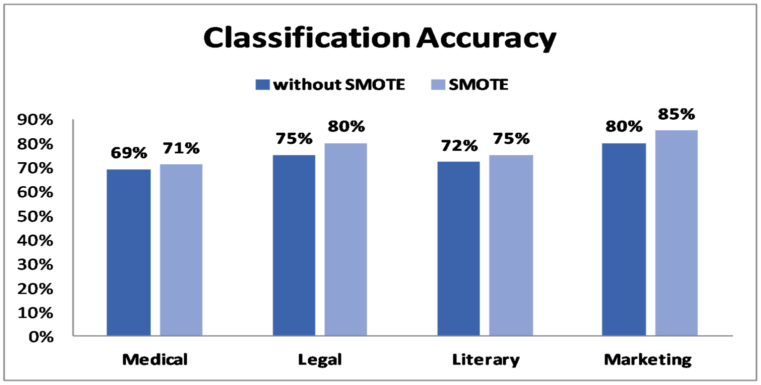


Fig. 3. Accuracy performance with and without SMOTE filter

for both metrics followed by the legal text, while the literary text received the lowest score. Interestingly, the scores are not in line with the results of NoDeeLe, i.e. the proposed deep learning schema, according to which the medical text received the worst classification accuracy and the marketing text received the best classification accuracy. In addition, although DeepL (*S1*) performed better in all cases according to NoDeeLe, Google Translate (*S2*) performed better in the case of the literary and marketing text according to BLEU, and in the case of the medical text according to METEOR. As far as the legal text is concerned, no difference was observed between the automatic metrics and NoDeeLe.

Apart from BLEU and METEOR, NoDeeLe was also compared to other methods. For that reason, additional experiments were carried out using different classifiers e.g. SVM and RF using the WEKA framework as backend. The SVM and the RF classifiers were trained on the same data and feature set as NoDeLee. As depicted in Figure 6, NoDeeLe achieves stronger correlation with the human judgments (78% average accuracy for the four texts), compared to

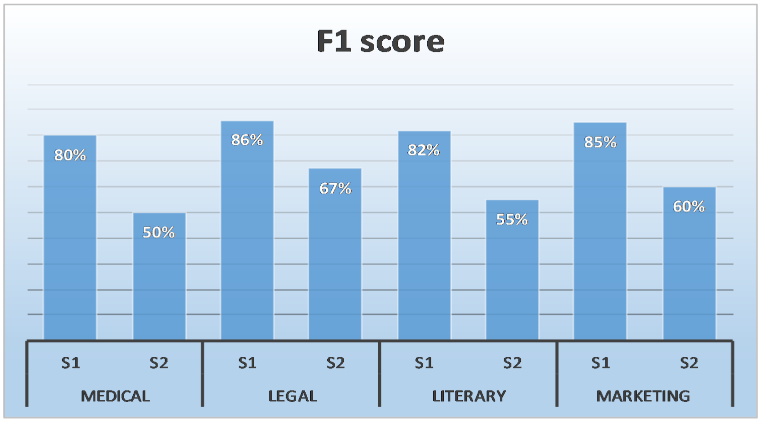


Fig. 4. F1 score per system and per text genre

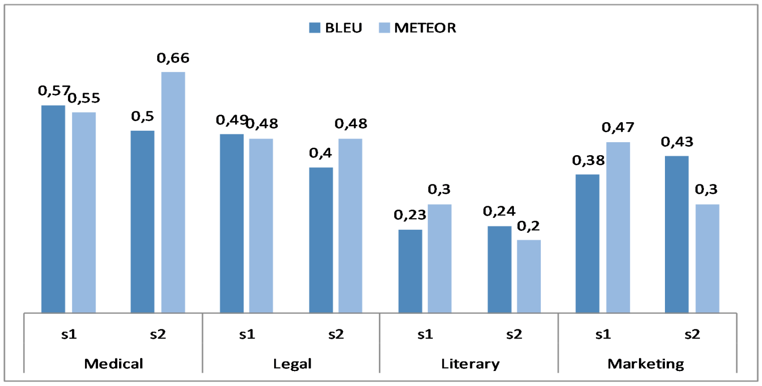


Fig. 5. Automatic Metrics BLEU and METEOR

the RF classifier (74% average accuracy for the four texts) and the SVM classifier (67% average accuracy for the four texts). Unlike BLEU and METEOR scores, NoDeeLe as well as the RF and SVM classifiers indicate that the marketing text has the best classification accuracy, followed by the legal text and the literary text, while the medical text has the worst classification accuracy. In addition, NoDeeLe as well as the RF and SVM classifiers reveal that DeepL (S1) performed better in all text genres in contrast with BLEU and METEOR, with the former showing that Google Translate (*S2*) performed better in the case of the literary and marketing text, and the latter showing that it performed better in the case of the medical text.

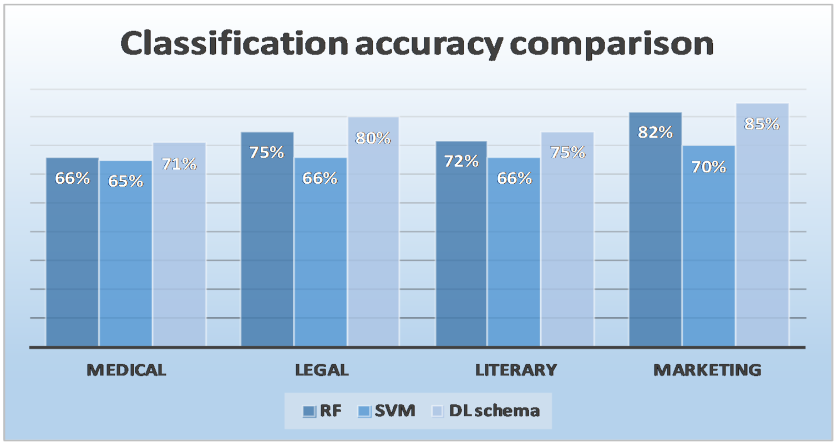


Fig. 6. Classification accuracy comparison with other algorithms

5 Conclusions and Future Work

In this paper, a deep learning novel schema for evaluating NMT systems and outputs is presented and discussed. The schema, i.e. NoDeeLe, used information from the source segments, the MT outputs and the reference translations as well as automatic metrics, and was applied in four different text genres: medical, legal, literary and marketing. Experimental results showed that NoDeeLe achieves stronger correlation with the human judgments compared to the RF classifier and the SVM classifier. Unlike BLEU and METEOR, NoDeeLe, as well as the RF and SVM classifiers, indicate *i.* that the marketing text has the best classification accuracy and the medical text the worst classification accuracy and *ii.* DeepL (*S1*) performed better in all text genres. These findings suggest that the BLEU and METEOR automatic metrics may not be appropriate for the evaluation of NMT output, as has been also indicated by other studies [4].

To complement and expand this study, we aim to explore if pre-trained embeddings e.g. fasttext, could improve classification accuracy, especially concerning texts with specialised terminology. In addition, we are planning to test: i. another neural network structure and ii. a learned evaluation metric, the BLEURT metric [27], on the same datasets. Finally, in order to further explore the observed difference between the BLEU and METEOR automatic metrics and NoDeeLe, we are planning to carry out a refined human error analysis to evaluate the linguistic quality of the MT outputs.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)

2. Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. pp. 65–72. Association for Computational Linguistics, Ann Arbor, Michigan (Jun 2005), https://www.aclweb.org/anthology/W05-0909
3. Bentivogli, L., Cettolo, M., Federico, M., Christian, F.: Machine translation human evaluation: an investigation of evaluation based on post-editing and its relation with direct assessment (2018)
4. Chatzikoumi, E.: How to evaluate machine translation: A review of automated and human metrics. Natural Language Engineering **26**(2), 137–161 (2020)
5. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321–357 (2002)
6. Cho, K., van Merrienboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. In: EMNLP (2014)
7. Chopra, S., Auli, M., Rush, A.M.: Abstractive sentence summarization with attentive recurrent neural networks. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 93–98 (2016)
8. Cohen, J.: A coefficient of agreement for nominal scales. Educational and psychological measurement **20**(1), 37–46 (1960)
9. Dauphin, Y.N., Fan, A., Auli, M., Grangier, D.: Language modeling with gated convolutional networks. In: International conference on machine learning. pp. 933–941. PMLR (2017)
10. Denkowski, M., Lavie, A.: Meteor universal: Language specific translation evaluation for any target language. In: Proceedings of the ninth workshop on statistical machine translation. pp. 376–380 (2014)
11. Devlin, J., Cheng, H., Fang, H., Gupta, S., Deng, L., He, X., Zweig, G., Mitchell, M.: Language models for image captioning: The quirks and what works. In: 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL-IJCNLP 2015. pp. 100–105. Association for Computational Linguistics (ACL) (2015)
12. Doddington, G.: Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In: Proceedings of the second international conference on Human Language Technology Research. pp. 138–145 (2002)
13. Duh, K.: Ranking vs. regression in machine translation evaluation. In: Proceedings of the Third Workshop on Statistical Machine Translation. pp. 191–194 (2008)
14. Gehring, J., Auli, M., Grangier, D., Yarats, D., Dauphin, Y.N.: Convolutional sequence to sequence learning. In: International Conference on Machine Learning. pp. 1243–1252. PMLR (2017)
15. Giménez, J., i Villodre, L.M.: Asiya: An open toolkit for automatic machine translation (meta-)evaluation. In: Prague Bull. Math. Linguistics (2010)
16. GRAHAM, Y., BALDWIN, T., MOFFAT, A., ZOBEL, J.: Can machine translation systems be evaluated by the crowd alone. Natural Language Engineering **23**(1), 3–30 (2017). https://doi.org/10.1017/S1351324915000339
17. Gupta, S., Agrawal, A., Gopalakrishnan, K., Narayanan, P.: Deep learning with limited numerical precision. In: International Conference on Machine Learning. pp. 1737–1746 (2015)

18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
19. Keras, K.: Deep learning library for theano and tensorflow. 2015 (2019), https://keras.io/
20. Koehn, P.: Statistical machine translation. Cambridge University Press (2009)
21. Li, J., Monroe, W., Ritter, A., Jurafsky, D., Galley, M., Gao, J.: Deep reinforcement learning for dialogue generation. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 1192–1202 (2016)
22. Mouratidis, D., Kermanidis, K.L.: Automatic selection of parallel data for machine translation. In: IFIP International Conference on Artificial Intelligence Applications and Innovations. pp. 146–156. Springer (2018)
23. Mouratidis, D., Kermanidis, K.L.: Ensemble and deep learning for language-independent automatic selection of parallel data. Algorithms **12**(1), 26 (2019)
24. Mouratidis, D., Kermanidis, K.L., Sosoni, V.: Innovative deep neural network fusion for pairwise translation evaluation. In: IFIP International Conference on Artificial Intelligence Applications and Innovations. pp. 76–87. Springer (2020)
25. Mouratidis, D., Kermanidis, K.L., Sosoni, V.: Innovatively fused deep learning with limited noisy data for evaluating translations from poor into rich morphology. Applied Sciences **11**(2), 639 (2021)
26. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
27. Sellam, T., Das, D., Parikh, A.: Bleurt: Learning robust metrics for text generation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7881–7892 (2020)
28. Shimanaka, H., Kajiwara, T., Komachi, M.: Ruse: Regressor using sentence embeddings for automatic machine translation evaluation. In: Proceedings of the Third Conference on Machine Translation: Shared Task Papers. pp. 751–758 (2018)
29. Specia, L., Raj, D., Turchi, M.: Machine translation evaluation versus quality estimation. Machine Translation **24**(1), 39–50 (2010). https://doi.org/10.1007/s10590-010-9077-2, https://doi.org/10.1007/s10590-010-9077-2
30. Su, K.Y., Wu, M.W., Chang, J.S.: A new quantitative quality measure for machine translation systems. In: COLING 1992 Volume 2: The 15th International Conference on Computational Linguistics (1992)
31. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in neural information processing systems. pp. 3104–3112 (2014)
32. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al.: Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144 (2016)
33. Zhang, Z., Chapman, S., Ciravegna, F.: A methodology towards effective and efficient manual document annotation: Addressing annotator discrepancy and annotation quality. In: Proceedings of the 17th International Conference on Knowledge Engineering and Management by the Masses. p. 301–315. EKAW'10, Springer-Verlag, Berlin, Heidelberg (2010)
34. Zhou, M., Wang, B., Liu, S., Li, M., Zhang, D., Zhao, T.: Diagnostic evaluation of machine translation systems using automatically constructed linguistic checkpoints. In: Proceedings of the 22nd International Conference on Computational

Linguistics (Coling 2008). pp. 1121–1128. Coling 2008 Organizing Committee, Manchester, UK (Aug 2008), https://www.aclweb.org/anthology/C08-1141

35. Ziganshina, L.E., Yudina, E.V., Gabdrakhmanov, A.I., Ried, J.: Assessing human post-editing efforts to compare the performance of three machine translation engines for english to russian translation of cochrane plain language health information: Results of a randomised comparison. In: Informatics. vol. 8, p. 9. Multidisciplinary Digital Publishing Institute (2021)

BLEU, METEOR, BERTScore: Evaluation of Metrics Performance in Assessing Critical Translation Errors in Sentiment-oriented Text

Hadeel Saadany[1] and Constantin Orăsan[2]

[1] University of Wolverhampton, UK, h.a.saadany@wlv.ac.uk
[2] University of Surrey, UK, c.orasan@surrey.ac.uk

Abstract. Social media companies as well as authorities make extensive use of artificial intelligence (AI) tools to monitor postings of hate speech, celebrations of violence or profanity. Since AI software requires massive volumes of data to train computers, Machine Translation (MT) of the online content is commonly used to process posts written in several languages and hence augment the data needed for training. However, MT mistakes are a regular occurrence when translating sentiment-oriented user-generated content (UGC), especially when a low-resource language is involved. The adequacy of the whole process relies on the assumption that the evaluation metrics used give a reliable indication of the quality of the translation. In this paper, we assess the ability of automatic quality metrics to detect critical machine translation errors which can cause serious misunderstanding of the affect message. We compare the performance of three canonical metrics on meaningless translations where the semantic content is seriously impaired as compared to meaningful translations with a critical error which exclusively distorts the sentiment of the source text. We conclude that there is a need for fine-tuning of automatic metrics to make them more robust in detecting sentiment critical errors.

Keywords: Automatic Metric · Critical Error · Sentiment Evaluation

1 Introduction

Facebook has once apologised after its machine-translation service lead to an arrest of a man from the West Bank whose profile posting in his native dialect that read "good morning" was mistranslated as "attack them", and later automatically detected by authorities as an incitement to violence[3]. The main danger in this type of MT error is that it changes the author's sentiment, here from positive to a negative or rather aggressive emotion. Research on translation of sentiment by MT systems has shown that users encounter similar mistakes where the sentiment polarity of the source is flipped to its exact opposite due to

[3] https://www.theguardian.com/technology/2017/oct/24/facebook-palestine-israel-translates-good-morning-attack-them-arrest

Translation and Interpreting Technology Online, pages 48–56,
July 05-07, 2021.
https://doi.org/10.26615/978-954-452-071-7_006

a mistranslation of a contronym, a dialectical expression, or a missed negation marker, especially in translation of online content of low-resource languages [17]. In machine translation research, the reliability of MT systems is conventionally measured by automatic quality metrics such as BLEU [13] and METEOR [1]. The aim of these automatic quality metrics is to evaluate a translation hypothesis (i.e. the automatic translation) against a reference translation, which is normally produced by a human translator. Good evaluation metrics should have a high correlation with human judgement on the quality of translation. Recently some automatic metrics have achieved a significant correlation with human judgement on the WMT Metrics task datasets (see [7,8,12]). However, research has reported weaker correlation with low human assessment score ranges for segment-level evaluation [20,19]. These findings point to the challenges involved in detecting low-quality translations by automatic metrics.

In this work, we focus on the problem of evaluating critical translation errors that can cause serious misunderstanding of the sentiment conveyed in the source text. To illustrate this point, suppose we are evaluating the MT output "*People are dead, starving in your presence, may God forgive you*" with its reference "*People are dead, starving in your presence, may God **not** forgive you*"[4]. The error in the MT output is only the missing of the word *not*, however, this omission causes the translation to convey the exact opposite sentiment of the source. We argue that such translation errors should be considered more critical than those which produce ungrammatical or low-quality translations, but do not significantly distort the message of the source. However, as we show in this paper, automatic quality metrics fail to give a penalty to this type of critical error proportional to its gravity and may equate this hypothesis with another that also has a uni-gram mistake, but transfers the affect message (e.g *People are dead, **hungry** in your presence, may God not forgive you*).

In this research we conduct an experiment with three canonical automatic quality metrics, BLEU, METEOR and BERTScore [22]. We measure the ability of each metric to penalise sentiment critical errors that severely distort the affect message as compared to translations which correctly transfer the correct sentiment as well as mistranslations that produce incomprehensible content in the target language. We first briefly present the three metrics in section 2. Then, in section 3, we explain our experiment and summarise the results. In section 4, we give our concluding remarks.

2 Related Work

The standard metric for assessing empirical improvement of MT systems is BLEU. Simply stated, the objective of BLEU is to compare n-grams of the candidate translation with n-grams of the reference translation and count the number of matches; the more the matches, the better the candidate translation.

[4] The hypothesis is the mistranslation of Twitter's Translate tab for an Arabic tweet https://twitter.com/ZPNyOawCRVTNBxu/status/878496659793170432, accessed 26 June 2021.

The final score is a modified n-gram precision multiplied by a brevity penalty to account for both frequency and adequacy. Due to its restrictive exact matching to the reference, BLEU does not accommodate for importance n-gram weighting which may be essential in assessing a sentiment-critical error. However, despite research evidence of its analytical limitations [9,16], BLEU, is still the *de facto* standard for MT performance evaluation because it is easy to calculate regardless of the languages involved. METEOR, on the other hand, incorporates semantic information as it evaluates translation by calculating either exact match, stem match, or synonymy match. For synonym matching, it utilises WordNet synsets [14]. More recent versions (METEOR 1.5 and METEOR++2.0) apply importance weighting by giving smaller weight to function words [3,6]. However, the METEOR weighting scheme would not allow for a great penalty of the missing negation marker in the hypothesis of our example above. In fact, the METEOR score for Twitter's MT wrong translation is 0.91, whereas the score for the correct translation (*People are dead, starving in your presence, may God* ***not*** *forgive you*) is 0.99. The main culprit for this proportionally inaccurate scoring is the function word weighting which causes the metric to be over permissive despite the MT engine missing of a negation marker crucial to the sentiment of the source tweet.

Both METEOR and BLEU assess the quality of translation in terms of surface n-gram matching between the MT output and a human reference(s). After the introduction of pretrained contextual word models, there has been a recent trend to use large-scale models like BERT [4] for MT evaluation to incorporate semantic contextual information of tokens in comparing translation and reference segments. A number of embedding-based metrics has proven to achieve the highest performance in recent WMT shared tasks for quality metrics (e.g. [7,8,12]). We take BERTScore as representative of this category. BERTScore computes a score based on a pair wise cosine similarity between the BERT contextual embeddings of the individual tokens for the hypothesis and the reference. Accordingly, a BERTScore close to 1 indicates proximity in vector space and hence a good translation. In the following section, we explain our experiment for assessing the performance of these three metrics with respect to critical translation errors that seriously distort the affect message of the source.

3 Experiment Set Up

3.1 Dataset Compiling

We measure the performance of the three metrics on two types of translated UGC data: synthetic and authentic. The synthetic dataset consists of 100 restaurant reviews extracted from the SemEval-2016 Aspect-Based Sentiment Analysis task where each review expresses mixed sentiment about a particular entity [15]. For this dataset we did not use machine translation, but we artificially modified the original texts in such a way that the original sentiment was distorted. Thus, we created hypothesis-reference pairs with changes only in sentiment-related words. The main objective of the synthetic data is to measure the sensitivity of each metric to sentiment-critical translation errors by making n-gram sentiment modifications to the hypothesis while keeping the other words intact. We made four types of sentiment modifications manually. For example, for the source review '*But the staff was so horrible to us*', we made the following modifications:

- One Non-Critical Error: a uni-gram change that does not affect the sentiment ('*But the staff was so horrible to* ***him***')
- One Critical Error: a uni-gram change that produced the opposite sentiment ('*But the staff was so* ***nice*** *to us*')
- Two Errors: a two-words change with one critical and one non-critical error ('*But the staff was so* ***nice*** *to* ***him***')
- Nonsense: a three-words change that produced a meaningless translation ('*But the team was so to him*')

Table 1: Distribution of Translation of Sentiment Errors for the Datasets

Dataset	**No Error**	**One Error**	**Two Errors**	**Nonsense**
Synthetic En to En		200	100	100
Total	**400**			
Authentic En to Sp/Ar/Pt/Ro	854	404	142	
Authentic Sp/Ar to En	150	150		
Total	**1700**			

The authentic dataset consisted of 1700 tweets collated from different emotion-detection and aggression-detection shared tasks ([11,10,2,21]). The source tweets were in three languages: English (1400), Arabic (200) and Spanish (100). This dataset was translated by Twitter's MT system (Google API). The Spanish and English source tweets were translated into English, and the English tweets were translated into Romanian, Arabic, Spanish and Portuguese. Five human annotators [5], native speakers of the respective languages, manually annotated the translations for sentiment errors. The annotation was straightforward: *Yes* the translation transfers the sentiment of the source (even though it can have non-sentiment related errors that do not seriously affect the overall sentiment/emotion) or *No*, it does not. If 'No', the annotators were asked to mark whether the mistranslation of sentiment is due to one or two linguistic errors. The linguistic error was either a missing negation marker, a mistranslation of a hashtag, an idiomatic expression or a polysemous word (table 1 shows the distribution of the datasets types used in the experiment). More details on how the errors were identified are discussed in [18].

We ran the three metrics on the hypothesis/reference pairs of the synthetic dataset and the hypothesis/reference[6] for Arabic and Spanish tweets, and the source/back-translations of the English tweets of the authentic dataset (The

[5] The annotators were computational linguists working on MT research.

[6] Reference translations were created by the two annotators native speakers of Arabic and Spanish.

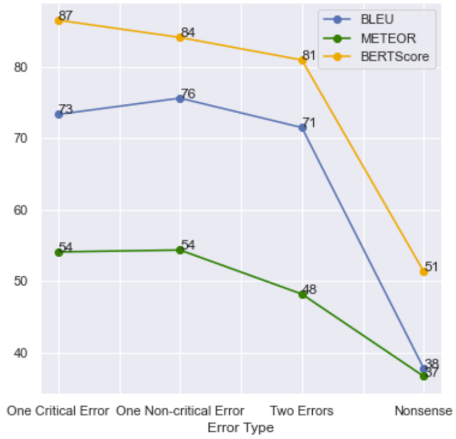


Fig. 1: Mean Scores for Synthetic Data[7]

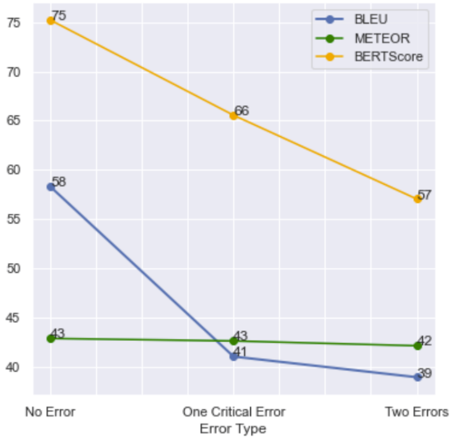


Fig. 2: Mean Scores for Authentic Data (en)

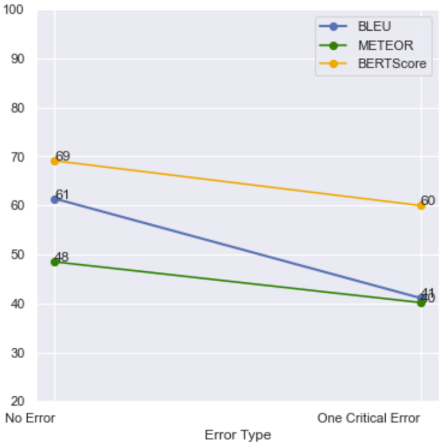


Fig. 3: Mean Scores for Authentic Data (ar/sp)

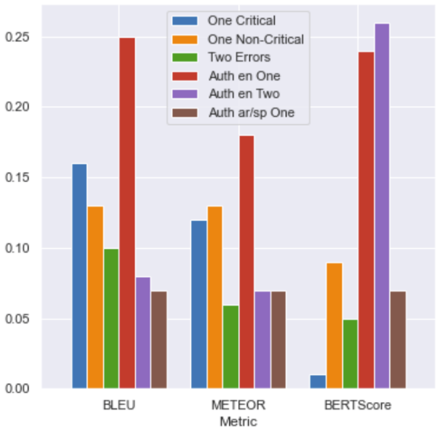


Fig. 4: Normalised Standard Deviation

back-translations were checked to make sure they reproduced the exact sentiment errors in the MT output). Accordingly, as shown in table 1, we evaluated 400 synthetic English hypothesis/reference pairs, 1400 English tweets translated into Romanian, Arabic, Spanish and Portuguese, and 300 Arabic and Spanish tweets translated into English. We used these datasets to calculate three measures for BLEU, METEOR and BERTScore: segment-level scores, mean segment-level scores and standard deviation for segment-level scores. Results of the experiment are explained in the next section.

[7] Scores in figures are standardised from 0 to 100 for easier display.

3.2 Results

The average segment scores of the three metrics for the four sentiment modifications we have conducted on the hypotheses of the synthetic dataset is shown in figure 1. As can be seen from the figure, the difference between the mean score for one critical error and one non-critical error is quite small for all the three metrics (max 3 points difference). This result essentially highlights the inability of the three metrics to distinguish between the mistranslation of a critical word that seriously distorts the affect message and the mistranslation of a non-critical word that does not affect the sentiment content (see table 2 for examples of such cases). The metrics, however, are able to distinguish low-quality translation with a highly distorted content as the average scores for the 'Nonsense' translations are far off from the other types of errors. Furthermore, the average BLEU score for one non-critical error is slightly higher than the one critical error. This is due to the fact that BLEU gauges the performance of an MT model by an indiscriminate n-gram matching, regardless of the semantic weight of each word. An error with a sentiment-critical word, therefore, is equally penalised as any other word. Also, for BERTScore the average score for one critical error is relatively high (0.85) due to what is known as the antonymy problem in contextual word embeddings [5]. Antonyms (e.g. 'great' and 'terrible') usually have similar contextual information and hence are closer in vector space. The change of one word to its exact opposite, therefore, is not adequately captured by the BERTScore metric. It can be claimed, therefore, that the embedding-based metric would generally struggle with hypotheses with only a uni-gram sentiment-critical error that flips the source sentiment to its opposite polarity.

Figure 2 shows a similar problem for the authentic English data. For METEOR, a translation that transfers the affect message has a similar average score as translations that have one or two linguistic errors that seriously distort the sentiment of the source. Note that in the authentic 'No Error' dataset, the hypothesis correctly transfers the main content but may have non-sentiment errors and hence METEOR scores may be lower for some hypotheses. However, the METEOR performance casts doubt on its ability to distinguish between a translation that can transmit the sentiment content despite other errors and another translation that has a critical error of the sentiment which would be unacceptable by human standards. By contrast, the average scores of the BERTScore metric correlate consistently with the degradation of the sentiment transfer in this authentic dataset. However, for the second language arc where Arabic/Spanish are the source languages, the difference between METEOR and BERTScore average scores for segments with no sentiment error and those with critical errors is relatively small (7 and 8 points, respectively as shown in figure3).

Finally, figure 4 shows the normalised standard deviation of the segment-level scores for the three metrics on the different datasets. The scores of the three metrics display the highest variation with the authentic dataset with one sentiment error and BERTScore displays a great variance with two sentiment errors in the same dataset. This indicates that translations with sentiment critical errors do not consistently receive low scores by the three metrics. Similarly, both the ME-

TEOR and BLEU metrics have a relatively higher deviation in segment-level scores for the synthetic dataset with one critical error. Therefore, hypotheses that are exact match to the reference but have only one critical error causing a misinterpretation of the affect message are not consistently penalised by the two metrics (see table 2 for examples of metric scores for references/hypotheses of the two datasets).

Table 2: Examples of Metric Scores for Different Error Types

Synthetic Data		Metric		
		BLEU	METEOR	BERTScore
Ref	Their pizza is the best, if you like thin crusted pizza.	1.0	1.0	1.0
Non-critical Error	Their pizza is the best, if you like thin ***layer*** pizza.	0.76	0.50	0.90
Critical Error	Their pizza is the ***worst***, if you like thin crusted pizza.	0.73	0.50	0.86
Authentic Data				
Ref	What is this amount of happiness, I don't understand!	1.0	1.0	1.0
One Error	What is this amount of ***anger***, I don't get it!	0.65	0.47	0.89
Ref	Sweetie like clouds, always fill me with joy.	1.0	1.0	1.0
No Error	***My love*** is like clouds, always fill me with joy.	0.65	0.44	0.52

4 Conclusion

In this research, we conducted an experiment with three canonical automatic quality metrics to evaluate their ability to penalise a critical translation error that seriously distorts the affect message of the source text. The average segment-level scores for the three metrics showed that sentiment-critical and non-critical errors are not appropriately distinguishable especially in our synthetic dataset. This shows that in scenarios where the MT output is an exact match to the reference except for one sentiment-pivotal word, the automatic quality metric becomes less sensitive to the mistranslation error. Similarly, with the authentic datasets, the average scores for METEOR showed that mistranslations with one or two critical errors are not appropriately penalised. Moreover, with both the authentic and synthetic data, the relatively high inconsistency of segment-level scores for hypotheses with one or two sentiment-critical errors suggests that a distortion of the sentiment content may misleadingly receive high scores by any of the three metrics. The results of the experiment call attention to the need for a sentiment-targeted evaluation measure that can adequately assess this type of critical translation errors that have can serious consequences in determining the sentiment stance of the author. Our future work will focus on fine-tuning the quality metrics to capture sentiment-critical lexicon to improve its performance with sentiment-oriented text.

Acknowledgements

Part of the research done by Hadeel Saadany was carried out in the context of the TranSent project at the University of Surrey.

References

1. Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)
2. Basile, V., Bosco, C., Fersini, E., Debora, N., Patti, V., Pardo, F.M.R., Rosso, P., Sanguinetti, M., et al.: Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In: 13th International Workshop on Semantic Evaluation. pp. 54–63. Association for Computational Linguistics (2019)
3. Denkowski, M., Lavie, A.: Meteor universal: Language specific translation evaluation for any target language. In: Proceedings of the ninth workshop on statistical machine translation. pp. 376–380 (2014)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
5. Etcheverry, M., Wonsever, D.: Unraveling antonym's word vectors through a siamese-like network. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 3297–3307 (2019)
6. Guo, Y., Hu, J.: Meteor++ 2.0: Adopt syntactic level paraphrase knowledge into machine translation evaluation. In: Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1). pp. 501–506 (2019)
7. Kepler, F., Trénous, J., Treviso, M., Vera, M., Martins, A.F.: Openkiwi: An open source framework for quality estimation. arXiv preprint arXiv:1902.08646 (2019)
8. Lo, C.k.: Extended study on using pretrained language models and YiSi-1 for machine translation evaluation. In: Proceedings of the Fifth Conference on Machine Translation. pp. 895–902 (2020)
9. Mathur, N., Baldwin, T., Cohn, T.: Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics. arXiv preprint arXiv:2006.06264 (2020)
10. Mohammad, S., Kiritchenko, S.: Understanding emotions: A dataset of tweets to study interactions between affect categories. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) (2018)
11. Mohammad, S.M., Bravo-Marquez, F.: Wassa-2017 shared task on emotion intensity. arXiv preprint arXiv:1708.03700 (2017)
12. Mukherjee, A., Ala, H., Shrivastava, M., Sharma, D.M.: Mee: An automatic metric for evaluation using embeddings for machine translation. In: 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA). pp. 292–299. IEEE (2020)
13. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
14. Pedersen, T., Patwardhan, S., Michelizzi, J., et al.: Wordnet:: Similarity-measuring the relatedness of concepts. In: AAAI. vol. 4, pp. 25–29 (2004)

15. Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., Al-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., et al.: Semeval-2016 task 5: Aspect based sentiment analysis. In: International workshop on semantic evaluation. pp. 19–30 (2016)
16. Reiter, E.: A Structured Review of the Validity of BLEU. Computational Linguistics **44**(3), 393–401 (2018)
17. Saadany, H., Orasan, C.: Is it great or terrible? preserving sentiment in neural machine translation of arabic reviews. In: Proceedings of the Fifth Arabic Natural Language Processing Workshop. pp. 24–37 (2020)
18. Saadany, H., Orasan, C., Quintana, R.C., do Carmo, F., Zilio, L.: Challenges in Translation of Emotions in Multilingual User-Generated Content: Twitter as a Case Study. arXiv preprint arXiv:2106.10719 (2021)
19. Sudoh, K., Takahashi, K., Nakamura, S.: Is this translation error critical?: Classification-based human and automatic machine translation evaluation focusing on critical errors. In: Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval). pp. 46–55 (2021)
20. Takahashi, K., Sudoh, K., Nakamura, S.: Automatic machine translation evaluation using source language inputs and cross-lingual language model. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 3553–3558 (2020)
21. Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., Çöltekin, Ç.: Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). arXiv preprint arXiv:2006.07235 (2020)
22. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with Bert. arXiv preprint arXiv:1904.09675 (2019)

Remote Interpreting: Platform Testing in a University Setting

Francesco Saina[1]

[1] SSML Carlo Bo, Rome and Bari, Italy
f.saina@ssmlcarlobo.it

Abstract. This work is based on the testing of a remote interpreting (RI) delivery platform conducted a year before the disruptive COVID-19 pandemic outbreak, and aimed at assessing the use and experience of such systems in a university setting. A survey was administered to the different groups of users (interpreters, audience, and speakers) involved in two tests to collect their responses and remarks, and assess trends and perceptions in their experience. According to the findings of the research project, the RI environment was already considered to be an indisputable yet burgeoning resource for conference settings with potential convenience and benefits for each group of users. However, participants' remarks early suggested that all the parties involved in the industry need to collaborate to effectively improve and enhance such services. Specific training on RI modalities would also appear to be increasingly necessary for interpreters to adapt to emerging working conditions and meet a thriving demand —and training institutions would ever more have to offer adequate solutions, while this technological shift also requires receptiveness and adaptability to an abruptly diversifying and evolving profession.

Keywords: Interpreting Technology, Distance Interpreting, Remote Interpreting.

1 Introduction

As in almost any professional field and communication setting, technology has taken a leading role in interpreting too. Over the last few years, remote interpreting (RI) specifically has become an ever-increasing modality being used in conference interpreting [14]—and in the time of global movement restrictions imposed by the COVID-19 pandemic, even the only modality enabling working continuity for professionals worldwide.

This work tests an RI delivery platform and aims at assessing the use and experience of such systems in a university setting—before their widespread use for teaching purposes following the Coronavirus outbreak.

After using the RI platform for two tests held at UNINT University in Rome, Italy, in April 2019, a survey was administered to the different groups of users involved in the proceedings (interpreters, audience, and speakers) to collect their responses and

Translation and Interpreting Technology Online, pages 57–67,
July 05-07, 2021.
https://doi.org/10.26615/978-954-452-071-7_007

remarks, and thus assess their experience with the use of such tools in an academic environment.

1.1 Research Methodology

The events took place in the 'Aula Magna' at UNINT University, where speakers and the audience gathered. A few participants followed the events from remote locations. A remote interpreting 'hub' was located in a classroom on the upper floor of the building, equipped with standard interpreting booths.

Interpretation for the events was provided by teams of volunteer interpreting students in the last year of their Master's Degree, with different levels of previous working experience. Only a limited number of students interpreted for both conferences.

After each of the two conferences, all the participants in the events (interpreters, audience, and speakers) were given a printed individual survey to fill out (remote listeners were sent a digital copy). Answers were then rigorously converted and entered into a specific digital database for a complete and accurate analysis.

Indeed, the survey was considered to be the most suitable instrument for the scope of this research, as it is a comprehensive 'means for gathering information about the characteristics, actions, or opinions of a large group of people' [15], and in accordance with the ultimate aim of any survey research, that of advancing academic knowledge in a scientific field [12].

Statistical Methodology. The various surveys contained 23 to 28 items, composed of closed-ended questions—either multiple-choice or yes/no. To collect more qualitative findings from each participant in the test and explore both general and individual attitudes, most questions were followed by a blank text box which the respondent could use for writing additional specifications or remarks and in-depth motivations.

A limited number of items invited the respondent to indicate a value on a Likert scale of 1 to 5. A final blank space for any further voluntary observation, comment, or personal impression was offered at the very end of the survey.

Clearly, personal values attributed by individual respondents to given parameters are not absolute and necessarily arbitrary. Therefore, average (or mean) values for all answers were calculated and collected to provide more representative results and compensate for such personal differences in grading.

Of course, it is also necessary to take into consideration the relatively limited size of the sample (i.e. groups of users participating in this research) when evaluating the accuracy of the results presented in this experimental study. A total amount of 98 surveys were administered to the three groups of users (interpreters, audience, and speakers) over the two conferences and 66 were returned.

Intuitively (but also according to the statistical notion of standard error), tests involving a larger sample would offer more representative estimations.

1.2 Survey Sample and Structure

Each group of users participating in the surveys was designed differently.

Interpreters were selected by the university's professors for the first test (a conference specially organized for the purpose of this experiment), while voluntary candidates were involved in the second event (and already planned conference with the use of the platform being proposed after the successful experience of a few days earlier).

Speakers participated in the two conferences following invitations from the organizing teams of each event, whereas audience members were both attendees spontaneously interested in the events' topics and specially invited guests.

According to the purposes of this study, the surveys were structured in three main sections.

The first aimed at collecting background information about the user, such as familiarity with conference interpreting settings, possible previous experience with RI, and self-assessed technological expertise.

The second part consisted of more detailed questions about the use of the platform during the event/test (the device employed to access it, evaluation of image and sound quality, RI-related issues such as concentration, sense of participation, fatigue, use of the features offered by the platform, etc.)

In the third and final section, perceptions, and opinions on RI both in general and in relation to traditional simultaneous interpretation were asked about.

1.3 Participants Data

Basic biographical data was collected from survey participants. However, given the reasonably lower number of speakers involved in the two conferences, they were not asked to indicate any personal information for granting their anonymity.

89% of participants from the remaining two groups examined (interpreters and audience) were female and 11% were male.

As previously mentioned, interpreters were all UNINT students in the final year of their Master's Degree in Conference Interpreting, thus their average (mean) age was 23.56, with standard deviation (the measure of dispersion of the values from the average value, i.e. the mean) 0.84.

Three language teams (English, Italian, and Spanish) with a total amount of 14 interpreters worked during first event, whereas 18 interpreters (offering two additional languages: French and Portuguese) provided their service for the second test. Only six interpreters participated in both events.

All the interpreters taking part in the two events were extremely collaborative and completed the surveys.

An approximate total amount of 80 people attended the two events held at UNINT University where the platform testing was carried out. 58 attendees listened to the interpreting service provided via the RI platform and, at the end of the events, they were handed the printed survey and encouraged to fill it out—while they were also reminded that participation in the survey was completely voluntary.

32 audience members returned the completed survey. Respondent audience age ranged from 17 to 47 years (M = 24.44, SD = 5.72).

The data obtained from the speakers' group is significantly limited as only two complete surveys were returned.

Since most of the speakers were preeminent professors and distinguished academics (alongside a few foreign guests speaking during the second event), it is to be taken into account that, despite their willingness, their activities can often limit their possibilities of fully participating in projects like this test. However, the data collected represents the perceptions and opinions on the RI platform of two qualified speakers and their considerations can be valuable even so to this research.

Since the floor source input (i.e. the presentations delivered by the speakers) was transmitted only from one computer, managed and controlled by UNINT's technical staff, speakers accessed the platform from their devices by using the audience token (access code), therefore in the survey they were mainly asked questions on their listening experience.

Survey and participants data is summarized in Table 1 below.

Table 1. Survey and participants data.

Group	**Administered**	**Returned**
Interpreters	32	32
Audience	58	32
Speakers	8	2
Total	98	66

Gender (interpreters + audience)	**%**	**Age** (group)	**Mean**	**SD**
Female	89	Interpreters	23.56	0.84
Male	11	Audience	24.44	5.72

1.4 Pilot Testing

On March 29, 2019, as suggested by Levy & Lemeshow's indications [6], a pilot test with survey administration to a restricted sample was conducted to validate the survey and assess the intelligibility and accuracy of the questions. Participants involved in this pilot survey were not the same as those who would participate in the actual test.

Two days before the first experiment, a final pilot test aiming at verifying the functioning of the platform and all technical requirements was conducted with all the interpreters who would participate in the research experiment, with the purpose of making them begin familiarizing with the system.

2 Findings

Due to the impossibility of presenting here the full report of the data collected from the surveys administered to the three groups of users, the respondents' most significant answers (and the remarks they added in the available blank spaces) will be analyzed and discussed to observe the scenario offered by the surveys, assess the results of this test, and outline general trends emerging from the users' experience.

2.1 Interpreter Survey Results

Despite interpreters testing the platform were students, survey results show that the majority among them (75%) had already had work experience, some of them (19%) even with RI.

Evaluations on video and audio quality were generally positive, but they show that there is still room for improvement of both, on the one hand, the quality of the video feed that event organizers can provide and, on the other hand, the reliability of the signal that the platform can ensure.

Further remarks expressed by the interpreters in the blank spaces available in the survey suggest that a video-mediated view of the speaker is only perceived as an additional asset when a zoomed-in view—which allows seeing nonverbal communication elements, gestures, and lip movements—is offered, thus providing interpreters with a better view than what they could usually see from an on-site booth.

Since remote interpreting is usually reported to have an impact on the interpreter's sense of fatigue and to produce an increased cognitive effort compared to an assignment in traditional simultaneous interpreting equivalent in time [4, 7, 2], interpreters were suggested to work in 15-minute turns, slightly shorter than average simultaneous interpreting turns—in accordance with recommendations by various guidelines and academic publications on RI [17, 2].

However, most of them seemed not to experience any increased mental and/or physical fatigue (91%) nor distraction (75%) caused by the use of RI, and this could be partly attributed to the fact that, in their training courses, they were already used to practicing with videos and in RI-like conditions.

Indeed, almost all of them (78%) perceived their performance the same as compared to traditional interpreting conditions—although half of them reported the platform itself to hinder their performance to a certain extent.

The notions of 'presence' and sense of participation in the communication event are also frequently mentioned when discussing RI in the academic and professional communities [9, 10, 8, 16].

Nevertheless, 78% of interpreting students—who are possibly more used to interacting and communicating in virtual environments—did not feel alienated due to the fact that they were not located in the main conference room.

A couple of them—among the less experienced—even added in the blank spaces for personal remarks that being in a remote location helped them cope with the stress and pressure caused by an on-site working environment, ultimately benefitting their performance.

Other respondents added that being in the hub created a ‘friendly and cooperative environment’ among the interpreting team, which helped interpreters ‘feel more comfortable and confident’ (comments reported from the interpreters’ surveys).

Furthermore, they did not seem to perceive the video-mediated view of the speaker and the absence of direct visual feedback from the audience to have a significant impact on the quality of their performance.

Almost no interpreter (3%) used the event chat feature offered by the platform, thus proving that not always the availability of more features coincides with a better environment for the interpreter, at least when boothmates are co-located as in this case, since the interpreting task requires extreme concentration and adding further elements could occasionally increase the cognitive load and interfere with the interpreter’s attention and performance.

Besides, the relay feature could appear seemingly more intuitive than its equivalent in traditional consoles at first (since it is also set before the beginning of the event), but most interpreters (67%) did not perceive any remarkable difference in its use.

One respondent suggested creating ‘a keyboard shortcut, e.g., the spacebar’ to make the use of the ‘mute’ microphone feature easier and more direct, without the need to use the mouse cursor.

Another interpreter reported the inconvenience of having to log in again when the web page is refreshed, and having the laptops inside the booth with less space for personal belongings and working materials (e.g., printed documents and glossaries) was mentioned as another element of discomfort.

78% of interpreters accepted favorably a few additional conditions required by the platform (such as the use of personal devices, the possibility of running out of charge, the need to download the mobile app to listen to their colleagues’ performances), without considering that an inconvenience.

Moreover, probably by virtue of their familiarity with information and communication technologies (ICTs), 91% of interpreting students found the use of the platform intuitive and immediate, and therefore did not need any additional informative material besides a guide they were provided with by the platform company and a whole morning to test the platform two days before the experiment.

A few of them (22%) complained about not having received presentations and materials shown to the audience by the speakers beforehand, since these are not always sharply visible in an RI setting, even when a dedicated screen is provided.

However, slightly more than half of them (56%) believe that specific training for learning how to use and work with such tools is necessary, at least (in the interpreters’ own words) ‘a couple of lessons’ should be introduced into regular courses to ‘familiarize with the use of such systems’, and ‘dedicated preparatory sessions’ before any assignment are considered to be ‘essential to verify the functioning of the platform’ and feel ‘confident on the day of the event’.

One respondent suggested the use of video tutorials for learning how to use the platform faster and more immediately.

The need for ‘available adequate equipment’ when practicing with such systems and when preparing for an RI assignment was also highlighted.

Above all, most of them expressed the awareness of the fact that such tools may have a preeminent role in ‘the future of the interpreting profession’ and therefore ‘training courses should take that into consideration’, since being able to master these systems can be ‘a valuable asset on the market’.

The importance of familiarizing with the platform is also supported by the fact that 5 out of the 6 interpreters who participated in both tests found the second experience to be better than the first on the whole—the one remaining considered it ‘the same’.

Table 2 outlines the most significant quantitative data collected from interpreters.

Table 1. Quantitative data from the interpreters’ surveys.

Interpreters	**Yes** (%)	**No** (%)	**N/A** (%)
Previous work/internship interpreting experience	75	25	/
Previous experience with RI	19	81	/
Increased mental and/or physical fatigue	9	91	/
Increased distraction	22	75	3
Platform-related obstacles	50	47	3
Feeling part of the event	78	9	13
Performance affected by lack of direct visual feedback	3	94	3
Use of the event chat	3	97	/
Platform conditions as inconvenience	22	78	/
Platform is easy-to-use, intuitive, and immediate	91	9	/
Need more materials	22	78	/
Specific training for RI	56	44	/

2.2 Audience and Speaker Survey Results

Moving the discussion to the data collected from the audience, in the first place it must be underlined that most attendees listening to the interpreting service were other university students in conference interpreting, alongside professors, researchers and professional interpreters of UNINT’s academic community, and a few participants from other faculties or even institutes.

Listeners followed the event both from the conference room and remote locations, with a couple of the former also moving from the event venue to other positions inside or outside the university while bringing their personal devices with them, and a limited group (13%) participating in both events where the platform was tested.

Half of them had never heard about RI before this test and only very few of them (9%) had participated in other events where RI services were provided. This indicates that even in events attended by the interpreting community, RI had not become a prevalent solution yet, but most of them (60%) only rarely take part in events where interpretation services of any kind are offered—or at least they do not need interpretation during the events they usually attend.

62% of them accepted downloading the app on their mobile devices to listen to the interpretation service, while those who accessed the platform via its webpage without downloading any software could also follow the video of the event.

Overall evaluations on video and audio quality (and synchronicity between what was happening in the event venue and the transmission of the signal) from the audience were positive again, but encouraged improvements to definitely make the RI platform an optimal solution. Including the video feed and the event chat on the mobile app too was mentioned by a few audience members as a recommend upgrade.

A considerable majority of respondents did not experience any increased fatigue or stress (88%), distraction (81%), and obstacle (72%) due to the platform, thus indicating a general positive experience for the users.

Saving time usually dedicated to the distribution and return of the receiving devices and headphones was also remarked by three respondents as an advantage of being able to use personal devices. Additional remarks praised both app and web user-friendly interfaces and the easy selection of language channels.

One of the remote listeners defined 'being able to follow the conference from a distant location while listening to the interpretation service in different languages' as 'revolutionary and simple at the same time'. Two different remote listeners remarked increased distraction since they were following the event from home instead.

Three respondents praised the possibility to continue listening to the audio feed also after exiting the app, and therefore being able to simultaneously use their mobile devices for separate needs. Nevertheless, two different respondents mentioned this same possibility as leading them to distraction.

Dissenting respondents agreed on some complaints about the platform, i.e., rapidly running out of charge on their personal devices (claimed by eight respondent audience members and also mentioned by a few interpreters listening to their colleagues via their mobile devices), and listening to the audio signal from their device speakers even after disconnecting their headphones (reported twice as an inconvenience). One respondent only also signaled slowness when changing language channels.

Additionally, features like a general event chat appear not to make a significant difference in the user experience, unless specific and individual chat options are offered.

The main quantitative data obtained from attendees is displayed in Table 3 below.

Table 1. Quantitative data from audience surveys.

Audience	**Yes** (%)	**No** (%)	**N/A** (%)
Awareness of RI before the test	50	47	3
Participation in events with RI	9	91	/
Increased fatigue or stress	12	88	/
Increased distraction	19	81	/
Platform-related obstacles	25	72	3
Use of the event chat	6	94	/
Platform conditions as inconvenience	25	69	6
Platform is easy-to-use, intuitive, and immediate	91	6	3

Finally the two speakers, albeit indicating a high degree of familiarity with ICTs, were at their first experience with an RI platform.

Members of this users group expressed moderate positive evaluations on audio quality and synchronicity and an overall favorable judgment on the platform—still taking into account that they had little direct involvement in the use of the platform as speakers since the technical staff entirely prepared and managed it for them.

However, their experience as users of the platform for listening to the interpreted audio did not significantly hinder their speaking task nor increase their stress or distraction in any reported way.

3 Conclusion

The objective of this research was to investigate the use and observe the experience of a remote simultaneous interpreting platform in a university setting. This study did not aim at evaluating the platform itself nor expressing an ultimate judgment on the implementation of remote interpreting as an accomplished working modality.

The purpose of the project was rather to collect the users' perceptions and remarks and assess trends in their experience, by testing a professional platform on a limited but representative sample of participants, composed of interpreting students, professional and non-specialized audience, and speakers.

3.1 Discussion

Interpreting students testing the platform are clearly short of extensive experience and competency to express a more accurate and comprehensive evaluation on RI advantages and disadvantages, however they also are the future professionals who will encounter these tools in their working environments more than any previous generation, thus collecting their impressions and inclinations towards such systems is a valuable standpoint. Results and tendencies emerging from their answers suggest a remarkably responsive and receptive approach, and openness to innovation and evolution in the profession.

Previous experiments on distance interpreting reported that professional interpreters often find difficulties in embracing RI solutions: since they are used to automated processing when performing their tasks in traditional interpreting conditions, they are therefore hindered when trying to accommodate new variables [8].

Notwithstanding, large-scale medical examinations did not find any evidence of additional stress, and a performance evaluation assessed that remote interpreters' outputs are slightly inferior compared to those of on-site interpreters, yet not enough to achieve statistical relevance [7, 13].

Furthermore, it is not to be forgotten that already in the middle of the twentieth century, most prominent consecutive interpreters refused to adapt to the then recently-born simultaneous modality [5]. As a matter of fact, Moser-Mercer [8] had already perceived that interpreters with years of professional practice 'may be less likely to adapt to a new working environment than less experienced colleagues who may exhibit a greater degree of adaptive expertise'.

Audience and speakers' responses show that RI has achieved offering a satisfactory overall experience by now. Nevertheless, there is still room for maneuver and collaboration among all the parties involved in such evolution appears to be the best way to developing and implementing increasingly better solutions—and RI delivery platforms have regularly and relentlessly been updating their systems during the COVID-19 emergency to keep pace with constantly evolving needs.

The technological development has completely shifted the paradigms of society and work as a whole: professionals in any field are adapting to new working modalities and conditions and, since communication patterns also are considerably evolving, the interpreting profession too will inevitably be involved in such virtual revolution and will diversify accordingly [9, 1, 11, 3].

The adaptability of junior interpreters clearly needs to be associated with rigorous and in-depth preparation. The need for specific training on RI systems during interpreting courses and the availability of adequate equipment in training institutions were both expressed by interpreting students participating in this test, thus providing support to what Ziegler and Gigliobianco [18] had already acutely underlined.

Most attention should be paid to the effect of additional practice and familiarization with the platform on the five interpreters who participated in both events and already considered the second experience better than the first.

3.2 Further Research

The aforementioned outcome motivates the proposition of the need to carry out more than one single experiment with the same participants (and on bigger samples, too) taking into account another parameter: time variation.

This could allow a comparison between subsequent sessions and further examination of new criteria, such as the evolution of trends over time, the consequences of increasing expertise of the users with the platform, or potential changes in the perception of the performance and the overall service.

Moreover, during the tests carried out for this research, same-language colleagues have been working in the same booth. Therefore, an additional challenge to be explored would be turn handover and communication between boothmates when interpreters are not placed in the same location.

The COVID-19 pandemic, besides its dramatic impact on global health, economy, and society, is also an unprecedented challenge in the history of interpreting, questioning several tenets of the profession itself. It will undoubtedly mark RI as one of the main subjects in the interpreting research field over the next years, as it deserves and offers space for extensive supplementary exploration.

The outcome of this work may provide training institutions with insights and indications on how to implement RI tools in their environment after the Coronavirus emergency when shaping the post-pandemic academic scenario, and the considerations in the last paragraphs could pave the way to only a few possible paths for additional research interest and further investigation.

References

1. AIIC (Private Market Sector Standing Committee): Remote simultaneous interpreting: Time to start a dialogue (2019), https://aiic.net/p/8755, last accessed 2019/10/09.
2. Braun, S.: Remote interpreting. In: Mikkelson, H., Jourdenais, R.: The Routledge handbook of interpreting. Routledge, New York, NY (2015).
3. European Commission (Directorate General for Interpretation): Interpreting platforms: Consolidated test results and analysis (2019), https://ec.europa.eu/education/knowledge-centre-interpretation/sites/kci/files/interpreting_platforms_-_consolidated_test_results_and_analysis_-_def.pdf, last accessed 2021/05/25.
4. European Parliament (Interpretation Directorate): Report on remote interpretation test. 22–25 January 2001. Brussels (2001), http://www.europarl.europa.eu/interp/remote_interpreting/ep_report1.pdf, last accessed 2021/05/25.
5. Gaiba, F.: The origins of simultaneous interpretation: The Nuremberg trial. University of Ottawa Press, Ottawa (1998).
6. Levy, P.S., Lemeshow, S.: Sampling of populations: Methods and applications. 4th edn. Wiley, Hoboken, NJ (2008).
7. Moser-Mercer, B.: Remote interpreting: Assessment of human factors and performance parameters (2003), last accessed 2019/10/09.
8. Moser-Mercer, B.: Remote interpreting: The crucial role of presence. Bulletin suisse de linguistique appliquée 81, 73–97 (2005).
9. Mouzourakis, P.: That feeling of being there: Vision and presence in remote interpreting (2003), https://aiic.net/p/1173, last accessed 2019/10/09.
10. Mouzourakis, P.: Remote interpreting: A technical perspective on recent experiments. Interpreting 8(1), 45–66 (2006).
11. Olsen, B.S.: Remote interpreting: Feeling our way into the future. The ATA Chronicle 46(3), 14–16 (2017).
12. Pinsonneault, A., Kraemer, K.L.: Survey research methodology in management information systems: An assessment. Journal of Management Information Systems 10, 75–105 (1993).
13. Roziner, I., Shlesinger, M.: Much ado about something remote: Stress and performance in remote interpreting. Interpreting 12(2), 214–247 (2010).
14. Seeber, K.G.: Distance Interpreting survey: Answers to seven questions (2018), http://members.aiic.net/p/8640, last accessed 2019/10/09.
15. Tanur, J.M.: Advances in methods for large-scale surveys and experiments. In: McAdams, R., Smelser, N.J., Treiman, D.J.: Behavioral and social science research: A national resource (Part II). National Academy Press, Washington, DC (1982).
16. Viaggio, S.: Interpretation: Theory and practice (2019), http://sergioviaggio.com/?p=4330, last accessed 2021/05/25.
17. Wisconsin Court System: Guidelines for using telephonic interpreting in court (2011), https://www.wicourts.gov/services/interpreter/docs/telephoneinterpet.pdf, last accessed 2021/05/25.
18. Ziegler, K., Gigliobianco, S.: Present? Remote? Remotely present! New technological approaches to remote simultaneous conference interpreting. In: Fantinuoli, C. (ed.): Interpreting and technology, pp. 119–139. Language Science Press, Berlin (2018).

Approaching Stress and Performance in RSI: Proposal for Action to Take Back Control

Dora Murgu[1]

[1] Interprefy AG, Bellerivestrasse 11, CH-8008 Zürich, Switzerland
Dora.Murgu@interprefy.com

Abstract. The relationship between stress and performance and Remote Interpreting (RI)/Remote Simultaneous Interpreting (RSI) has been widely studied in academic, professional and corporate research during the past fifty years. Most of such research has attempted to correlate RI/RSI with changes in stress levels and performance, with little to no relevant results to suggest causality. While no significant clinical causality has been found between RI/RSI and stress, self-perceived stress during RI and especially RSI among practicing conference interpreters is consistently high and recent studies suggest a tendency on the increase. Similar results have been observed with performance, which has been and is consistently self-assessed as poorer during RI/RSI by practicing interpreters compared to in-person interpreting, however no significant decrease in performance was observed by independent reviewers. Several scholars have suggested a correlation between such low self-perceived performance / high self-perceived stress and a lack of control which might result from being exposed to unknown factors during RI/RSI, prominently technological elements, the performance of which no longer relies on third parties but lies with the interpreters themselves. This paper is centered on the same hypothesis and suggests a proposal for action that interpreters can undertake to help regain control and thus improve their attitude toward RI/RSI.

Keywords: Remote simultaneous interpreting, performance, stress management, interpreter education, risk management, PMI.

"(remote interpreters) require different problem-solving and capacity management strategies in order to be better prepared to face new situations"
Andres Dörte & Stefanie Falk

1 Stress and Remote Interpreting in Literature

The relationship between stress, performance and remote interpreting was first identified in the 70s in experiments conducted during several meetings held by UN organizations. They all resulted in interpreters complaining of an increase in stress levels. (Dörte & Falk, 2009). The perception of increased stress has been consistently identified in subsequent studies performed by United Nations and European Institutions

Translation and Interpreting Technology Online, pages 68–75,
July 05-07, 2021.
https://doi.org/10.26615/978-954-452-071-7_008

(Mouzourakis, 2006) with interpreters generally concluding that remote interpreting, as compared to on-site interpreting, causes higher levels of stress, fatigue and results in lower performance.

2 The focus on RSI

With the rise of remote simultaneous interpreting delivery platforms in the mid 2010s and especially the hypergrowth of remote simultaneous interpreting assignments in 2020 as a result of the world pandemic, research has shifted from remote interpreting to remote *simultaneous* interpreting to reanalyze its impact on the health, wellbeing and performance of conference interpreters working remotely.

Studies conducted on (RSI) have highlighted, among others, (…) psychological factors, such as fatigue, higher levels of stress and loss of motivation and concentration (Fantinuoli, 2018; Moser-Mercer, 2011). Some of the most recent studies and their most relevant results within the scope of this paper are summarized below.

DG SCIC interpreters who participated in the European Parliament's test of four interpreting platforms in April and May 2019 correlated the high number of errors and issues to a lack of training on platform use before becoming familiar with its layout. Interpreters noted that "getting accustomed to the new tools would certainly decrease stress and fatigue" (European Commission, 2019).

In a survey conducted to 66 interpreting students from the University of International Studies in Rome during April 2019, interpreters complain about the lack of training in RSI and highlight the familiarity with the RSI platform as an essential factor to boost performance (Saina, 2021). There was no mention to stress or fatigue, however it is important to note that interpreting students were co-located in an interpreting hub and supported by a team of technicians.

A survey conducted in 2020 to 27 conference interpreters in Turkey aimed at exploring trends in the perception of remote interpreting revealed that the majority of the interpreters interviewed were comfortable with troubleshooting with internet connection however they were largely undecided when asked about their ability to handle connection problems or other technical problems during an assignment, in line with their claim of not being very knowledgeable in computer hardware and peripherals (Kincal & Ekici, 2020).

A research project on RSI conducted by two researchers at the École supérieure d'interprètes et de traducteurs (ESIT) to 946 professional interpreters between March and April 2021, revealed that 50% of the 857 eligible respondents believe that their performance is worse, compared to on-site interpreting, while 83% consider that RSI is more difficult (Collard & Bujan, 2021).

In a survey conducted by the Canadian Association of Professional Employees to 73[1] professional interpreters registered with the Association 93% have interrupted the interpretation during an ongoing assignment because of poor sound quality, nearly half of which (43%) resumed even if the sound quality issue was not resolved. 87% of

[1] Responses were gathered from 43 participants

respondents experienced high or very high levels of stress (Canadian Association of Professional Employees, 2021).

An article produced by an AIIC interpreter and voice researcher claimed that, compared to the sound produced in traditional conferencing audiovisual systems, the sound delivered by RSI and videoconferencing platforms is of poor quality due to its frequency range being limited to no more than one third of the audible spectrum and noise suppression, feedback cancelling and other algorithms. The author argues that high and very high frequency information, crucial to understanding speech in complex acoustic environments, is suppressed, resulting in overworking of the interpreter's ears and nervous system (Caniato, 2020).

In a focus group conducted by the author with four practicing interpreters in June 2021, the items identified as important factors which make RSI more difficult than in-person interpreting were technical accountability as interpreters are responsible for their own equipment, poor incoming audio and not being co-located to their boothmate.

3 Stress, performance, and control

While none of the recent studies have included a clinical approach in their evaluation of stress and performance during RSI, many of the early studies did indeed focus not only on self-perception but also on medical and physiological examination to determine changes in stress levels as well as independent expert evaluation to measure performance.

Early and recent studies alike resulted in a generally negative and more stressful perception of remote interpreting when compared to in-person interpreting, however studies which included a clinical approach did not find significant changes in stress hormone values nor increases in stress level were observed. With regards to quality, independent reviews found that performance was not negatively affected by RSI despite a generally negative self-perception of performance from the interpreters themselves (European Parliament, 2002; Moser-Mercer, 2003; Roziner & Shlesinger, 2010; Seeber & AIIC, 2018; Fantinuoli, 2019)

Such results and a lack of correlation between RI/RSI and physiological changes lead researchers to assume that the underlying cause of a negative perception is lack of control of the situation during a RI/RSI assignment (Moser-Mercer, 2003; Moser-Mercer, 2005; Mouzourakis, 2006; Roziner & Shlesinger, 2010; Ziegler & Gigliobianco, 2018).

Control is much associated with the ability to anticipate, predict and respond to the unknown and is a known construct to ameliorate stress responses (Steptoe & Poole, 2016) as well as increased optimism which subsequently resulted in active coping (Fontaine, Manstead, & Wagner, 1993). In their review of the concept of psychological stress and in the most relevant literature, in an attempt to find a correlation with RI/RSI, Ricardi et al. note that "uncontrollable or unpredictable events are more stressful than controllable or predictable ones". The authors also identify the concept

4

of "perception of the consequences of failure" as a relevant factor in psychological stress levels (Ricardi, Marinuzzi, & Zecchin, 1998).

It can be assumed, therefore, that actions aimed at increasing control over unpredictable situations that arise during RSI (including, but not limited to, issues with sound, network failure, unscheduled software updates) as well as the unknown associated with the use of technology, in terms of software (including, but not limited to sound control applications, videoconferencing software), hardware (including, but not limited to, peripherals such as headsets, microphones, adapters, cables) and RSI platforms (multiple platforms have their own particular functions, dynamics and interfaces) will increase the interpreters' ability to anticipate, predict and respond to the unknown and therefore will improve their self-perceived stress and performance levels.

4 Proposal for action

In order to define, address and manage unpredictable situations in RSI I propose applying the Project Management Institute (PMI)'s approach to Risk Management as it is consistent with most modern risk management standards (Weaver, 2008; Mulcahy, 2003). Risk is defined by the PMI as "an uncertain event or condition, that if it occurs, has a positive or negative effect on a project's objective" (Project Management Institute, 2017) and is managed following six processes[2]. In Project Management, risk management is aimed at systematically and proactively addressing unknown and unpredictable situations in order to take control of the project (Mulcahy, 2003).

I propose that interpreters apply this approach to manage risks following the six standard processes, with the addition of a seventh process which is aimed at increasing even further the level of control of the risk[3]. By performing the 7-process approach to risk management, interpreters can identify those situations which are most likely to occur, plan an appropriate response and be better prepared if the risk occurs,

[2] For a comprehensive analysis of Risk Management please see Mulcahy (2003).

[3] This assumption is made on the basis of the author's experience as a telephone interpreting Training and Quality Manager for the Spanish LSP Interpret Solutions between 2008 and 2016, where interpreters were subject to unpredictable situations during their onboarding training process. Part of the onboarding process was a series of role-plays where the trainers introduced stress-inducing elements such as background noise, interruption in the call or bad speaker attitude, all of which were moderately frequent occurrences during telephone interpretation calls. When faced with similar situations in real calls, monitored as part of the quality management process of the company, interpreters that had been onboarded using this methodology were able to effectively manage the situation, proved to be resourceful and kept calm. The same role-plays were used as part of telephone interpreting training modules taught at several universities in Spain with significant differences in performance, self-perceived and observed stress and general attitude toward telephone interpreting. While a study was not performed at the time to determine correlation it can be assumed that interpreters who had been subject to unpredictable situations during their onboarding would show an overall better response when they occurred in real calls as opposed to interpreters who had not.

therefore likely experiencing lower levels of psychological stress and drops in performance in the event of a risk taking place.

The six processes as well as the additional seventh process are detailed below and include RSI-specific examples for illustration purposes.

1. Plan Risk Management. Actions include determining categories of risk, analyzing lessons learned (i.e. issues that occurred in past RSI assignments) or determining which stakeholders to involve in risk management (i.e., other interpreters, family members, technical support from the preferred RSI platform who can help identifying risks)
2. Identify Risks. This is the most important process in this approach as it is aimed at recording as many risks as possible, appropriately categorize them, and, where applicable, identify triggers or early warning signs. There are several methods to identify risks such as using a prompt list of standard categories (i.e., network, sound, computer software, RSI platform, personal, environment), cause and effect diagrams, root cause analysis, or interviewing or brainstorming with stakeholders determined in the previous process. Risks are more effectively identified if they clearly state their effect (i.e. the risk of a glass of water spilling over the keyboard will have the effect of damaging the computer during an ongoing RSI assignment). A useful tool for documenting risk is an electronic risk register using a simple spreadsheet software, as illustrated in the sample below:

Table 1. Sample Risk Register

ID	Category	Risk	Effect	Trigger	Probability	Impact

3. Perform Qualitative Risk Analysis. During this process, each risk is evaluated for its *probability* to occur and *impact* on the situation by allocating a subjective numerical probability and impact score, such as 1 to 5 or 1 to 10, 1 being the lowest impact and probability and 5 or 10 the highest. Scoring risks will help identifying the top priority risks that need especial attention, i.e. those with the highest probability of occurring and the highest impact such as a dog barking loudly or a power outage due to temporary construction works in the building.
4. Perform Quantitative Risk Analysis. This process involves assigning objective values to those risks that allow it, by determining their probability in terms of percentage and their impact in terms of monetary value. For example, one interpreter has identified the probability of headset failure as 2% based on the issues occurred in assignments during the past three months and assigned the monetary value of the impact at 175€ which is the cost of purchasing an additional headset to use in case of failure. Many risks cannot have a monetary value added nor can their probability

of occurring be measured precisely, therefore this process is very likely to be applicable to a few risks. Nevertheless, Quantitative Risk Analysis can also be a useful exercise to forecast possible costs and help budgeting, especially after adding up the monetary values of all quantified risks.

5. Plan Risk Response. The goal of this process is to determine actions to reduce the probability and impact of a negative risk and increase the likelihood of a positive risk. In the case of threats, are four response strategies: avoid (eliminate the cause), mitigate (reduce the probability or impact), transfer (assign the risk to someone else, typically by subcontracting an action or buying insurance) or accept (do nothing). The accept strategy is usually allocated to risks that cannot be avoided, transferred or mitigated, such as a natural disaster leading to a power outage. It is important to identify such risks as they can elicit contingency plans (i.e. switch to a battery operated device such as a mobile phone or tablet). In the case of opportunities, or risks with a positive outcome, the four strategies are embrace, enhance, share or accept. For example, outsourcing administrative and invoicing tasks to an accountant would be a sharing response to the opportunity of freeing up time spent on such tasks.
6. Monitor Risks. Once risks are identified and responses are planned, it is important to review the risk register and the planned responses regularly as new situations may lead to new risks or temporary situations have ceased to occur. Personal changes might even lead to including a brand-new category of risks, such as adopting a pet.
7. Practice Risk Response. This process focuses on live testing and practicing the risk response, especially to high probability, high impact risks, insofar as the risk identified allows it, in order to increase familiarity with the risk response and better coping with its application it in real-life situations. For example, one interpreter has identified interruption in the broadband internet connection as a high probability and high impact risk due to some temporary roadworks on her street. Her planned response is sharing the mobile data from her mobile phone and having the phone at hand in case an immediate switch to the phone's connection is needed. By practicing the risk response, she can actually test this situation by performing RSI on a practice platform or speech repository engine and turn off the broadband router at any given moment or, even better, asking a family member to turn off the broadband unexpectedly. This will not only automate the risk response and increase familiarity with the risk but may very likely lead to identifying secondary risks associated to the response which were missed during the planning process, such as having the phone connected to a power source, turning on the do-not-disturb mode to prevent interruptions or ensuring the phone's data plan supports high traffic.

Because risks arise from a varied set of circumstances which are unique to each interpreter's equipment, location, experience with the use of technology and personal circumstances among many other factors, it is important that the 7-process approach is performed individually by each interpreter for the best outcome. An interpreter who is tech-savvy, has three pets and works in an open-plan apartment will have identified different risks from an interpreter who is not familiar with technology, has small children and works from a home office which is independent from the rest of the house.

While the PMI's approach to Risk Management is an industry recognized methodology to reduce unknowns and regain control on projects across any industry or field of specialization and has several associated certifications that are globally recognized, further research to measure self-perceived stress and quality upon the application of the 7-process approach to remote interpreters is encouraged, the results of which would be of great value to an increasingly demanded profession.

References

1. AIIC Taskforce on Distance Interpreting, https://aiic.org/site/TFD0, last accessed 2020/06/03.
2. Braun, S.: Remote interpreting. In Mikkelson, H., Jourdenais, R. (eds) The Routledge Handbook of Interpreting, pp. 352–367. Routledge (2015).
3. Canadian Association of Professional Employees: *Survey of Interpreters Experience with Virtual Sittings of Parliament.* Ottawa, Ontario: ACEP-CAPE (2021).
4. Caniato, A: *The Proposed Pathodynamics of the Junk Sound Syndrome: Why RSI sound is bad for the interpreter's ears (2020, July 14).* Retrieved June 2021, from LinkedIn: https://www.linkedin.com/pulse/proposed-pathodynamics-junk-sound-syndrome-why-rsi-bad-andrea-caniato/
5. Collard, C., & Bujan, M.: Research Project on Remote Simultaneous Interpreting. Paris: ESIT (2021).
6. Congress Rental Australia: The state of Remote Simultaneous Interpretation in 2020-21: an insight into the remote simultaneous interpretation experience for interpreters and how it can be improved. Sidney: Congress Rental Network (2021).
7. Dörte, A., Falk, S.: Download Information and Communication Technologies (ICT) in Interpreting Remote and Telephone Interpreting. Spürst Du wie der Bauch rauf-runter? Fachdolmetschen im Gesundheitsbereich (InterPartes), 9–27 (2009).
8. Dranch, K.: Report for Conference Interpreting after Covid-19 . Prague (2021).
9. European Commission: *Interpreting Platforms Consolidated test results and analysis.* Brussels: European Commission DG Interpretation (2019).
10. European Parliament: *Report on the 2nd EP Remote Interpretation Test.* European Parliament, Working Party on New Technologies of the Interpretation Directorate, Brussels (2002).
11. Fantinuoli, C.: Interpreting and technology: The upcoming technological turn. Interpreting and technology, 1–12 (2018).
12. Fantinuoli, C.: The Technological Turn in Interpreting: The Challenges That Lie Ahead. In: BDÜ Conference Translating and Interpreting 4.0, Bonn (2019).
13. Fontaine, K., Manstead, A., Wagner, H.: Optimism, perceived control over stress, and coping. European Journal of Personality 7(4), 267–281 (1993).
14. Gile, D.: Basic Concepts and Models for Interpreter and Translator Training. John Benhamins, Amsterdam/Philadelphia (2009).
15. International Organization for Standardization: *ISO/PAS 24019 Simultaneous interpreting delivery platforms — Requirements and recommendations (2019).* Retrieved June 2020, from https://www.iso.org

16. Jimenez Serrano, O.: Foto fija de la interpretación simultánea remota al inicio del 2020. Revista Tradumática. Tecnologies de la Traducció 17, 59–80 (2019).
17. Kincal, Ş., Ekici, E.: Reception of remote interpreting in Turkey: A pilot study. RumeliDE Dil ve Edebiyat Araştırmaları Dergisi 21, 979–990 (2020).
18. Moser-Mercer, B.: Remote interpreting: Assessment of human factors and performance parameters. *Joint project International* (2003).
19. Moser-Mercer, B.: Remote Interpreting: Issues of Multi-Sensory Integration in a Multilingual Task. Meta 50(2), 727–738 (2005).
20. Moser-Mercer, B.: Remote interpreting. In: Gambier, Y., Van Doorslaer, L. (eds.) Handbook of Translation Studies, vol. 2. John Benjamins Publishing Company, Amsterdam/ Philadelphia (2011).
21. Mouzourakis, P.: Remote interpreting A technical perspective on recent experiments. Interpreting 8(1), 45–66 (2006).
22. Mulcahy, R.: Risk Management, Tricks of the Trade for Project Managers. RMC Publications, United States of America (2003).
23. Project Management Institute: A guide to the Project Management Body of Knowledge (PMBOK guide). Project Management Institute, Pennsylvania (2017).
24. Ricardi, A., Marinuzzi, G., Zecchin, S.: Interpretation and Stress. The Interpreters' Newsletter 8, 93–106 (1998).
25. Roziner, I., Shlesinger, M. Much ado about something remote: Stress and performance in remote interpreting. International Journal of Research and Practice in Interpreting 12(2), 214–247 (2010).
26. Saina, F. Remote Interpreting: Platform Testing in a University Setting to Shape a Post-COVID Scenario. In: ITA 2021 Conference: The new Workd - Translation in an Age of Uncertainty. Israel Translators Association, Tel Aviv (2021).
27. Seeber, K., & AIIC.: *Interpreting from the sidelines: Attitudes towards remote interpreting at the 2014 FIFA World Cup.* University of Geneva, Faculty of Translation and Interpreting, Geneva (2018).
28. Steptoe, A., Poole, L.: Control and Stress. In: Fink, G., Editor, Stress: Concepts, Cognition, Emotion, and Behavior. Academic Press, London (2016).
29. United Nations: *A joint experiment in remote interpretation. UNHQ-UNOG-UNOV*. Geneva: United Nations, Department of General Assembly Affairs and Conference Services (1999).
30. Ünlü, C.: Remote Simultaneous Interpretation in Pre and Post- COVID-19: An Overview of the Profession Revisited from Sectoral, Ethical, and Pedagogical Perspectives. Interdisciplinary Debates on Discourse, Meaning and Translation, 311–337 (2021).
31. Weaver, P.: The meaning of risk in an uncertain world. In: PMI Global Congress 2008. Malta (2008).
32. Ziegler, K.: Gigliobianco, S. Present? Remote? Remotely present! New technological approaches to remote simultaneous conference interpreting. In: Fantinuoli, C., Editor, Interpreting and technology, pp. 119-139. Language Science Press, Berlin (2018).

A Comparison between Named Entity Recognition Models in the Biomedical Domain

Maria Carmela Cariello[1][0000−0001−5001−0360], Alessandro Lenci[2][0000−0001−5790−4308], and Ruslan Mitkov[3][0000−0002−6074−2749]

[1] University of Pisa, Pisa 56126, Italy
m.cariello1@studenti.unipi.it
[2] University of Pisa, Pisa 56126, Italy
alessandro.lenci@unipi.it
[3] University of Wolverhampton, Wolverhampton, WV1 1LY, UK
r.mitkov@wlv.ac.uk

Abstract. The domain-specialised application of Named Entity Recognition (NER) is known as Biomedical NER (BioNER), which aims to identify and classify biomedical concepts that are of interest to researchers, such as *genes*, *proteins*, *chemical compounds*, *drugs*, *mutations*, *diseases*, and so on. The BioNER task is very similar to general NER but recognising Biomedical Named Entities (BNEs) is more challenging than recognising proper names from newspapers due to the characteristics of biomedical nomenclature. In order to address the challenges posed by BioNER, seven machine learning models were implemented comparing a transfer learning approach based on fine-tuned BERT with Bi-LSTM based neural models and a CRF model used as baseline. Precision, Recall and F1-score were used as performance scores evaluating the models on two well-known biomedical corpora: JNLPBA and BIOCREATIVE IV (BC-IV). Strict and partial matching were considered as evaluation criteria. The reported results show that a transfer learning approach based on fine-tuned BERT outperforms all others methods achieving the highest scores for all metrics on both corpora.

Keywords: Biomedical NER · Deep Learning · Transfer Learning.

1 Introduction

Named Entity Recognition (NER) is a task that aims to recognise and classify mentions of named entities in unstructured text into pre-defined semantic categories such as *person*, *organisation*, *location*, *time expression*, *monetary value*, and so on. In Natural Language Processing (NLP), NER not only acts as a tool for information extraction (IE), but plays an essential role in a variety of downstream applications such as information retrieval [7], text summarisation [13], machine translation [2], question answering [14], and many other NLP tasks.

The interest in NER is not a novelty, but it has been increasing in recent years due to the exponential growth of digital information that stimulated domain-specific applications of NER in order to extract entity mentions not only from

Translation and Interpreting Technology Online, pages 76–84,
July 05-07, 2021.
https://doi.org/10.26615/978-954-452-071-7_009

general texts, such as newspaper articles, but also from specialised texts. In many applied research domains, NER is directly used to alleviate the problem of the search and discovery of information, becoming an invaluable tool particularly in those research areas where it is difficult for researchers to keep up with relevant publications [20].

One specialised application of NER is known as Biomedical named entity recognition (BioNER), which is defined as the task of identifying and classifying Biomedical Named Entities (BNEs), technical terms referring to key concepts that are of interest to biomedical researchers, such as *gene*, *protein*, *chemical compound*, *drug*, *mutation*, *disease*, and so on. BioNER has gained increasing attention from the research community. In fact, many works in medicine focus on the analysis of scientific articles to find out hidden relationships between BNEs, such as *gene* and *protein*, in order to drive experimental research [20]. Although a large body of systems are dedicated to extract BNEs in scientific literature, BioNER tools can be applied to find all kinds of entities in any kind of health related text, including radiology reports and clinical notes [19].

Generally, BioNER is considered a more challenging task compared to domain-independent NER due to the characteristics of biomedical nomenclature. The lack of standardised naming conventions, the frequent crossover in vocabulary, the excessive use of abbreviations, synonyms and variations, the morphological complexity due to the use of unusual characters such as Greek letters, digits, punctuation – these are just some of the factors making the recognition of BNEs particularly difficult for BioNER systems. Moreover, biomedical text often contains complex multi-word BNEs and, especially in the area of gene and protein names, multi-word BNEs are rather the rule than the exception. Not only multi-word BNEs are more difficult to identify, but in many cases there is also no agreement on the exact borders of such names, making the evaluation of BioNER tools complex [11]. For example, many BNEs may contain verbs and adjectives that are embedded in names, making a legitimate gene or protein name hard to distinguish from the general language text surrounding it. Lastly, the biomedical domain is an expanding field where new concepts emerge daily and new names are coined on a daily basis. In addition, new variants are always created for already existing concepts since biomedical concepts are studied in different branches of medicine which use different naming conventions.

To address these challenges, seven Machine Learning (ML) models were implemented following a Sequence Tagging (ST) approach[4]. A transfer learning approach based on fine-tuned BERT is compared to Bi-LSTM-based neural models and a CRF model used as baseline. The impact of pre-trained word embedding models on the performances of neural models is also investigated. The comparison between models is carried out by evaluating the performances on two well-known BioNER corpora.

The rest of the paper is structured as follows. Section 2 presents the data used in this study. Section 3 outlines the models employed in our experiments,

[4] The Colab notebooks used for running the experiments are available here: https://github.com/cariello1/BioNER.

which are described in Section 4. Section 5 discusses the results obtained and finally section 6 summarises the conclusions of this study.

2 Data

BioNER models were evaluated using two benchmark corpora released during well-known and popular shared competitions. The first one is the corpus of the JNLPBA 2004 shared task, which is derived from the popular GENIA corpus. The second one is the BIOCREATIVE (BC-IV) corpus used for the Track 2 of BioCreative IV shared task. Both corpora were made publicly available in the IOB2 annotation format.[5] According to this schema, tokens are labelled with a *B-class* tag at the beginning of every sequence that represents an entity, with an *I-class* tag if the tokens are inside a sequence and with an *O* tag if the tokens are outside of a sequence that represents an entity.

2.1 JNLPBA 2004 Shared Task Corpus

Derived from the GENIA corpus, JNLPBA [8] is a manually annotated collection of articles extracted from the MEDLINE database. Compared to the 36 classes of the original corpus, JNLPBA has 5 classes: *protein*, *DNA*, *RNA*, *cell line* and *cell type*, and does not contain any nested or discontinuous entities. The training set includes entirely the GENIA corpus, while the test set consists of 404 newly annotated MEDLINE abstracts from the GENIA project. The training set contains 18,546 sentences for a total of 472,006 words, while the test set contains 3,856 sentences for a total of 96,780 words.

2.2 BioCreative IV CHEMDNER Corpus

BioCreative IV CHEMDNER (BC-IV) [10] is a collection of PubMed abstracts which contains chemical entity mentions labelled manually by experts in the field, following annotation guidelines specifically defined as part of the BioCreative IV competition. No nested annotations or overlapping entity mentions are included. The original fine-grained annotation schema including seven classes was collapsed into one generic class, *CHEMICAL*. The training set contains 30,682 sentences for a total of 891,948 words, while the test set contains 26,364 sentences for a total of 766,033 words.

3 Models

Considering BioNER as a Sequence Tagging (ST) task, seven models were implemented in order to solve BioNER and compare performances of a traditional ML algorithm used as a baseline with the latest advanced neural models.

[5] MTL-Bioinformatics-2016: https://github.com/cambridgeltl/MTL-Bioinformatics-2016.

3.1 Conditional Random Field

Conditional Random Fields (CRF) is a probabilistic graphical model, which provides a framework for modelling global probabilities based on some observations of the local functions and representing a distribution over labels [3]. CRF has the advantage over other ML algorithms to efficiently model dependencies between observations and labels, taking context into account. Among traditional ML algorithms, CRF is known as the most popular solution for solving ST tasks such as BioNER [9].

Given the morphological complexity behind BNEs, which are rich of unusual characters, applying features that have been used for traditionally named entities to identify biomedical instances could be insufficient. A specific set of features that exploit biomedical nomenclature characteristics needs to be engineered, in order to allow the algorithm to efficiently recognise BNEs [1].

3.2 Bi-LSTM Based Neural Networks

Bi-LSTM is a type of Recurrent Neural Network (RNN) that is widely used as context-encoder for ST tasks such as BioNER. RNNs are able to model context dependencies storing information during the sequential processing implementing units with self-connections [15]. However, standard RNNs suffer from the exploding gradient problem, which is responsible for the reduction in the ability to learn long-distance relationships, so that they have a limited application to real-world ST. LSTMs extend RNNs with a memory cell unit consisting of several gates to store and access information over long periods of time, efficiently modelling dependencies between far apart sequence elements as well as consecutive elements. Since LSTMs can access context only in one direction, Bidirectional LSTMs (Bi-LSTMs) are used instead, in order to scan the data in both directions and provide access to all surrounding context. Bi-LSTM combines the benefits of long-range memory and bidirectional processing, which make this model perfectly suitable for ST [6].

3.3 Fine-tuned BERT

BERT [5] is a pre-trained system based on Transformer that can be fine-tuned to solve specific language tasks. The Transformer [18] is a neural architecture which dispenses with recurrence entirely relying only on the attention mechanism to draw global dependencies between input and output. Since Transformers do not rely on sequential processing, they can process an input sequence of words all at once, allowing for much more parallelisation and requiring significantly less time to train compared to Bi-LSTM-based models [17]. Since Transformers allowed for a more efficient training on larger datasets than it was possible before they were introduced, they drastically improved the prospects of using Transfer Learning for Natural Language Processing (NLP). Indeed, in the last years a shift has occurred from the use of pre-trained word vectors for feature extraction to the use of pre-trained systems such as BERT, that has been trained on a huge general language dataset and can be fine-tuned to solve a wide variety of NLP tasks [12].

4 Experiments

The first experiment is aimed at training a CRF model, a traditional ML method widely used for solving BioNER that is easy to implement, provides reasonable results, and does not require much expertise and time to build. The CRF performance is considered as a baseline for evaluating the performance of the other models. For the CRF model, a specific set of features is used to allow the algorithm to recognise BNEs. Linguistic features are selected exclusively to exploit the characteristics of biomedical BNEs such as the morphological complexity. To enable the model to capture contextual information, context features are also provided in a 5-word window.

The first neural model implemented is a Bi-LSTM-based architecture that uses Softmax layer as the decoding layer. The embedding layer is initialised with random weights and computes word vectors during the learning process. Learned representations of data are fed into a Bi-LSTM layer which extracts contextual information. The output is then passed to another Bi-LSTM layer so that the model learns even deeper, more abstract representations from data. The decoding layer uses a Softmax function to transform scores into a probability distribution over classes. Labels for each word are independently predicted without taking into account dependencies between labels. In a second model, Softmax was replaced with CRF to make the model capable of capturing relationships between entity labels. It has been shown indeed that for ST tasks it is more beneficial to jointly decode label sequences using CRF than decoding each label independently [4].

In the next models, the Bi-LSTM+CRF architecture was enhanced replacing randomly initialised word vectors with different pre-trained distributed representation models. First, Bi-LSTM+CRF was combined with pre-trained vectors from FastText.[6] Next, these vectors were replaced with a concatenation of word-level and character-level representations using pre-trained word embedding from GloVe[7] and character embedding learned using an LSTM model. Character representations are able to capture sub-word level information such as prefix, suffix and orthographic characteristics enabling the model to handle the Out-Of-Vocabulary (OOV) problem, which causes GloVe to return many zero values. The last Bi-LSTM based neural model uses contextual embedding incorporating into the embedding layer a pre-trained ELMo [16] model. This model does not consider an additional character-level embedding, unlike the model with GloVe, since ELMo already provides context-dependent character-level representations.

[6] https://fasttext.cc/docs/en/crawl-vectors.html

[7] https://nlp.stanford.edu/projects/glove/

Finally, a fine-tuned BERT-Large model is employed. In the pre-processing step, a WordPiece tokenizer is used in order to allow the model to process words that it has never seen before by decomposing them into known sub-words. For restoring the original tokenisation, a post-processing step is needed in order to compare BERT outputs with those of the other models. The hyper-parameter settings and all the fine-tuning procedure rely on the indication provided on the original paper by Devlin et al. [5]. Due to the high number of parameters, the model is trained on an NVIDIA Tesla K80 16GB GPU.

5 Results

The results obtained on the test set for both corpora are shown in Tables 1 and 2. Precision, Recall and F1-score are reported according to the strict matching criterion, and the overall scores for JNLPBA are computed using micro-average. The F1-score computed according to the partial matching criterion is also reported. For what concerns the neural models, each model is run five times and the final reported result is the average among the runs.

JNLPBA				
Model	**Precision**	**Recall**	**F1**	**F1 (partial)**
CRF	0.68	0.69	0.69	0.78
Bi-LSTM+Softmax	0.68	0.69	0.69	0.78
Bi-LSTM+CRF	0.68	0.70	0.69	0.77
FastText+Bi-LSTM+CRF	0.67	**0.74**	**0.70**	0.77
GloVe+Char+Bi-LSTM+CRF	**0.68**	**0.75**	**0.71**	**0.79**
ELMO+Bi-LSTM+CRF	0.63	**0.77**	0.69	0.78
Fine-tuned BERT	**0.68**	**0.77**	**0.72**	**0.79**

Table 1. Overall performance of the models on JNLPBA.

BIOCREATIVE IV				
Model	**Precision**	**Recall**	**F1**	**F1 (partial)**
CRF	0.86	0.73	0.79	0.83
Bi-LSTM+Softmax	0.85	0.74	0.79	0.83
Bi-LSTM+CRF	0.77	0.83	0.80	0.87
FastText+Bi-LSTM+CRF	0.82	0.77	0.80	0.88
GloVe+Char+Bi-LSTM+CRF	0.83	**0.82**	**0.83**	0.88
ELMO+Bi-LSTM+CRF	0.77	**0.87**	**0.82**	0.86
Fine-tuned BERT	**0.89**	**0.87**	**0.88**	**0.93**

Table 2. Overall performance of the models on BIOCREATIVE IV.

The overall results show that BERT outperforms the other models, since it achieves the highest scores on both corpora. BERT proves to be able to effectively recognise and classify BNEs, despite being a model trained on text different from the target domain. For JNLPBA the scores do not differ in terms of precision compared to the other models but recall shows an improvement of 8% over the baseline model. A slight increase is also recorded for the F1-score compared to the baseline model, achieving 72% (against 69%) and 79% (against 78%) according to, respectively, the strict and partial matching evaluation criteria. The second and the third best performing models are respectively the Bi-LSTM+CRF that incorporates the GloVe+Character embedding and the model that incorporates the FastText embedding. A significant increase for the recall and a slight increase for the F1-score are recorded for both strict and partial matching over the baseline model. For BC-IV, instead, BERT stands out significantly over the other models, achieving outstanding scores on all metrics. Specifically, BERT outperforms the baseline model by 14% on recall and achieves an F1-score of 88% (against 79%) and 93% (against 83%) according to, respectively, the strict and partial matching evaluation criteria. For what concerns the precision the increase on the baseline is instead less remarkable. The BERT model outperforms also the second and the third best performing models that in this case are, respectively, the Bi-LSTM+CRF that incorporates the GloVe+Character embedding and the model that incorporates the ELMo embedding. A significant increase is recorded for all the metrics with the exception of the recall, where the ELMo model achieves a score comparable to BERT.

Using the GPU, the training of BERT on the BC-IV corpus requires only 20 minutes, while the Bi-LSTM models require more then 30 minutes. Therefore, even if the use of a GPU is required to fine-tune BERT, this model clearly outperforms the other approaches for the recognition of BNEs on both the biomedical test corpora.

6 Conclusion

Seven Machine Learning (ML) models were implemented following a Sequence Tagging (ST) approach for solving BioNER on two well-known corpora. A transfer learning approach based on fine-tuned BERT was compared with Bi-LSTM-based neural models and a CRF model used as baseline. The fine-tuned BERT model achieved the highest scores for all metrics on both corpora. Thus, according to what emerged from these experiments, the use of pre-trained vectors has a significant impact on the performance of the Bi-LSTM models, leading to an OOV error reduction and an increase of the recall. In addition, the inclusion of sub-word level information into the models proved to be particularly beneficial for solving BioNER on both corpora. Based on these results, the use of pre-trained transformer-based neural models such as BERT for solving BioNER looks promising. Specifically, the advantage of using BERT for BioNER lies in the fact that it can be employed as a ready-to-use model that can be easily fine-tuned for solving the task, requiring significantly less time to train and achieving superior performance scores compared to other approaches.

References

References

1. Alshaikhdeeb, B., Ahmad, K.: Biomedical Named Entity Recognition: A Review. Artificial Intelligence Review, 6(6), pp. 889, (2016)
2. Babych, B., Hartley, A.: Improving Machine Translation Quality with Automatic Named Entity Recognition. In: Proceedings of the 7th International EAMT workshop on MT and other language technology tools, Improving MT through other language technology tools, Resource and tools for building MT at EACL 2003, pp. 1-8. Association for Computing Machinery, Budapest, Hungary (2003)
3. Bengong Y., Zhaodi F.: A comprehensive review of conditional random fields: variants, hybrids and applications. Artificial Intelligence Review, 6(53), pp. 4289–4333, (2020)
4. Cho, H., Lee, H.: Biomedical named entity recognition using deep neural networks with contextual information. BMC Bioinformatics, 20, 735 (2019)
5. Devlin, J.,Chang, M., Lee, K., Toutanova, K. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota, USA (2019)
6. Graves, Alex: Supervised Sequence Labelling with Recurrent Neural Networks. Springer, Berlin (2012)
7. Guo, J., Xu, G., Cheng, X., Li, H.: Named Entity Recognition in Query. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 267–274. Association for Computing Machinery, Boston, MA, USA (2009)
8. Kim, J., Ohta, T., Tsuruoka, Y., Tateisi, Y., Collier, N.: Introduction to the Bio-Entity Recognition Task at JNLPBA. In: Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications, pp. 70–75. Association for Computational Linguistics, Geneva, Switzerland (2004)
9. Kocaman V., Talby D.: Biomedical Named Entity Recognition at Scale. ICPR Workshops (2020)
10. Krallinger, M., Rabal, O., Leitner, F., Vazquez, M., Salgado, D., Lu, Z., Leaman, R., Lu, Y., Ji, D., Lowe, D., Sayle, R., Batista-Navarro, R., Rak, R., Huber, T., Rocktäschel, T., Matos, S., Campos, D., Tang, B., Xu, H., Valencia, A.: The CHEMDNER corpus of chemicals and drugs and its annotation principles. Journal of Cheminformatics, 7(1), S2 (2015)
11. Leser, U. and Hakenberg, J.: What makes a gene name? Named entity recognition in the biomedical literature. Briefings in bioinformatics, 6(4), pp.357-369 (2006)
12. Malte A., Ratadiya P.: Evolution of transfer learning in natural language processing. ArXiv, abs/1910.07370 (2019)
13. McDonald, D. M., Chen, H.: Summary in Context: Searching versus Browsing. ACM Transactions on Information Systems (TOIS) **24**(1), 111–141 (2006)
14. Molla-Aliod, D., Zaanen, M., Smith, D.: Named entity recognition for question answering. In: Proceedings of the Australasian Language Technology Workshop 2006, pp. 51–58. Australasian Language Technology Association, Sancta Sophia College, Sydney, Australia (2006)
15. Nayel, H., Shindo, H., Shashirekha, H., Matsumoto, Y.: Improving Multi-Word Entity Recognition for Biomedical Texts. International Journal of Pure and Applied Mathematics, 118(16), pp. 301-320 (2018)

16. Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep Contextualized Word Representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 2227–2237. Association for Computational Linguistics, New Orleans, Louisiana, USA (2018)
17. Li, J., Sun, A., Han J., Li, C.: A Survey on Deep Learning for Named Entity Recognition. IEEE Transactions on Knowledge & Data Engineering, 1, pp. 1-1, (2020)
18. Vaswani A., Shazeer N.M., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I.: Attention is All you Need. ArXiv, abs/1706.03762 (2017)
19. Zhang S., Elhadad N.: Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. Journal of Biomedical Informatics, 46(6), pp. 1088-1098 (2013)
20. Zweigenbaum, P., Demner-Fushman, D., Yu, H., Cohen, K. B.: Frontiers of biomedical text mining: current progress. Briefings in bioinformatics, 8(5), pp. 358–375 (2007)

Cross-lingual Named Entity Recognition via FastAlign: a Case Study

Ali Hatami[1], Ruslan Mitkov[1], and Gloria Corpas[2]

[1] University of Wolverhampton, UK
{a.hatami, r.mitkov}@wlv.ac.uk
[2] University of Málaga, Spain
gcorpas@uma.es

Abstract. Named Entity Recognition is an essential task in natural language processing to detect entities and classify them into predetermined categories. An entity is a meaningful word, or phrase that refers to proper nouns. Named Entities play an important role in different NLP tasks such as Information Extraction, Question Answering and Machine Translation. In Machine Translation, named entities often cause translation failures regardless of local context, affecting the output quality of translation. Annotating named entities is a time-consuming and expensive process especially for low-resource languages. One solution for this problem is to use word alignment methods in bilingual parallel corpora in which just one side has been annotated. The goal is to extract named entities in the target language by using the annotated corpus of the source language. In this paper, we compare the performance of two alignment methods, Grow-diag-final-and and Intersect Symmetrisation heuristics, to exploit the annotation projection of English-Brazilian Portuguese bilingual corpus to detect named entities in Brazilian Portuguese. A NER model that is trained on annotated data extracted from the alignment methods, is used to evaluate the performance of aligners. Experimental results show the Intersect Symmetrisation is able to achieve superior performance scores compared to the Grow-diag-final-and heuristic in Brazilian Portuguese.

Keywords: Named Entity Recognition · Word Alignment · Cross-lingual.

1 Introduction

Word alignment is a Natural Language Processing (NLP) task that can be applied to a parallel text corpus to find the word-to-word correspondences in a sentence pair. It can be used in the Machine Translation (MT) pipeline to analyse the output of MT or to improve the quality of translation memory [1]. Although word alignment is not a necessary step in the MT pipeline, in special cases it helps the model to improve the translations. For example, word alignment is useful to translate domain-specific terminology and low-frequency content words [2]. Word alignment can also be used to match the alignments with the source annotations to determine the projection on the target text. It helps the

Translation and Interpreting Technology Online, pages 85–92,
July 05-07, 2021.
https://doi.org/10.26615/978-954-452-071-7_010

Named Entity Recognition (NER) system to detect Named Entities (NEs) on the target text based on the annotation information of the source text.

NER is a classification task that tries to identify and tag NEs in a given sentence. The benefits of performing NER during translation tasks range from improved translation quality to customer data protection. The performance of a machine learning-based NER system depends on multiple factors such as the amount of labelled data used to train the model. The availability of labelled data is one of the key points for training the NER systems. It is a basic challenge, especially for low-resource languages. One of the possible solutions for this challenge is to use cross-lingual approaches. These approaches use parallel bilingual corpus to extract the annotation information from languages with rich labelled data.

There are two different approaches to extract NEs from bilingual corpora: direct transfer approach, and annotation projection approach. Approaches based on the direct transfer try to use language-independent features to train the model on the source side and then directly apply it on the target side. Different cross-lingual features such as word-embeddings [3], word clusters [4] and Wikifier [5] can be used in the training process. These approaches suffer from sense ambiguity and word order differences that lead to noise in the output of the model [6].

Methods based on the annotation projection use word alignment information to project annotations from the source language to the target side in bilingual parallel corpora. Different approaches can be used to extract the alignments between sequences of words in the source and target languages. The most common approaches of word alignments are based on the IBM approach that is a classic word alignment model [7]. In the IBM models, every word in the source language can be aligned at most with one word in the target language that is called many-to-one mapping. But real-word alignment approaches that are based on the IBM model support different types of mappings such as many-to-many. The simplest approach to produce a many-to-many mapping is the symmetrization heuristic [8]. The symmetrization heuristic uses alignment in both directions. There are different methods to implement a symmetrization heuristic. Intersection and Union alignments are the most popular methods that merge alignments of both directions. The Intersection of two models expresses just a one-to-one relationship between words and it misses some of the alignments. So, it has a higher precision of alignment points but at the cost of losing in recall. Union of two models can capture all complementary information of both models. Unlike the Intersection heuristic, Union has a higher recall but lower precision.

There are some extended versions for Intersection to improve its performance such as the GDFA heuristic. It grows the Intersection heuristic by adding neighbouring alignment points from the union and unaligned points to the intersection [9]. The GDFA heuristic includes three steps. The first step is Grow-diag that intersects two-directional alignments and gradually considers the neighbourhood of each alignment point between the source and target languages. The second step is (-final) covers non-neighbour alignment points of intersection alignment points. The final step (-and) adds alignment points between two unaligned words.

This paper is organized as follows. Section 2 introduces NER and alignment methods. Section 3 describes experimental settings and results. Finally, Section 4 presents our conclusions and future work.

2 Methods

Word alignment provides useful information for several applications of NLP. MT is one of these applications that can use the benefits of word alignment methods to improve translations. Although the Neural MT models (NMT) do not rely on the word alignment approaches, they still play an important role to improve the output quality of the model. They can be used to extract an external lexicon and apply it in the inference process of MT [10]. It can help the MT model to better use domain-specific terminology to adapt the model with a new domain or to improve the translations of out-of-vocabulary content words. NER is another application to use word alignments in MT to improve its performance. NER is an information extraction task to automatically detect NEs in text and classifying them into predefined entity types such as PERSON, ORGANIZATION, LOCATION, TIME, DATE, etc. Today's NER systems are based on supervised machine learning models including Maximum Entropy Markov Models (MEMMs) [11], Conditional Random Fields (CRFs) [12], and neural networks [13]. Although the NER model that is based on neural architectures show high performance, they need a large amount of manually annotated NER data that is not available for low-resource languages [14]. In addition, the process of annotating by a human is a time- and money-consuming task.

Cross-lingual NER is an effective solution to tackle these challenges. It means transferring annotated information from a high-resource language that has enough annotated resources to a low-resource language with less or no annotated data. There are two main groups of cross-lingual NLP, direct transfer and annotation projection [6]. In direct transfer approaches, the NER model is trained on a language with rich labelled data and then applied to a text in a different language to detect NEs [4]. These approaches attempt to use language-independent features. But some features are dependent on the type of language. So, selecting a suitable set of features plays an important role in the quality of these approaches.

Annotation projection is another approach that is based on a parallel corpus between source and target languages [15]. These methods attempt to annotate the target side by using the annotation information of the source language. The quality of the annotation task is related to the quality of labelled data on the source side, the quality of alignments, and the size of the parallel data. This paper is a part of our research on cross-lingual NER transfer with minimal resources in the pipeline of MT. We focus on evaluating the performance of alignment methods in the annotation projection approach, where there is only one source language with rich label and no labelled data in the target language. In this paper, the accuracy of two alignment heuristics, Grow-diag-final-and (GDFA) and Intersect Symmetrisation are evaluated on the English-Brazilian Portuguese parallel corpus.

3 Experiments

In this paper, the NER model is based on the FLAIR [1] framework to train it on the training dataset and extract NER annotations from the test dataset. Before training the NER model, we used aligner methods to extract the annotation projection for the target language, Brazilian Portuguese, because we want to figure out ways of taking annotation benefits of the source language, English, that already is available.

3.1 Experimental Settings

The NER model of our experiment is on the FLAIR framework. The main idea of this framework is based on the word and document embeddings. It uses a simple GloVe embedding for 150 epochs to train the model [16]. We trained the model on our datasets (Logitech, Rakuten, TomTom and Udemy) in the English-Brazilian Portuguese language pair. Then performance of the trained model is evaluated on the test dataset. About 10% of the total data has been selected for each set, validation and test. Table. 1 shows the statistical information of training, validation and test datasets for the English-Brazilian Portuguese corpus. The entire dataset including training and test, has been manually annotated in the source side (it is called Gold-standard reference) as follows:

'text': "Use the Windows calibration utility."
$['token\_end' : 3,\ 'start' : 10,\ 'label' :'\ PRS',\ 'end' : 17,\ 'token\_start' : 3]$

In this research, the fast-align [2] model has been used to extract correspondence words (or multi-words) for the English-Brazilian Portuguese language pair. The main focus of this part of the project is to extract annotations of the target text by matching the alignments with the source annotations. Parameters of the alignment model were trained based on the generic data. Because the alignment model needs the tokenized source and target sentences as input. For the source side of the test dataset, we have tokenized sentences and manual annotation was provided by a linguistics expert. After tokenizing the target sentences, alignment algorithms were applied to the test dataset for extracting the annotation projection between the source and target languages. For example, in the sentence pair of "Go to Kobo.com. || Vá para Kobo.com.", the term of "Kobo.com" is a named entity with a tagged as URL label. An aligner tries to find correspondence words of the source text in the target text. In this example, the output of aligner is: Go → Vá || to → para || Kobo.com → Kobo.com. So we can detect the named entity of "Kobo.com" using projection between the source and target languages.

The GDFA and Intersect symmetrisation heuristics are used to obtain alignments. These heuristics use different alignment approaches in both directions (EN→BR-PR and BR-PR→EN). One of the notable differences between the outputs of GDFA and Intersect heuristics comes back to the approach of aligning

[1] $https : //github.com/flairNLP/flair$
[2] $https : //github.com/clab/fast_align$

multi-words to a single word. GDFA heuristic finds all correspondences between words, but Intersect just aligns one of the multiple words (first word) to a single word. For example, in the sentence pair of "Thank you for waiting. || Obrigado por esperar.", the term of "Thank you" in the source language must be aligned to "Obrigado" in the target language. The GDFA heuristic aligns both words, "Thank" and "you" to the target word (0-0 1-0 2-1 3-2 4-3) but the intersect heuristic just aligns "Thank" and omits the second word (0-0 2-1 3-2 4-3).

	Source Language			**Target Language**		
	Training	**Dev.**	**Test**	**Training**	**Dev.**	**Test**
Number of sentences	2811	281	363	2811	281	363
Number of words	21358	3858	5035	42146	4404	5106
Unique words	2251	1174	776	3645	1082	929
Number of NE	1153	112	88	1153	112	88

Table 1. Data statistics for the EN-BR PR corpus.

3.2 Results & Discussion

To evaluate the performance of the aligner algorithms, the accuracy of annotation projection can be investigated. Gold-standard annotation is used as a reference to evaluate the results of the projection. Table. 2 shows the accuracy of each label and overall accuracy for both heuristics based on Gold-standard reference. The accuracy of Intersect heuristic is 75% which is 10% higher than the accuracy of the GDFA heuristic. PRS and ORG are two labels that make this difference between the accuracy of two heuristics. The accuracy of the Intersect heuristic in PRS and ORG tags are 69% and 86%, respectively, while those of GDFA heuristic are 58% and 73%, respectively.

To evaluate the performance of the alignment heuristics, we use the NER model that trained on the annotated dataset. The annotated information for the training dataset can be extracted from the annotation projection by the aligners, GDFA & Intersect. The GDFA & Intersect heuristics were used to extract annotation in the target side by using annotation projection between the source language and the target language. Table. 3 shows the results for the NER model based on GDFA & Intersect heuristics respectively.

For the test dataset, Gold standard reference was used to evaluate the performance of the NER model as well as annotation data provided by heuristics. So, we trained the NER model on corpora that annotated using two different heuristics. For each model, we used a test dataset which was labelled in two different ways, relevant heuristic method and the Golden reference. The results show that the intersect-based NER model has higher f1-score than the NER model which trained on the GDFA heuristic. Based on the results of the alignment experiments (Table. 2), NEs in our test dataset can be grouped into 7 classes

	GDFA			**Intersect**		
label	**det.**	**no-det.**	**Acc (%)**	**det.**	**no-det.**	**Acc (%)**
NAME	8	2	80	8	2	80
PRS	32	23	58	38	17	69
URL	6	4	60	6	4	60
ORG	11	4	73	13	2	86
REFNUM	1	3	25	1	3	25
EMAIL	0	1	0	0	1	0
CRR	0	1	0	0	1	0
Overal	**58**	**38**	**65**	**66**	**30**	**75**

Table 2. Accuracy of GDFA & Intersect based on the source side's annotation.

including: NAME, PRS, URL, ORG, REFNUMBER, EMAIL and CRR. The NER model for both aligners detected some wrong classes that do not have correct named entity labels. This problem comes back to the setup of the aligners that aligned a named entity into the wrong label on the target side. The output of the aligners without correcting the wrong labels has been used to train the NER model. The results show that the Intersect-based NER model provides a better f1-score than the GDFA-based model in 6 classes out of 7. Only the f1-score for the class of "PRS" in the GDFA-based model (0.4138) is a bit better than that of the Intersect-based (0.4).

	GDFA-based NER		**Intersect-based NER**	
Reference	**Grow**	**Gold**	**Intersect**	**Gold**
NAME	0.3636	0.3636	0.3636	0.6154
PRS	0.3571	0.4138	0.3860	0.4
URL	0.4	0.8	0.8571	0.8
ORG	0.7619	0.7692	0.7692	0.8889
REFNUM	0.2857	0.3333	0.6667	0.3333
EMAIL	0.667	0.0	0.0	0.0
CRR	0.0	0.0	0.0	0.0
weighted avg.	**0.9788**	**0.9818**	**0.9861**	**0.9865**

Table 3. F1-score of NER model based on GDFA & Intersect.

4 Conclusion & Future Work

During this study, we focused on NER as a crucial task in the machine translation pipeline. The availability of labelled data for training the model is a main challenge of the NER systems. The focus of this project was to address this problem by using the aligner algorithms. The aligners can extract annotations in a target side (low-resource language) using annotation projection from the source side (high-resources language). The experiment shows the NER model that was trained on the Intersect heuristic has a better performance than GDFA. It seems that the performance of the aligners directly impacts on the performance of the NER model. Using state-of-art approaches for the alignment part can be a potential plan for future projects. In this project, restrictions on access to parallel bilingual dataset in English-Brazilian Portuguese by the source side's annotations impacts on the performance of the NER model as well as aligners. This project can be considered as a starting point for our study on the aligner approaches for annotation projection in low-resource languages.

References

1. Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio: Neural Machine Translation by Jointly Learning to Align and Translate. International Conference on Learning Representations, (2016).
2. Philip Arthur, Graham Neubig, Satoshi Nakamura: Incorporating Discrete Translation Lexicons into Neural Machine Translation. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Pages 1557–1567 (2016).
3. Chen-Tse Tsai, Dan Roth: Cross-lingual wikification using multilingual embeddings. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics. Pages 589–598 (2016).
4. Oscar Täckström, Ryan T. McDonald, and Jakob Uszkoreit.: Cross-lingual word clusters for direct transfer of linguistic structure.Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Pages 477–487 (2012).
5. Chen-Tse Tsai, Stephen Mayhew, Dan Roth: Cross-lingual named entity recognition via wikification. Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning. Association for Computational Linguistics, Pages 219–228 (2016).
6. Stephen Mayhew, Chen-Tse Tsai, Dan Roth: Cheap Translation for Cross-Lingual Named Entity Recognition. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Pages 2536–2545 (2017).
7. Chris Dyer, Victor Chahuneau, Noah A. Smith: A Simple, Fast, and Effective Reparameterization of IBM Model 2. Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Pages 644–648 (2013).
8. Franz Josef Och, Hermann Ney: A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics, Volume 29, Number 1, Pages 19–51 (2003).
9. Philipp Koehn, Franz J. Och, Daniel Marcu: Statistical Phrase-Based Translation. Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. Pages 127–133 (2003).

10. Stephen Mausam, Soderland, Oren Etzioni, Daniel S Weld, Kobi Reiter, Michael Skinner, Marcus Sammer, Jeff Bilmes, et al.: Panlingual lexical translation via probabilistic inference. Artificial Intelligence, 174(9-10):619–637 (2010).
11. Andrew McCallum, Dayne Freitag, and Fernando C. N. Pereira.: Maximum entropy markov models for information extraction and segmentation. In Proceedings of the Seventeenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pages 591–598 (2000).
12. John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the Eighteenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pages 282–289 (2001).
13. Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer: Neural architectures for named entity recognition. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, pages 260–270 (2016).
14. Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A. Smith, Jaime Carbonell: Neural Cross-Lingual Named Entity Recognition with Minimal Resources. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Pages 369–379 (2018).
15. Mengqiu Wang and Christopher D Manning.: Cross-lingual projected expectation regularization for weakly supervised learning. Transactions of the Association for Computational Linguistics, Volume 2, Pages 55–66 (2014).
16. Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, Roland Vollgraf: FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations). Association for Computational Linguistics, Pages 54–59 (2019).

Building a Corpus for Corporate Websites Machine Translation Evaluation

A Step by Step Methodological Approach

Irene Rivera-Trigueros[1][0000-0003-4877-4083] and María-Dolores Olvera-Lobo[2][0000-0002-0489-7674]

[1]University of Granada, Department of Translation and Interpreting,
Faculty of Translation and Interpreting, C/ Buensuceso, 11, 18002, Granada, Spain
ireneriver a@ugr.es

[2] University of Granada, Department of Information and Communication, Colegio Máximo
de Cartuja, Campus Cartuja s/n, 18071, Granada, Spain
molvera@ugr.es

Abstract. The aim of this paper is to describe the process carried out to develop a parallel corpus comprised of texts extracted from the corporate websites of southern Spanish SMEs from the sanitary sector which will serve as the basis for MT quality assessment. The stages for compiling the parallel corpora were: (i) selection of websites with content translated in English and Spanish, (ii) downloading of the HTML files of the selected websites, (iii) files filtering and pairing of English files with their Spanish equivalents, (iv) compilation of individual corpora (EN and ES) for each of the selected websites, (v) merging of the individual corpora into a two general corpus one in English and the other in Spanish, (vi) selection a representative sample of segments to be used as original (ES) and reference translations (EN), (vii) building of the parallel corpus intended for MT evaluation. The parallel corpus generated will serve to future Machine Translation quality assessment. In addition, the monolingual corpora generated during the process could as a base to carry out research focused on linguistic−bilingual or monolingual−analysis.

Keywords: Parallel Corpora, Monolingual Corpora, Machine Translation, Machine Translation Quality Assessment, Corporate Websites.

1 Background

Nowadays, thanks to the development of information and communication technologies, companies are able to spread messages globally, allowing them to open new markets. Web 2.0 tools, such as websites, provides enterprises, especially Small and Medium-sized ones (SMEs) with great opportunities for internationalization [1]. In fact, in the European Union (EU), more than 99% of all enterprises − save for the financial business sector − are SMEs [2] and 77% of them have a website [3]. However, language barriers often pose a challenge for companies when it comes to the multilingual dissemination of corporate information and that is why Machine Transla-

Translation and Interpreting Technology Online, pages 93–101,
July 05-07, 2021.
https://doi.org/10.26615/978-954-452-071-7_011

tion (MT) can be a resource with great potential for solving this problem. Nevertheless, MT quality is generally inferior to that reached by professional human translations. Consequently, MT evaluation, by means of automatic and human metrics, plays a key role for determining MT quality as well as for MT systems to be improved. However, the assessment of MT systems implies cognitive linguistic, social, cultural and technical processes [4]. As a result, assessing MT can present difficulties as in the majority of cases there is not just one correct translation [5]. In addition, it is important to note that there is a great lack of consensus in relation to translation quality assessment and approaches may differ according to the individuals, groups or contexts in which quality is assessed. Therefore, there are a number of metrics and criteria for undertaking the evaluation of MT systems. However, generally speaking, there are two main types of MT evaluation: human and automatic.

On the one hand, most automated metrics establish comparisons between the output of an MT system and one or more reference translation [4, 6]. Some of these measures such as WER, PER or TER are based on the Levenshtein or edit distance [7], the difference among this metrics is that some of this metrics consider word or phrases reordering as and edit operation [8]. Other measures, such as BLEU [9]–the most popular metric–are based on precision and carried out at the level of n-grams, indivisible language units. BLEU employs a modified precision that considers the maximum number of each n-gram appearance in the reference translation and applies a brevity penalty that is added to the measurement calculation. Other precision-centred metrics, for example, are NIST [10], ROUGE [11], F-measure [12] and METEOR [13].

On the other hand, human evaluation revolves around adequacy–based on semantic quality–and fluency–based on syntactic quality. For adequacy, evaluation reference translations or the original text, if the evaluators have language knowledge, are required. In the case of fluency evaluation, as the evaluation is monolingual, no reference translation nor the original text are necessary. Human evaluation can be carried out by means of ranking, Likert-type ordinal scales, gap filling tasks or by identifying, annotating, classifying and correcting translation errors, amongst others [8].

Human evaluation, despite demanding more time, effort and costs, is considered to be more reliable than automatic metrics, as their capacity to evaluate syntactic and semantic equivalence is limited [4, 6]. However, human evaluation cannot be reproduced, is less objective than automatic metrics and requires evaluators to fulfill certain criteria and to be trained prior the evaluation task. Therefore, it is advisable to combine various metrics that evaluate different aspects in order to assure the reliability of the results.

In the light of this scenario, the aim of this paper is to describe the process carried out to develop a parallel corpus (Spanish – English) comprised of texts extracted from corporate websites of southern Spanish SMEs from the sanitary sector which will serve as the basis for future MT quality assessment tasks.

2 Methodology

Prior research set the basis for the corpus presented in this paper [14]. The latter study analyzed 1425 Andalusian SMEs, belonging to what is referred to as Group Q: Healthcare and social services activities according to the CNAE-2009 classification (Spanish National Classification of Economic Activities). This economic sector was chosen as the healthcare sector is the second biggest group as regards business creation in net terms according to official reports [15]. The sample was selected using information from the Sectoral Ranking of Companies by Turnover offered by the Spanish source elEconomista.es, a daily newspaper with special focus on economics, finance, and business. The data from this Company Ranking comes from the INFORMA D&B S.A.U. (S.M.E.) database − which boasts the Spanish Association for Standardization and Certification (AENOR) quality certificate − and is fed from several public and private sources. This study concluded that around a half of the analyzed SMEs had a website, but only 10% of them offered their content translated to one or more languages.

The final goal of the research project is to evaluate MT applied to corporate information available on SMEs websites, hence, reference translations were needed to build the parallel corpus. Therefore, those companies offering translated content to English and Spanish served to build the corpus described in this paper, which responds to a sequential sampling strategy meaning that first phase results [14] determined the methodology of the next phase [16, 17].

The stages for compiling the parallel corpora were: (i) selection of websites with content translated in English and Spanish, (ii) downloading of the HTML files of the selected websites, (iii) files filtering and pairing of English files with their Spanish equivalents, (iv) compilation of individual corpora (EN and ES) for each of the selected websites, (v) merging of the individual corpora into a two general corpus one in English and the other in Spanish, (vi) selection a representative sample of segments to be used as original (ES) and reference translations (EN) and, (vii) building of the parallel corpus intended for MT evaluation.

2.1 Selection of the sample

Previous research [14] showed that 64 companies offered their contents translated into English from Spanish. A technique of stratified random sampling [18, 19] was applied for selecting the websites which will comprise the corpus. Medical specialties were considered as the base for weighting adjustment. Therefore, the final sample was comprised of 45 websites (Table 1).

Table 1. Websites sample selection

Medical specialties	N	%	Sample (N)
Polyclinics and hospitals	7	10,94	5
Plastic Surgery	8	12,50	5
Radiology-Diagnostic	5	7,81	3

Obstetrics and Gynecology	9	14,06	6
Ophthalmology	1	1,56	1
Physical Medicine & Rehabilitation	4	6,25	3
Dentistry	19	29,69	12
Healthcare Transport	1	1,56	1
Oncology	1	1,56	1
Cardiology	1	1,56	1
Gastroenterology	1	1,56	1
Psychology	1	1,56	1
Otorhinolaryngology	1	1,56	1
Surgery	2	3,13	1
Neuroscience research	1	1,56	1
Neurology	1	1,56	1
Addiction treatment	1	1,56	1
Total	**64**	**100**	**45**

After selecting the 45 websites which will comprise the sample a professional scientific English native translator certified that all selected websites met the quality standards of professional translation

2.2 Downloading of websites

Once the websites were selected, they were downloaded with Cyotek Webcopy tool. A website is made of great volumes of files, that is why the download was limited to the first three depth levels. The reason behind this decision is that it is usually recommended placing the most relevant information in the first levels so that users do not have to click several times to access it and three levels are sufficient to meet this requirement [20, 21]. In total, 3.31 GB were downloaded, comprising 52,734 files and 15,741 folders.

2.3 Filtering of files and pairing

The downloaded files were filtered, and the HTML English files were paired to their equivalents in Spanish so that it was possible to obtain the reference translations, being Spanish the source text and English the target text. To this end, the files were named and stored to facilitate their identification. Two folders –English and Spanish−were created for each website and the files were named so that the Spanish version of the homepage of a given website was named as *Web1* and stored in the folder *Spanish* while its equivalent in English–named *Web1* as well–was stored in the folder *English*. Once all the files were paired those files not being useful for corpus compilation (HTML files without English equivalent, JavaScript files, etc.) were deleted.

2.4 Compilation of individual corpora for each website

Two corpora were compiled for each of the websites with Sketch Engine corpora analysis tool [22]. One of the corpora was built using the English files and the other using the Spanish files and once compiled, the resulting TXT files were downloaded. In total, 90 TXT files were obtained, half of them in English and the other half in Spanish.

2.5 Compilation of the general corpora

In order to know how many translation segments–sentences–were in total in the sample two general monolingual corpora, one in Spanish and the other in English, were built using the TXT files obtained in the previous stage. Table 2 show corpora description. The difference in the volume of tokens, words, sentences and paragraphs, besides the linguistic features of each language, is due to the fact that some of the Spanish files contained more information that their equivalents in English**.**

Table 2. General corpora description

	ES_Health SMEs websites	*EN_Health SMEs websites*
Tokens	726,093	613,524
Words	638,202	536,226
Sentences	43,450	38,053
Paragraphs	29,514	24,581

2.6 Sample selection

The sample selection process started once the two monolingual corpora were built. Given the corpus purpose–serve as the basis to perform MT evaluation−it was determined that sentences will be the reference unit to select translation segments. The number of sentences of the English corpus, 38,053, was established as a reference to estimate sample size. The sample was calculated for a finite population (N = 38,053) for a confidence level of 99% with margin of error of 5% [23]. Thus, the final sample was comprised of 654 segments.

Given the variability of the size and length of the companies websites, the segments extracted from each website ranged from 4,907−largest website–to 23–smallest website−. For his reason, a technique of stratifed random sampling [20, 21] was applied again for selecting the segments which will form the parallel corpus. In this case, the amount of segments of each website was considered as the base for weighting adjustment. It is important to note that two websites did not have sufficient percentual weight with regard to the total population, so they were not supposed to add any segment to the sample. However, in order not to leave two companies without representation a translation segment from each website was selected. As a result, the final sample was comprised of 656 segments (Table 3).

6

Table 3. Segments table selection

Company ID	Segments (N)	%	Sample (N)
8622_MAM10	4907	12,90	84
8690_MAM14	4321	11,36	74
8623_MAM08	2403	6,31	41
8690_MAP55	2134	5,61	37
8622_COM06	1843	4,84	32
8690_MAP42	1620	4,26	28
8622_MAP01	1613	4,24	28
8623_GRP10	1280	3,36	22
8690_CAP03	1250	3,28	21
8622_MAM06	1185	3,11	20
8610_CAP02	1123	2,95	19
8610_MAM08	1083	2,85	19
8690_MAM02	1067	2,80	18
8610_SEM01	1027	2,70	18
8622_SEM26	979	2,57	17
8690_MAP72	909	2,39	16
8621_MAP17	807	2,12	14
8610_MAM06	687	1,81	12
8690_CAM04	679	1,78	12
8610_MAP04	584	1,53	10
8622_MAP55	565	1,48	10
8623_MAP80	531	1,40	9
8622_MAP27	456	1,20	8
8621_MAP03	445	1,17	8
8622_MAP17	422	1,11	7
8622_ALP08	401	1,05	7
8623_SEP23	376	0,99	6
8621_ALP07	373	0,98	6
8623_ALM01	358	0,94	6
8623_MAP16	344	0,90	6
8622_MAP18	324	0,85	6
8622_CAM09	323	0,85	6
8623_MAP14	293	0,77	5
8623_MAP27	209	0,55	4
8690_COP10	194	0,51	3

8622_MAP22	189	0,50	3
8690_MAP47	172	0,45	3
8621_CAP16	134	0,35	2
8622_MAP47	112	0,29	2
8690_JAP09	101	0,27	2
8623_MAP88	75	0,20	1
8623_CAP11	61	0,16	1
8622_MAP28	49	0,13	1
8623_MAP52*	23	0,06	0 + 1
8690_CAP19*	22	0,06	0 + 1
TOTAL	**38,053**	**100**	**656**

3 Conclusion and future research

The aim of this paper was to describe step by step the methodological approach followed to build a parallel corpus which will be used for MT quality evaluation − including both human and automatic assessment – of corporate websites belonging to SMEs from the healthcare sector. To this end, to the final XLS file, containing the original Spanish segments and their equivalents in English, one more column was added containing MT output, therefore, the XLS file contains automatic translations (EN) together with their original translation (ES) and reference translations (EN), which are both essentials to MT quality assessment. However, more columns could be added in the future to include automatically generated translations from more MT systems to compare their performance. Further analysis concerning MT error identification, annotation and classification could also be carried out using the parallel corpus as a base, along with the evaluation of the post-editing process. In addition, the parallel corpora can be easily enlarged by adding segments from the monolingual corpora, which are already formatted and numbered in order to make the process as efficient as possible. The parallel corpus can also be uploaded to corpus analysis tools such as Sketch Engine for further linguistic analysis.

On another note, the monolingual corpora generated in English and Spanish can also serve to carry out linguistic research, including comparison between languages or monolingual analysis. These two corpora, given its considerable volume, could be used to train purpose-built MT systems and they can also serve to enlarge the knowledge concerning the features of corporate texts from the healthcare sector.

Acknowledgements

Study supported by the Spanish Ministry of Science, Innovation and Universities (MCIU), the State Research Agency (AEI) and the European Regional Development

Fund (ERDF) via the RTI2018.093348.B.I00 project and by the MCIU via the University Staff Training Program (FPU17/00667).

References

1. Alcaide, J.C., Bernués, S., Díaz-Aroca, E., Espinosa, R., Muñiz, R., Smith, C.: Marketing y Pymes. Las principales claves de marketing en la pequeña y mediana empresa (2013).
2. Muller, P., Julius, J., Herr, D., Koch, L., Peucheva, V., McKiernan, S.: Annual Report on European SMEs 2016/2017: Focus on self employment. Unión Europea, Bruselas (2017).
3. Eurostat: Internet advertising of businesses-statistics on usage of ads (2019).
4. Castilho, S., Doherty, S., Gaspari, F., Moorkens, J.: Approaches to Human and Machine Translation Quality Assessment. In: Moorkens, J., Castilho, S., Gaspari, F., and Doherty, S. (eds.) Translation Quality Assessment. pp. 9–38. Springer, Cham 2018.
5. Shaw, F., Gros, X.: Survey of Machine Translation Evaluation. , Saarbrücken (2007).
6. Han, L.: Machine Translation Evaluation Resources and Methods: A Survey. arXiv Comput. Lang (2016).
7. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions and reversals. Sov. Phys. Dokl. 10, 707–710 (1966).
8. Chatzikoumi, E.: How to evaluate machine translation: A review of automated and human metrics. Nat. Lang. Eng. 26, 137–161 (2020).
9. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: BLEU: A Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. pp. 311–318. Association for Computational Linguistics, Stroudsburg, PA, USA (2002).
10. Doddington, G.: Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. In: HLT '02: Proceedings of the second international conference on Human Language Technology. pp. 138–144 (2002).
11. Lin, C.-Y., Hovy, E.: Automatic Evaluation of Summaries Using N-gram Co-Occurrence Statistics. In: Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. pp. 150–157 (2003).
12. Turian, J.P., Shen, L., Melamed, I.D., Melamed, I.D.: Evaluation of Machine Translation and its Evaluation. In: Proceedings of MT Summit IX. , New Orleans (2003).
13. Banerjee, S., Lavie, A.: METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. pp. 65–72. , Ann Arbor (2005).
14. Rivera-Trigueros, I., Olvera-Lobo, M.D.: Internet Presence and Multilingual Dissemination in Corporate Websites: A Portrait of Spanish Healthcare SMEs. J. Glob. Inf. Manag. 29, 1–17 (2021).
15. Dirección General de Industria y de la Pequeña y Mediana Empresa: Retrato de la Pyme (2020).

16. Hernández-Sampieri, R., Fernández Collado, C., Baptista Lucio, P.: Metodología de la Investigación. McGrah-Hill (1991).
17. Baltar, F., Gorjup, M.T.: Muestreo mixto online: Una aplicación en poblaciones ocultas. Intang. Cap. 8, 123–149 (2012).
18. Babbie, E.: Fundamentos de la investigación social. Ediciones Paraninfo, Méjico (2000).
19. Clairin, R., Brion, P.: Manual de muestreo. La Muralla, Madrid (2001).
20. US Department of Health and Human Services: Research-Based Web Design & Usability Guidelines. US Government Printing Office, Washington, DC (2006).
21. Powell, T.A.: Web Design: The Complete Reference. , Berkley, CA (2000).
22. Kilgarriff, A., Rychlý, P., Smrz, P., Tugwell, D.: The Sketch Engine. In: Williams, G. and Vessier, S. (eds.) Proceedings of the 11th EURALEX International Congress. pp. 105–115. Lorient, Francia (2004).
23. Martínez Bencardino, C.: Estadística y muestreo. Ediciones ECOE, Bogotá (2019).

SmarTerp: A CAI System to Support Simultaneous Interpreters in Real-Time *

Susana Rodríguez[1], Roberto Gretter[2], Marco Matassoni[2], Daniele Falavigna[2], Álvaro Alonso[3], Oscar Corcho[3], and Mariano Rico[3]

[1] Independent researcher, Madrid, Spain
[2] Fondazione Bruno Kessler, Trento, Italy
[3] Universidad Politécnica de Madrid, Madrid, Spain

Abstract. We present a system to support simultaneous interpreting in specific domains. The system is being developed thanks to a strong synergy among technicians, mostly experts on both speech and text processing, and end-users, i.e. professional interpreters who define the requirements and will test the final solution. Some preliminary encouraging results have been achieved on benchmark tests collected with the aim of measuring the performance of single components of the whole system, namely: automatic speech recognition (ASR) and named entity recognition.

Keywords: Computer-Assisted Interpretation · Multilingual Knowledge Graphs · Automatic Speech Recognition

1 Introduction

Simultaneous interpreting is a very cognitively demanding task consisting in the execution of different processing sub-tasks in parallel. As an example, if we take the interpretation of numbers, a high error or omission rate is observed, especially in the case of interpreters working in isolation (without a booth-mate, as in remote simultaneous interpreting or RSI), ranging from 70% in the case of students to as much as 40% in the case of professional interpreters [2]. As a further example, a study reported in [9] shows that the number of disfluencies (i.e. hesitations) produced by interpreters is significantly higher than that produced by non interpreters, mainly due to the lexical richness of interpreters themselves. The SmarTerp project aims to develop a Computer-Assisted Interpretation (CAI) system to support the simultaneous interpreter, especially in the RSI modality, by addressing the entire workflow of the interpreting activity, from the preparation of specialised multilingual glossaries that will serve to feed and train the ASR and AI built into the system and extract and propose terminology (e.g. named entities, numerals, etc) to assist the interpreter in real-time, to the post-event validation of new entries by the interpreter that will be fed back into the system to perpetuate a virtuous circle of generating and accumulating specialised knowledge for recurrent use by the interpreter/team of interpreters and the end-customer of the interpreting services.

Although the proposal of using both natural language processing (NLP) and ASR technologies is not new for developing CAI tools (see e.g. the works reported in [3] for a good review) and, at the same time, there are projects, such

* Under the aegis of the EIT Digital, supported by the European Institute of Innovation and Technology (EIT), a body of the EU

Translation and Interpreting Technology Online, pages 102–109,
July 05-07, 2021.
https://doi.org/10.26615/978-954-452-071-7_012

as EABM[1] that aim to use extensively ASR technology to create user-friendly interpreting interfaces, we believe that the strong synergistic effort produced in the SmarTerp project among NLP/ASR experts, software developers and end-users, aimed at both defining the requirements and evaluating and refining the performance of the resulting CAI system, can provide a significant step forward in the development of such tools.

2 Automatic Transcription of Audio

One of the requirements of the ASR systems used in the SmarTerp project is that they have to perform well on specific application domains. More precisely, the source language to be translated by the interpreter may contain a large number of technical terms and morphological variations that are usually not present (or occur with low frequencies) in "general purpose" training text corpora. The result is that a general purpose language model (LM) exhibits on in-domain data high values of both out-of-vocabulary (OOV) word rates and perplexities, worsening the word error rate (WER) of the ASR system that utilises it. To alleviate this effect we propose a procedure, described in section 2.2, that extracts from a given corpus the texts that are "closest", in some way, to a glossary of terms furnished by an interpreter. This one is assumed to contain most of the important words of the subject of a given interpretation session. Then, taking advantage from previous experience for estimating the proficiency of second language learners (see [6]), we developed a procedure, summarised in section 2.2, to adapt a general purpose LM to the domain of each interpretation session. This way we are able to instantiate an ASR engine specific to each interpretation session. Note that this has a strong impact on the whole architecture of the system, since it requires to update, on demand by the interpreters, the LM of each ASR engine.

2.1 Acoustic Models

The acoustic models are trained on data coming from CommonVoice [1] and Euronews transcriptions [7] , using a (Kaldi) standard *chain* recipe based on lattice-free maximum mutual information (LF-MMI) optimisation criterion [8]. In order to be more robust against possible variations in the speaking rate of the speakers, the usual *data augmentation* technique for the SmarTerp models has been expanded, generating time-stretched versions of the original training set (with factors 0.8 and 1.2, besides the standard factors 0.9 and 1.1).

Table 1 summarises the characteristics of the audio data used for the models in our five working languages.

2.2 Language Models

As previously mentioned, we assume a glossary will be available from which to derive some *seed words* that will be used, in turn, both to update the dictionary of the ASR system and to select LM adaptation texts from the available training corpora. These ones are derived both from Internet news, collected from about 2000 to 2020, and from a Wikipedia dump. Table 2 reports some statistics related

[1] see https://www.eabm.ugent.be/EABM

Table 1. Audio corpora for training the acoustic models.

Language	CV (h:m)	EuroNews (h:m)	Total Speakers	Running words
English	781:47	68:56	35k	5,742k
French	432:07	59:42	14k	3,637k
German	426:30	70:47	13k	3,196k
Italian	148:40	74:22	9k	1,727k
Spanish	322:00	73:40	16k	2,857k

Table 2. Text corpora for training the LMs for ASR in SmarTerp. Mw means millions of running words.

Language	Lexicon size	Total running words	Internet News	Wikipedia 2018
English	9,512,829	3790,55 Mw	1409,91 Mw	2380,64 Mw
French	4,422,428	1442,85 Mw	536,06 Mw	906,79 Mw
German	8,767,970	2015,47 Mw	972,89 Mw	1042,58 Mw
Italian	4,943,488	3083,54 Mw	2458,08 Mw	625,46 Mw
Spanish	4,182,225	2246,07 Mw	1544,51 Mw	701,56 Mw

to the training corpora used in this work for 5 different languages. Note that the huge lexicon size is due to the fact that Internet data have a very long queue of questionable terms (typos, etc.). To accomplish the task of text selection we implement the following steps:

- selection of the **seed words**, i.e. technical words that characterise the topic (i.e. the interpretation session) to be addressed; they are simply the words, in the glossary provided by the interpreter, that are not in the initial lexicon (composed by the most frequent 128 Kwords of that language);
- selection of the **adaptation text**, i.e. sentences in the training corpus that contain at least one of the seed words. Note that we hypothesise not having additional texts related to the topic to be addressed;
- creation of both the **adapted lexicon** and **adapted LM**.

Since several approaches can be employed to obtain and use the seed words (e.g. based on texts' distance, texts' semantic similarity, etc) we define the following indicators that allow to measure their effectiveness on benchmark tests (see section 5) collected and manually transcribed within the SmarTerp project.

- OOV rate. Since OOV words cannot be part of the ASR output, they will certainly be errors. We try to get a low OOV rate without increasing too much the lexicon size.
- WER of the ASR system.
- Precision, Recall and F-measure on a subset of technically significant words (hereafter called important words), manually marked in the benchmarks.

3 Semantic Interpretation

Once the transcripts are generated from the audio input, the role of the semantic interpretation module is to detect relevant entities that appear on these transcripts and that may be of interest for the interpreters. Examples of such entities are those that may be difficult for them to translate during the interpretation session, such as terms that are very specific to the domain or numerical values, which are known to be hard to translate since they require an additional cognitive effort due to the transcoding exertion they require, etc.

The main challenge in this context is that we are not dealing with a typical Named Entity Recognition problem, where elements like persons, organisations, places, etc., need to be detected. That is, recognising the entity "United States" in the text and offering its potential translation into Spanish "Estados Unidos" may not make much sense in the context of the whole system, since this is commonly a well-known term for interpreters. Using an example of the dentistry domain, it is rather more useful for an interpreter to identify the noun "flap" and provide its translation into Spanish ("colgajo"), or to identify a numerical value ("nineteen_seventy_six") and transform it into the Arabic numeral (the year 1976) the interpreter will recognise and introduce in the interpreted speech (in the target language) with little or no effort. Therefore, we need to talk about Interpreter-relevant Term Recognition and their translation into the target language.

To perform this type of task, the module is based on the usage of a layered set of multilingual general purpose, domain-specific and user-specific knowledge graphs, following best practices in the representation of multilingual linked data, as described in Section 3.1. The translation of numerical entities is discussed in Section 3.2.

3.1 Multilingual Knowledge Graphs

The terms and entities that are used by the system are represented using common practices for multilingual Linked Data [5]. These ensure that given an entity or term identified in the knowledge graph (e.g., https://www.wikidata.org/wiki/Q30 for Wikidata's term for the United States of America), the labels in different languages would be easily available using simple SPARQL queries.

As discussed in Section 2, for the overall system to work adequately it is important to adapt the underlying resources (in the case of this module, the multilingual knowledge graphs) to the interpreting sessions that are going to be performed. In our case, the resource management strategy of the multilingual terminologies that are used by this module differs slightly from the approach followed for the adaptation of language models used in the ASR component.

Instead of adapting a single resource, in our case we maintain three layers of multilingual knowledge graphs, which are used as the basis for the identification of terms to be translated and presented to the interpreter. The first layer contains rather small knowledge graphs that are generated from the multilingual glossaries (dictionaries) that are commonly maintained by interpreters, with domain-specific and event-specific terms. These glossaries are commonly edited by interpreters using spreadsheets, where each column contains terms, acronyms, etc., in every language of interest, and are commonly used by them when working on an interpreting session. The second layer contains domain-specific knowledge graphs (e.g., from the medical domain) that are generated from publicly available resources. This second layer is activated after the first one, when there are potentially-relevant terms that have not been identified in the first layer. The final layer contains an extract of existing knowledge graphs like Wikidata and DBpedia (with the relevant languages used during the interpretation session and only containing term URIs and their labels in different languages) that can be used in case that none of the previous ones are activated. In order to provide a very fast access to these multilingual knowledge graphs with a low memory consumption, we have generated Header-Dictionary-Triples (HDT) versions of them [4]. For each group of n words coming from the transcripts, we obtain the tokens and use combinations of 1 to 5 n-grams (words) so as to look for these terms in the different layers. Although this may seem like a brute force approach, our initial experiments have shown that it allows identifying relevant terms in the generated transcripts.

3.2 Numerical Entity Translation

Numbers are identified in the transcriptions provided by the ASR system using a special notation with underscores (e.g., `_sixty nine_`). This allows the semantic interpretation system to identify these terms easily, so that they do not need to be submitted to the knowledge-graph-based structure that was presented in the previous section. The transformations for numerical entities are implemented following a simple rule-based approach where the typical types of transformation across languages have been identified by a group of interpreters (transformations for years across languages, transformations for units used in quantities, etc.)

4 System Architecture

Fig. 1 shows the block diagram of the integration between the following modules:

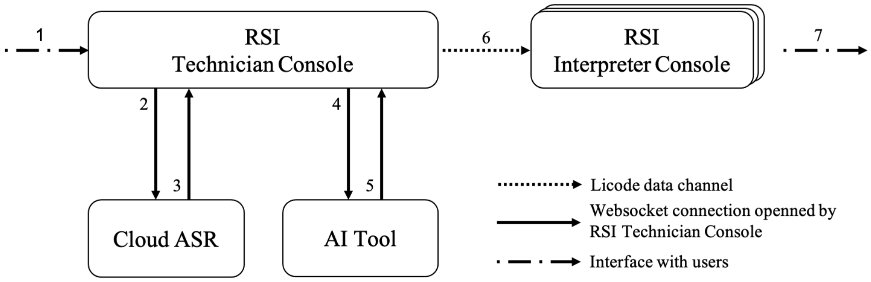


Fig. 1. Block diagram of RSI - ASR - AI integration

- **RSI Technician Console** represents the web-based interface used by the technical staff managing an interpreting session. Thanks to this interface, the technician is able to introduce in the system the audio and video flows of the conference speakers. On the other hand, **RSI Interpreter Console** represents the set of software consoles used by interpreters to visualise both the speakers' video and the materials (e.g. presentation slides) shared by them as well as to receive their audio. Interpreters also use these consoles to manage the input and output audio channels for developing the necessary tasks during an interpreting session. They can switch the input channels

from the technician one (i.e. floor) to the ones shared by other interpreters and modify their output language channel. In this console interpreters also see the AI tool output as explained below.
- **Cloud ASR** represents the on-demand cloud service in charge of transcribing in real-time an audio input in different languages for producing a text output. Its interface supports the receipt of a set of consecutive audio chunks (having a duration that will be optimised during the test phase of the SmartErp project) extracted from an audio flow through a websocket. As a result the Cloud ASR system sequentially respond with the text transcription in the same websocket connection using a JSON document.
- **AI Tool** receives an audio transcription and generates a set of terms for helping interpreters to perform their job. Its interface support the receipt of a set of consecutive JSON documents with a transcription through a websocket. As a result the AI tool generates the terms and sends them through the websocket interface when ready.

The complete interaction flow between the modules can be summarised as follows: 1) Using the JavaScript MediaStream API, the interface asks permission to access web camera and microphone of the PC used by the technician. This generates a video stream and an audio stream. These streams are shared with the interpreters using Licode [2], an open source multi videoconferencing platform based on WebRTC. 2) Thanks to the AudioContext API, the interface extracts audio chunks from the audio stream and sends them to the Cloud ASR using a previously opened websocket. 3) Cloud ASR synchronously answers with the text transcription using the format described above. 4) After receiving the transcription, the technician interface sends each JSON object with the transcription to the AI Tool using a second websocket connection previously created. 5) The AI Tool process the transcriptions and asynchronously generates the terms for interpreters. These terms are sent to the RSI technician console through the websocket connection. 6) Using Licode data channel the technician console multicasts the terms generated by the AI tool to the consoles of the interpreters connected to the same session. These terms are displayed, without a significant delay, to the interpreters in the console. 7) Audio output in the different available languages is sent to the assistants of the conference and to other interpreters in the same session.

5 System Evaluation

As mentioned above, in SmartTerp we prepared benchmarks for the 3 languages of the project (English, Italian, Spanish) plus two important European languages, French and German. Table 3 reports duration and number of words of the benchmarks; French and German are still in a processing stage. Data were collected and manually transcribed using Transcriber[3], a tool for segmenting, labelling and transcribing speech. In addition to time markers and orthographic transcription of the audio data, we decided to label with parenthesis Important Words (IWs), which represent content words that are significant for the selected domain (i.e. dentistry) and are a fundamental part of the desired output of the automatic system.

[2] https://lynckia.com/licode

[3] http://trans.sourceforge.net/

Table 3. Benchmarks collected and annotated in SmarTerp.

language	recordings	raw duration	transcribed duration	running words	running IWs
English	5	04:02:34	03:03:06	28279	3343
French	12	03:22:07	–	–	–
German	∼16	∼03:00:00	–	–	–
Italian	33	05:29:34	04:10:31	31001	4560
Spanish	13	03:09:53	03:01:59	25339	3351

Preliminary ASR results on the completed benchmarks are reported in Table 4, with and without the adaptation stage. Together with OOV rate and lexicon size, we report WER computed on all the uttered words (including functional words, which are useless for this task), and precision/recall computed only on IWs that, since they represent the most technically significant words in the domain, are more related to the output desired by interpreters. It is worth noting that the adaptation system is effective for all of the three languages and for all the considered metrics. Low WER for English is partly due to a scarce audio quality in the recordings, that mainly affects functional words: this explains the English high precision, which is computed on IWs only.

Table 4. Preliminary results for baseline and adapted systems. Both WER on all words and precision/recall/F-measure on isolated IWs are reported.

language	Lexicon size	OOV rate	WER	IWs: P / R / F
English baseline	128041	1.93%	26.39%	0.96 / 0.59 / 0.73
English adapted	213237	0.79%	23.34%	0.97 / 0.71 / 0.82
Italian baseline	128009	3.51%	15.14%	0.95 / 0.67 / 0.79
Italian adapted	1197995	1.02%	11.73%	0.98 / 0.82 / 0.90
Spanish baseline	128229	4.09%	22.60%	0.93 / 0.56 / 0.69
Spanish adapted	236716	1.14%	17.74%	0.98 / 0.75 / 0.85

6 Conclusions

The SmarTerp project is an on-going innovation action funded by the EIT Digital aiming to develop a Computer-Assisted Interpretation system to support the cognitively demanding task of simultaneous interpretation with state-of-the-art language technology. To do so, the consortium, created to solve the many challenges the real-time constraints impose on the system, has obtained so far encouraging results. In particular: *a)* good performance on specific application domains by the ASR systems thanks to a procedure that extracts from a given corpus the texts that are closest to a typical interpreters' glossary and adapts a general purpose LM to the domain of each interpretation session; and *b)* devising of a semantic interpretation module to detect relevant entities that appear on the ASR transcripts and that may be of interest for the interpreters, such as named entities and terms that are very specific to the domain, or numerical values, which are known to be difficult to interpret since they require an additional cognitive effort due to the transcoding exertion they require.

References

1. Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Morais, R., Saunders, L., Tyers, F., Weber, G.: Common voice: A massively-multilingual speech corpus. In: Proceedings of the 12th Language Resources and Evaluation Conference. pp. 4218–4222. European Language Resources Association, Marseille, France (May 2020)
2. Desmet, B., Vandierendonck, M., Defrancq, B.: Simultaneous interpretation of numbers and the impact of technological support. In: Multilingual Natural Language Processing. pp. 13—27. No. 11, C. Fantinuoli ed. Berlin: Language Science Press (2018)
3. Fantinuoli, C.: Interpreting and technology. In: Multilingual Natural Language Processing. No. 11, C. Fantinuoli ed. Berlin: Language Science Press (2018)
4. Fernández, J.D., Martínez-Prieto, M.A., Gutiérrez, C., Polleres, A., Arias, M.: Binary rdf representation for publication and exchange (hdt). Web Semantics: Science, Services and Agents on the World Wide Web **19**, 22–41 (2013), http://www.websemanticsjournal.org/index.php/ps/article/view/328
5. Gracia, J., Montiel-Ponsoda, E., Cimiano, P., Gómez-Pérez, A., Buitelaar, P., McCrae, J.: Challenges for the multilingual web of data. Journal of Web Semantics **11**, 63–71 (2012)
6. Gretter, R., Matassoni, M., Allgaier, K., Tchistiakova, S., Falavigna, D.: Automatic assessment of spoken language proficiency of non-native children. In: Proc. of ICASSP (2019)
7. Gretter, R.: Euronews: a multilingual speech corpus for ASR. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). pp. 2635–2638. European Language Resources Association (ELRA), Reykjavik, Iceland (May 2014)
8. Manohar, V., Hadian, H., Povey, D., Khudanpur, S.: Semi-supervised training of acoustic models using lattice-free MMI. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) pp. 4844–4848 (2018)
9. Plevoets, K., Defrancq, B.: The effect of informational load on disfluencies in interpreting: A corpus-based regression analysis. The Journal of the American Translation and Interpreting Studies Association **11**(2), 202–224 (2016)

Sign Language Translation in a Healthcare Setting*

Floris Roelofsen[1], Lyke Esselink[1], Shani Mende-Gillings[1], and Anika Smeijers[2]

[1] Institute for Logic, Language and Computation, Amsterdam, the Netherlands
{f.roelofsen,l.d.esselink,s.e.mendegillings}@uva.nl
[2] Amsterdam University Medical Centre, Amsterdam, the Netherlands
a.s.smeijers@amsterdamumc.nl

Abstract. Communication between healthcare professionals and deaf patients is challenging, and the current COVID-19 pandemic makes this issue even more acute. Sign language interpreters can often not enter hospitals and face masks make lipreading impossible. To address this urgent problem, we developed a system which allows healthcare professionals to translate sentences that are frequently used in the diagnosis and treatment of COVID-19 into Sign Language of the Netherlands (NGT). Translations are displayed by means of videos and avatar animations. The architecture of the system is such that it could be extended to other applications and other sign languages in a relatively straightforward way.

Keywords: text-to-sign translation · healthcare translation technology

1 Introduction

Communication between healthcare professionals and deaf patients is challenging [11], especially during the current COVID-19 pandemic [20]. Sign language interpreters can often not enter hospitals and clinics, interpreting via video relay is not always viable, and face masks make lipreading impossible [14].

We conducted an online questionnaire to understand how these difficulties are perceived by deaf people in the Netherlands. Questions were video recorded in Sign Language of the Netherlands (NGT) and also presented in written Dutch. 179 people participated, aged 20 to 84. The results—which will be presented in detail elsewhere—show that the general inability of healthcare professionals to communicate in sign language is indeed perceived as a very significant threat. Specifically, 88% of participants stated that they were worried about communication barriers should they need to be hospitalised with COVID-19, while, for

* We are grateful to Onno Crasborn, Manny Rayner, Marijke Scheffener, Irène Strasly, Roos Wattel, and Inge Zwitserlood for helpful discussions at various stages of this project. Special thanks to Bastien David, John Glauert, and Richard Kennaway for their very generous help with JASigning and SiGML. Finally, we gratefully acknowledge financial support from the Netherlands Organisation for Innovation in Healthcare (ZonMw, grant number 10430042010027) and the European Research Council (ERC, grant number 680220).

Translation and Interpreting Technology Online, pages 110–124,
July 05-07, 2021.
https://doi.org/10.26615/978-954-452-071-7_013

comparison, only 33% stated that they were worried about the fact that friends and relatives would not be allowed to visit them in the hospital.

To address this problem, we have developed a modular system which allows healthcare professionals to translate phrases that are frequently used in the diagnosis and treatment of hospital patients, in particular phrases related to COVID-19, from Dutch or English into NGT. For a limited number of sentences, translations are displayed by means of pre-recorded videos. In addition, the system is able to generate translations that are displayed by means of a signing avatar. The present paper focuses on describing the latter part of the system.

Although we have concentrated on NGT as the target sign language, the problem we aim to address manifests itself worldwide.[1] Therefore, in contrast to most existing text-to-sign translation systems (which are tailor-made for a specific target sign language and not easily portable to other languages, see Sections 3 and 4), we have aimed to design the system in such a way that it could be extended to include other source and target languages in a relatively straightforward way.

We should emphasise that a qualified human sign language interpreter should, whenever available, always be preferred over a machine translation system. Still, it is worth investigating the extent to which a machine translation system can be of help in situations in which a human interpreter cannot be employed, especially in the medical setting where effective, instantaneous communication between healthcare professionals and patients can be of critical importance.

2 Brief background on sign languages

Evidently, we cannot provide a comprehensive overview here of the linguistic properties of sign languages in general (see, e.g., [1]), nor of NGT in particular (see [19]). We will, however, highlight some important features which any text-to-sign translation system needs to take into account.

First of all, sign languages have naturally evolved in deaf communities around the world. This means that, contrary to a rather common misconception, there is not a single, universal sign language used by all deaf people worldwide, but many different sign languages, just as there are many different spoken languages.

Second, deaf people often have great difficulties processing spoken language even if presented in written form. The median reading level of deaf adolescents when finishing high-school is comparable to that of 8-year-old hearing children [27,17]. This may be surprising at first sight but not so much upon reflection. Imagine what it would be like as a native speaker of, say, English, to learn Hebrew or Thai just by looking at the characters, without being told how these characters are pronounced. Thus, for healthcare professionals to communicate with deaf patients exclusively through written text would not be satisfactory. This is especially true in a medical setting, where it is critical to avoid miscom-

[1] The World Federation of the Deaf estimates that there are around 70 million deaf people around the world (see https://wfdeaf.org/).

munication, to obtain reliable informed consent for interventions, and to foster an environment in which patients feel maximally safe.

Third, there is generally no direct correspondence between the sign language used in a given country and the spoken language used in that same country. For instance, while English is the mainstream spoken language both in the US and in the UK, American Sign Language (ASL) and British Sign Language (BSL) differ considerably from each other, as well as from spoken English. Such differences do not only pertain to the lexicon, but also to grammatical features such as word order. This means in particular that, to translate a sentence from English to ASL or BSL it does not suffice to translate every word in the sentence into the corresponding sign in ASL/BSL and then put these signs together in the same order as the words in the English sentence.

Fourth, signs are generally not just articulated with the hands, but often also involve facial expressions and/or movements of the head, mouth, shoulders, or upper body. These are referred to as the *non-manual* components of a sign. A text-to-sign translation system has to take both manual and non-manual components of signs into account.

Fifth, related to the previous point, non-manual elements are not only part of the *lexical* make-up of many signs, but are also often used to convey certain *grammatical* information (comparable to *intonation* in spoken languages). For instance, raised eyebrows may indicate that a given sentence is a question rather than a statement, and a head shake expresses negation. Such non-manual grammatical markers are typically ‘supra-segmental’, meaning that they do not co-occur with a single lexical sign but rather span across a sequence of signs in a sentence. Sign language linguists use so-called *glosses* to represent sign language utterances. For instance, the gloss in (1) represents the NGT translation of the question *Are you going on holiday?*.

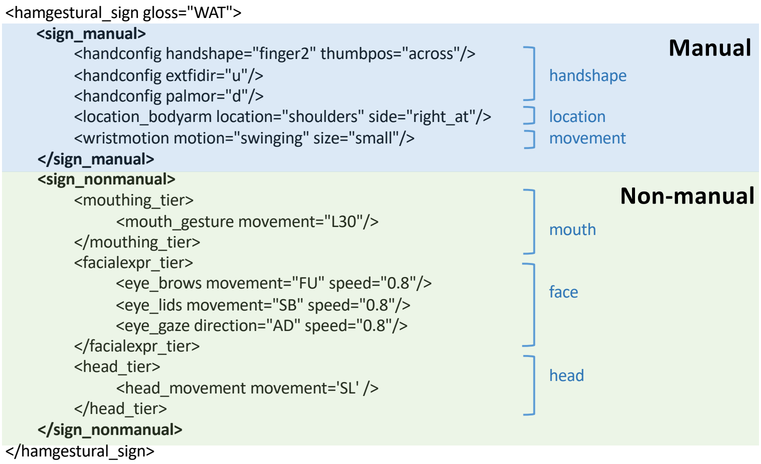
(1) $\overset{\text{brow raise}}{\overline{\text{YOU HOLIDAY GO}}}$

Lexical signs are written in small-caps. They always involve a manual component and often non-manual components as well. The upper tier shows non-manual grammatical markers, and the horizontal line indicates the duration of these non-manual markers. In this case, ‘brow raise’ is used to indicate that the utterance is a question. A text-to-sign translation system should be able to integrate non-manual elements that convey grammatical information with manual and non-manual elements that belong to the lexical specification of the signs in a given sentence [28]. This means that a system which translates sentences word by word, even if it re-orders the corresponding signs in accordance with the word order rules of the target sign language, cannot be fully satisfactory. More flexibility is needed: word by word translation can be a first step, but the corresponding signs as specified in the lexicon, must generally be adapted when forming part of a sentence to incorporate non-manual markers of grammatical information.

3 Sign synthesis

A crucial prerequisite for text-to-sign translation is sign *synthesis*: the ability to create sign language avatar animations. Broadly speaking there are two ways to achieve this: key-frame animation (e.g., [7]) and motion capture (e.g., [12]).

While motion capture makes it possible to obtain a library of high-quality animations for lexical signs, a disadvantage of this technique is that animations for lexical signs obtained in this way are difficult to modify so as to incorporate non-manual grammatical markers [5]. In principle, the same problem also applies to libraries of lexical signs obtained by means of key-frame animation. However, in this case, there is a promising strategy to overcome the problem. Namely, rather than directly animating each lexical sign, it is possible to generate key-frame animations of lexical signs procedurally from structured specifications of the phonetic properties of these signs [9]. Such phonetic properties include (but are not limited to) the initial location, shape and orientation of the hands, possibly movements of the hands and other body parts, and facial expressions. Several formalisms have been developed to specify the phonetic properties of signs in a structured, computer-readable fashion (see [5] for an overview). Arguably the most extensively developed and most widely used formalism is the Sign Gesture Markup Language (SiGML) [9,13], which is based on the HamNoSys notation originally developed for the annotation of sign language corpora [22,15]. For illustration, our SiGML encoding of the NGT sign WHAT is given in Figure 1. As can be seen in the figure, both manual components (handshape, location, movement) and non-manual features (mouth, face, head) are encoded.



```
<hamgestural_sign gloss="WAT">
    <sign_manual>
        <handconfig handshape="finger2" thumbpos="across"/>
        <handconfig extfidir="u"/>
        <handconfig palmor="d"/>
        <location_bodyarm location="shoulders" side="right_at"/>
        <wristmotion motion="swinging" size="small"/>
    </sign_manual>
    <sign_nonmanual>
        <mouthing_tier>
            <mouth_gesture movement="L30"/>
        </mouthing_tier>
        <facialexpr_tier>
            <eye_brows movement="FU" speed="0.8"/>
            <eye_lids movement="SB" speed="0.8"/>
            <eye_gaze direction="AD" speed="0.8"/>
        </facialexpr_tier>
        <head_tier>
            <head_movement movement='SL' />
        </head_tier>
    </sign_nonmanual>
</hamgestural_sign>
```

Fig. 1. SiGML encoding of the NGT sign WAT ('what').

SiGML specifications can be converted into key-frame animations by the JASigning avatar engine [9,18,16]. This approach makes it possible, in principle, to integrate non-manual grammatical markers with the lexical signs that make up a sentence, although such functionality has not yet been thoroughly implemented in systems based on SiGML and JASigning to our knowledge.

Given these considerations, we opted to use SiGML and JASigning as a basis for sign language synthesis, and to implement a new functionality to automate the integration of non-manual grammatical markers with lexical signs. A basic library of SiGML specifications of around 2000 lexical signs in NGT was already compiled in the course of previous projects ([10], see also [18,23,10]). While we have had to extend this library with healthcare-related as well as some general-purpose signs, the availability of an initial repertoire of signs encoded in SiGML was essential for a timely development of the system.

4 Text-to-sign translation: A modular approach

In text-to-sign translation, two general approaches can be distinguished, differing mainly in the type of intermediate representation that is employed in going from text to sign.

In the first approach, which we will refer to as the **gloss approach**, a given input sentence is transformed into a gloss of the corresponding sign language utterance. Next, based on this gloss representation, an avatar animation is generated.

(2) **Gloss approach**: text $\Longrightarrow$ gloss $\Longrightarrow$ animation

This approach is taken, for instance, by HandTalk, a Brazilian company that provides an automated text-to-sign translation service with Brazilian Portuguese and English as possible source languages, and ASL as well as Brazilian Sign Language as possible target languages. HandTalk uses machine learning techniques to map input texts to the corresponding glosses, and a combination of key-frame animation and motion capture techniques to generate animations based on a given gloss.

In the second approach, which we refer to as the **phonetic approach**, the given input sentence is transformed into a sequence of phonetic representations of signs. Next, based on these phonetic representations, an avatar animation is generated.

(3) **Phonetic approach**: text $\Longrightarrow$ phonetic rep. $\Longrightarrow$ animation

This approach has been taken in work based on SiGML and JASigning (see, e.g., [30,18,23,2,8,6]). Unlike in the gloss approach, applying machine learning techniques to carry out the first step, from text to phonetic representations, is not feasible because it would require the availability of large parallel corpora of texts and the corresponding phonetic sign representations, which do not exist and would be very costly to create. The process of manually generating phonetic representations is highly time-consuming and requires expert knowledge of SiGML or a similar formalism. Rayner et al. [24] have created a framework to ease this process, which is especially helpful if the sentences that need to be translated are all variations of a limited set of templates. For instance, the framework has been used successfully to develop an application for translating railway announcements [6]. In less restricted domains, however, generating phonetic representations still requires expert knowledge of SiGML or similar formalisms and remains very time-intensive.

The gloss approach and the phonetic approach have complementary pros and cons. An advantage of the gloss approach is that it enables the use of machine learning technology to carry out the first part of the translation process. Disadvantages are that (i) the animation of each individual sign involves a lot of manual work, (ii) grammatical non-manual elements cannot be straightforwardly integrated with lexical signs, and (iii) all components of the system are tailor-made for a particular target sign language, i.e., no part of the system can be re-used when a new target language is considered. In particular, since no gloss-based system currently exists for NGT, this approach was not viable for our purposes.

Advantages of the phonetic approach are that (i) grammatical non-manual features can in principle be integrated with lexical signs (though this possibility remains largely unexplored) and (ii) part of the system, namely the software that generates avatar animations based on phonetic representations (i.e., JASigning or a similar avatar engine) is not language-specific and can be used for any target sign language. The main disadvantage is that the initial step from text to phonetic representations involves a lot of manual work.

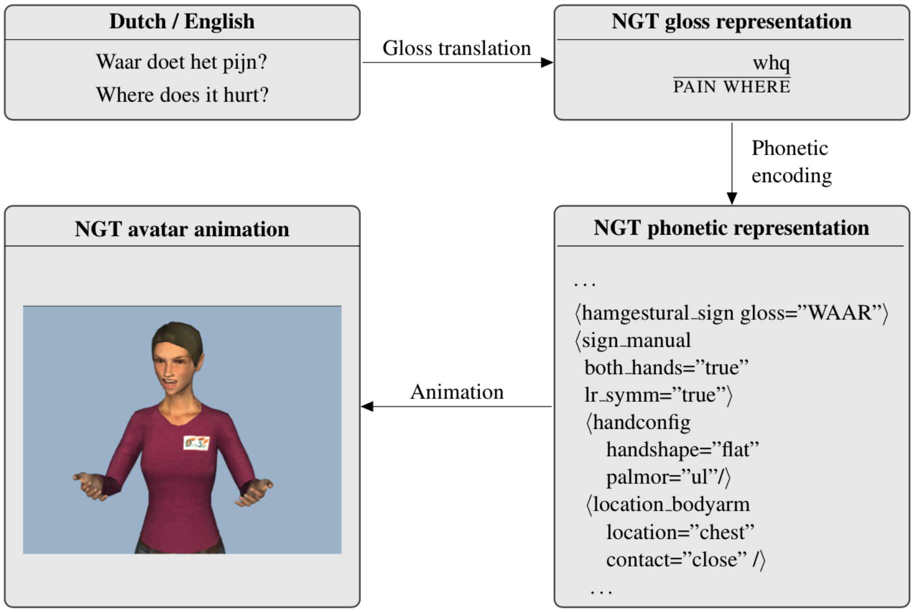
Given these considerations, we propose a **modular approach**, which employs *both* a gloss representation *and* a phonetic representation in going from a given input text to an avatar animation of the corresponding sign language utterance. As depicted in Figure 2, our modular approach breaks the translation process up into three steps:

1. **Gloss translation**
   In this step, the given Dutch or English input sentence is mapped to a gloss representation of the corresponding NGT sentence.
2. **Phonetic encoding**
   In this step, the NGT gloss is transformed into a computer-readable phonetic representation, in our case formulated in SiGML.
3. **Animation**
   In this step, an avatar animation is generated based on the given phonetic representation.

Consider, for instance, the Dutch/English input sentence in (4):

(4) Waar doet het pijn?
Where does it hurt?

The first step is to convert this sentence into the corresponding NGT gloss in (5), where 'whq' stands for the non-manual marking that is characteristic for constituent questions in NGT. While empirical studies have found quite some



Waar doet het pijn?
Where does it hurt?

$\overline{\text{PAIN WHERE}}^{\text{whq}}$

```
...
⟨hamgestural_sign gloss="WAAR"⟩
⟨sign_manual
 both_hands="true"
 lr_symm="true"⟩
 ⟨handconfig
   handshape="flat"
   palmor="ul"/⟩
 ⟨location_bodyarm
   location="chest"
   contact="close" /⟩
 ...
```

Fig. 2. Overview of the modular translation pipeline.

variation in the actual realisation of 'whq' in NGT [4,26], furrowed eyebrows are seen as the most canonical realisation [19].

(5) $\overline{\text{PAIN WHERE}}^{\text{whq}}$

The second step is to map this gloss representation to a phonetic representation in SiGML, a fragment of which is displayed in Figure 2. Finally, this SiGML representation is fed into the JASigning avatar engine, which generates an animation (see Appendix A for a visualisation).

5 Implementation

The implementation choices we have made were driven by the specific objective to address the urgent need for a translation tool to aid healthcare professionals in communicating with deaf patients, ensuing from the current pandemic. Two requirements follow from this objective: (i) the system had to be developed within a short time-frame, and (ii) high accuracy of the delivered translations was more important than broad approximate coverage.

Our aim has therefore *not* been to automate the entire translation process. In particular, automating the process of mapping input sentences to the corresponding NGT glosses using machine learning techniques would not have been feasible within a short time-frame, and would, even in the somewhat longer term, most likely result in an unacceptably low accuracy rate for use in a healthcare setting.[2] We therefore mainly focused on automating the phonetic encoding step, something that significantly reduces the manual labor needed in the overall translation pipeline. Automating the mapping from glosses to phonetic representations has not been done in previous work on NGT [23] and, to the best of our knowledge, not in work on other sign languages either.

5.1 Collecting phrases for translation

We collected a set of phrases that are commonly used during the diagnosis and treatment of COVID-19, based on consultation with healthcare professionals at the Amsterdam University Medical Centre (AUMC) as well as direct experience (one of the authors is a medical doctor). We also consulted a list of phrases that was used in the SignTranslate system [21].[3]

The resulting corpus was then divided into three categories: video-only, avatar-only, and hybrid. The first category, video-only, consisted mainly of sentences that could be divided into three further categories: emotional, complex, and informed consent. Sentences concerning the patient's emotional well-being require a high level of empathy to be conveyed, which is difficult to achieve in a satisfactory way with an avatar given the current state of the art. We therefore deemed that video translations were necessary for these sentences. Sentences were classified as complex when they involved a combination of several statements and/or questions, or required a demonstration of pictures or diagrams along with an explanation (see Appendix B for an example). Finally, in the case of questions and statements concerning informed consent it is especially important to leave no room for potential misunderstandings. To ensure this, we chose to always offer video translations of these sentences.

The second category, avatar-only, consisted of sentences with many variations differing by only one word or phrase, indicating for instance the time of day or a number of weeks. It would not have been feasible to record a video translation for each version of these sentences.

The third category, hybrid, consisted of sentences that do not fall into one of the other two categories. For these, the system offers both a video translation and an avatar translation. In some cases, the avatar translation is slightly simplified compared to the video translation.

After categorising all of the sentences, those from the first and third category were translated into NGT and recorded by a team consisting of a sign language interpreter and a deaf signer. The deaf signer who is visible in the videos was

[2] An recent study investigating the feasibility of automated sign language translation for children television programs in the Netherlands [23] drew the same conclusion.

[3] The SignTranslate system was developed in the UK around 2010 to translate phrases common in a healthcare setting from English to British Sign Language. Translations were displayed by means of videos, not by avatar animations. Evidently, the system was not specifically targeted at COVID-19 healthcare. However, many general-purpose phrases are also relevant in the diagnosis and treatment of COVID-19.

chosen for her clear signing style without a specific dialect, and her neutral reputation within the deaf community. Translations were checked by one of the authors (Smeijers), who is a sign linguist and a medical doctor. This resulted in a collection of 139 video translations. The sentences from the second and third category (including all variations) together comprised 7720 sentences for avatar translation.

5.2 Constructing SiGML representations

In order for the system to operate fast at run-time, we pre-processed all sentences and stored SiGML representations of their translations in a database. At run-time, the system only queries this database and does not compute any translations on the fly.

To construct the SiGML representations of full sentences, we implemented a program that, when given the gloss representation of a sentence in NGT, creates the SiGML code for that sentence. It first retrieves the SiGML code for all lexical signs in the given gloss from a lexical database, and then adapts this code to add non-manual grammatical elements. For instance, in the case of yes/no questions, the program makes sure that the sentence ends with the general interrogative sign in NGT (palms up, raised eyebrows) and changes the non-manual component of the last sign before this general interrogative sign to include raised eyebrows, in line with what we observed in our collection of video translations. In the case of wh-questions, the general interrogative sign was also always appended at the end of the sentence. Although the use of this sign in questions is in fact *optional* in NGT [4], we expect that it increases comprehension in the case of avatar translations.

5.3 User interface

We developed an online user interface. The user chooses a translation format (video or avatar) and enters a sequence of search terms. Based on their input they are presented with a list of available sentences from the database. These sentences may differ depending on the translation format chosen (video/avatar). After selecting a sentence the translation is offered in the chosen format.

As mentioned earlier, some of the possible input sentences differ only in one word or phrase. These sentences can be thought of as involving a general template with a variable that can take several values, such as a day of the week, a time of day, or a number of times / minutes / hours / days / weeks / months. When a user wants to translate such a sentence, they first select the template and then provide the intended value for the variable. For example, they may select the template "I am going to explain more at *time*", and then select a particular time (as illustrated in Appendix C).

While JASigning in principle offers a number of different avatars for sign language animation, there are differences in execution between these avatars. Our user interface therefore only makes use of one of the avatars, Francoise, and does not allow the user to choose between different options. We intend to further optimise the visualisation of the avatar in future work.

6 Discussion

As a first step in evaluating the system we have consulted extensively with a prominent member of the deaf community in the Netherlands who has years of experience in advising organisations (especially museums and hospitals) on how to make their services more accessible to deaf people. Based on these consultations and our own experiences in developing the system, we believe that the following considerations will be helpful in guiding further work in this direction.

The main advantage of avatar technology over video translation is that it provides flexibility and scales up more easily. Once a library of animated signs has been created, and a procedure to integrate non-manual grammatical markers has been implemented, translations for many sentences can be generated. This makes it particularly straightforward to provide translations for sentences that differ only slightly from each other (e.g., in a phrase indicating the time of day).

A disadvantage, however, of avatar translations is that they can be less natural and more difficult to comprehend. While several empirical studies have reported promising comprehension rates for JASigning avatars (see, e.g., [18,25]), our consultant indicates that certain avatar translations offered by our system may be difficult to understand for some users. Certain signs differ from each other only in rather subtle ways, and may be indistinguishable when produced by the JASigning avatar. Certain facial expressions and body movements of the avatar are quite unnatural, which can add to the difficulty of understanding translations. Certainly, the avatar's ability to display emotional empathy is very limited. This makes it undesirable to use avatar translations in situations where such empathy is required, as is often the case in medical settings.

Video translations, on the other hand, have their own benefits and drawbacks. They are better than avatar translations in terms of naturalness and comprehensibility, especially in the case of complex sentences. Moreover, our consultant indicates that patients are likely to feel more comfortable watching a video of a human signer rather than an animated avatar in a situation in which their physical well-being is at stake.

The main disadvantage of a video translation system is its inability to scale up efficiently. All translations have to be recorded separately, even ones that are almost identical. Cutting and pasting video fragments of individual signs to create new sentences does not yield satisfactory results.

A general advantage that a machine translation system (using either prerecorded videos, or an avatar, or both) may sometimes have over a human interpreter, especially in the healthcare domain, concerns privacy (see also [3]). A patient may receive sensitive information, and may not want this information to be known to anyone else (the deaf community in the Netherlands is relatively small, which makes it relatively likely that a patient and an interpreter are personally acquainted). In this case, employing a human interpreter has a disadvantage (though this may of course be outweighed by the higher level of translation accuracy and empathy that can be provided by a human interpreter).

It is important to emphasise that constructing sign language translations in either format is a time-consuming affair, though for different reasons. Building a corpus of video translations is time intensive because every translation has to be recorded separately. For avatar translations, it takes time to encode individual signs. The latter are reusable, however, which becomes especially attractive as the number of required translations grows. However, the overall preference for one method over another is context-dependent: pros and cons should be carefully weighed in each specific context.

Finally, we note that one clear limitation of the current system is that it only translates text into sign language, not the other way around. This means, for instance, that if a doctor uses the system to ask a deaf patient an open-ended question such as *How do you feel?*, and if the patient gives an elaborate answer in NGT, the doctor will most likely not be able to understand the answer and our system will not be of help in this case. Overcoming this limitation would require incorporating sign recognition technology (see, e.g., [29]), which has been beyond the scope of our project so far. Note, however, that if a doctor uses our system to ask a more specific yes/no question such as *Do you feel dizzy?*, then the answer in NGT—involving a head nod in the case of *yes* and a head shake in the case of *no*—will most likely be perfectly clear for the doctor even without a general understanding of NGT. Thus, the current system is able to support relatively simple dialogues, but it is limited in scope and certainly does not (yet) offer a full-fledged dialogue system. We view it as a first, but critical step toward a more comprehensive solution.

7 Conclusion and future work

We have investigated the potential of automated text-to-sign translation to address the challenges that the current pandemic imposes on the communication between healthcare professionals and deaf patients. We have motivated a modular approach to automated text-to-sign translation, and have built a first prototype system following this approach. We have discussed various prospects and limitations of the system.

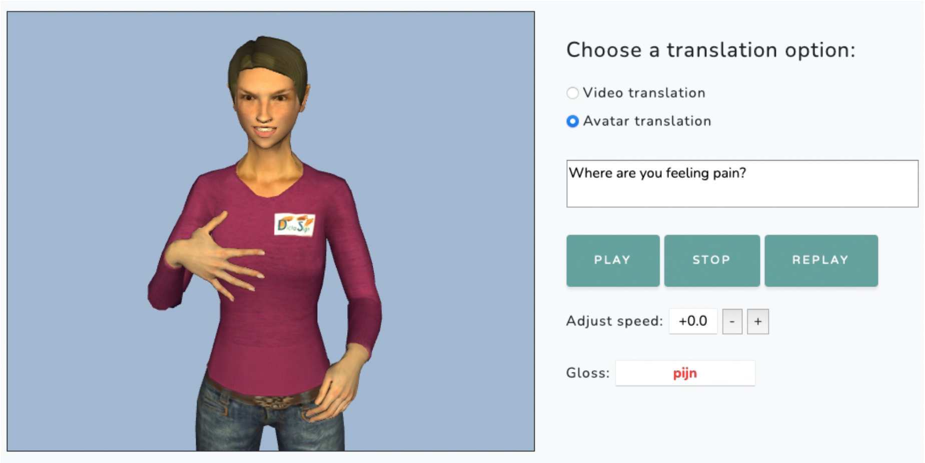
In future work, we intend to evaluate the system more thoroughly and to employ the modular approach motivated here to develop text-to-sign translation systems for different domains, e.g., for announcements at airports or railway stations, a use case which has already been explored to some extent for other sign languages [2,8]. In addition, we also intend to improve the avatar visualisation.

References

1. Baker, A., van den Bogaerde, B., Pfau, R., Schermer, T.: The linguistics of sign languages: An introduction. John Benjamins Publishing Company (2016). https://doi.org/10.1075/z.199
2. Battaglino, C., Geraci, C., Lombardo, V., Mazzei, A.: Prototyping and preliminary evaluation of a sign language translation system in the railway domain. In: Antona, M., Stephanidis, C. (eds.) Universal Access in Human-Computer Interaction. pp. 339–350 (2015). https://doi.org/10.1007/978-3-319-20681-3_32
3. Bouillon, P., David, B., Strasly, I., Spechbach, H.: A speech translation system for medical dialogue in sign language—Questionnaire on user perspective of videos and the use of Avatar Technology. In: Proceedings of the 3rd Swiss Conference on Barrier-free Communication (BfC 2020). pp. 46–54 (2021)
4. Coerts, J.: Nonmanual grammatical markers: an analysis of interrogatives, negations and topicalisations in Sign Language of the Netherlands. Ph.D. thesis, University of Amsterdam (1992)
5. Courty, N., Gibet, S.: Why is the creation of a virtual signer challenging computer animation? In: International Conference on Motion in Games. pp. 290–300. Springer (2010). https://doi.org/10.1007/978-3-642-16958-8_27
6. David, B.V.C., Bouillon, P.: Prototype of Automatic Translation to the Sign Language of French-speaking Belgium. Evaluation by the Deaf Community. Modelling, Measurement and Control C **79**(4), 162–167 (2018). https://doi.org/10.18280/mmc_c.790402
7. Delorme, M., Filhol, M., Braffort, A.: Animation Generation Process for Sign Language Synthesis. Advances in Computer-Human Interactions, ACHI **9**, 386–390 (2009). https://doi.org/10.1109/ACHI.2009.29
8. Ebling, S., Glauert, J.: Building a Swiss German Sign Language avatar with JASigning and evaluating it among the Deaf community. Universal Access in the Information Society **15**(4), 577–587 (2016). https://doi.org/10.1007/s10209-015-0408-1
9. Elliott, R., Glauert, J., Jennings, V., Kennaway, R.: An overview of the SiGML notation and SiGML signing software system. In: Workshop on the Representation and Processing of Sign Languages at the Fourth International Conference on Language Resources and Evaluation (LREC). pp. 98–104 (2004)
10. Esselink, L.: Lexical resources for sign language synthesis: The translation of Dutch to Sign Language of the Netherlands (2020), bachelor's thesis. University of Amsterdam, https://scripties.uba.uva.nl/search?id=715792
11. Fellinger, J., Holzinger, D., Pollard, R.: Mental health of deaf people. The Lancet **379**(9820), 1037–1044 (2012). https://doi.org/10.1016/S0140-6736(11)61143-4
12. Gibet, S., Courty, N., Duarte, K., Naour, T.L.: The SignCom system for data-driven animation of interactive virtual signers: Methodology and evaluation. ACM Transactions on Interactive Intelligent Systems (TiiS) **1**(1), 1–23 (2011). https://doi.org/10.1145/2030365.2030371
13. Glauert, J., Elliott, R.: Extending the SiGML notation–a progress report. In: Second International Workshop on Sign Language Translation and Avatar Technology (SLTAT) (2011)
14. Grote, H., Izagaren, F.: COVID-19: The communication needs of D/deaf healthcare workers and patients are being forgotten. British Medical Journal **369** (2020). https://doi.org/10.1136/bmj.m2372

15. Hanke, T.: Hamnosys-representing sign language data in language resources and language processing contexts. In: LREC. vol. 4, pp. 1–6 (2004)
16. Jennings, V., Elliott, R., Kennaway, R., Glauert, J.: Requirements for a signing avatar. In: Workshop on Corpora and Sign Language Technologies at the 7th International Conference on Language Resources and Evaluation (LREC 2010). pp. 33–136 (2010)
17. Kelly, L.P., Barac-Cikoja, D.: The comprehension of skilled deaf readers. In: Cain, K., Oakhill, J. (eds.) Children's comprehension problems in oral and written language: A cognitive perspective, pp. 244–280. The Guilford Press (2007)
18. Kennaway, R., Glauert, J., Zwitserlood, I.: Providing signed content on the internet by synthesized animation. ACM Transactions on Computer-Human Interaction **14**(3), 1–29 (2007). https://doi.org/10.1145/1279700.1279705
19. Klomp, U.: A descriptive grammar of Sign Language of the Netherlands. Ph.D. thesis, University of Amsterdam (2021)
20. McKee, M., Moran, C., Zazove, P.: Overcoming additional barriers to care for deaf and hard of hearing patients during covid-19. JAMA Otolaryngology–Head & Neck Surgery **146**(9), 781–782 (2020). https://doi.org/10.1001/jamaoto.2020.1705
21. Middleton, A., Niruban, A., Girling, G., Myint, P.K.: Communicating in a healthcare setting with people who have hearing loss. Bmj **341** (2010). https://doi.org/10.1136/bmj.c4672
22. Prillwitz, S., Leven, R., Zienert, H., Hanke, T., Henning, J.: HamNoSys version 2.0: an introductory guide (1989)
23. Prins, M., Janssen, J.B.: Automated sign language (2014), TNO technical report
24. Rayner, M., Bouillon, P., Ebling, S., Gerlach, J., Strasly, I., Tsourakis, N.: An open web platform for rule-based speech-to-sign translation. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. pp. 162–168. Association for Computational Linguistics, Berlin, Germany (2016). https://doi.org/10.18653/v1/P16-2027, `https://www.aclweb.org/anthology/P16-2027`
25. Smith, R.G., Nolan, B.: Emotional facial expressions in synthesised sign language avatars: a manual evaluation. Universal Access in the Information Society **15**(4), 567–576 (2016). https://doi.org/10.1007/s10209-015-0410-7
26. de Vos, C., van der Kooij, E., Crasborn, O.: Mixed signals: Combining linguistic and affective functions of eyebrows in questions in Sign Language of the Netherlands. Language and Speech **52**(2-3), 315–339 (2009). https://doi.org/10.1177/0023830909103177
27. Wauters, L.N., Van Bon, W.H., Tellings, A.E.: Reading comprehension of Dutch deaf children. Reading and writing **19**(1), 49–76 (2006). https://doi.org/10.1007/s11145-004-5894-0
28. Wolfe, R., Cook, P., McDonald, J.C., Schnepp, J.: Linguistics as structure in computer animation: Toward a more effective synthesis of brow motion in American Sign Language. Sign Language & Linguistics **14**(1), 179–199 (2011). https://doi.org/10.1075/sll.14.1.09wol
29. Zhou, Z., Chen, K., Li, X., Zhang, S., Wu, Y., Zhou, Y., Meng, K., Sun, C., He, Q., Fan, W.: Sign-to-speech translation using machine-learning-assisted stretchable sensor arrays. Nature Electronics **3**(9), 571–578 (2020). https://doi.org/10.1038/s41928-020-0428-6
30. Zwitserlood, I.: Synthetic signing. In: The World of Content Creation, Management, and Delivery (IBC 2005). p. 352–357 (2005)

A UI example: avatar translation of a simple question



Choose a translation option:

○ Video translation

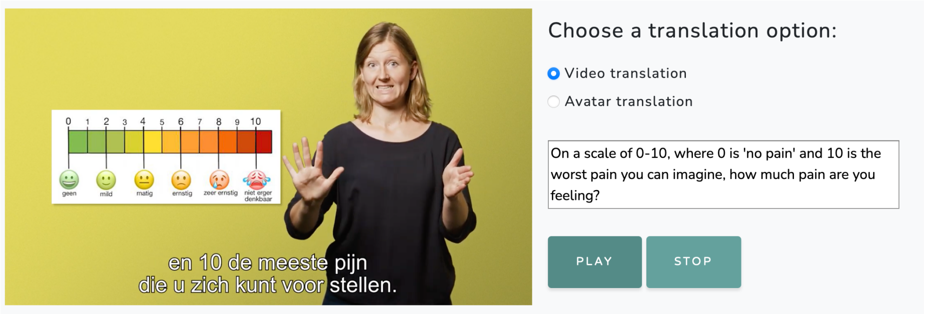
● Avatar translation

Where are you feeling pain?

Adjust speed: +0.0 - +

Gloss: pijn

B UI example: video translation of a complex question



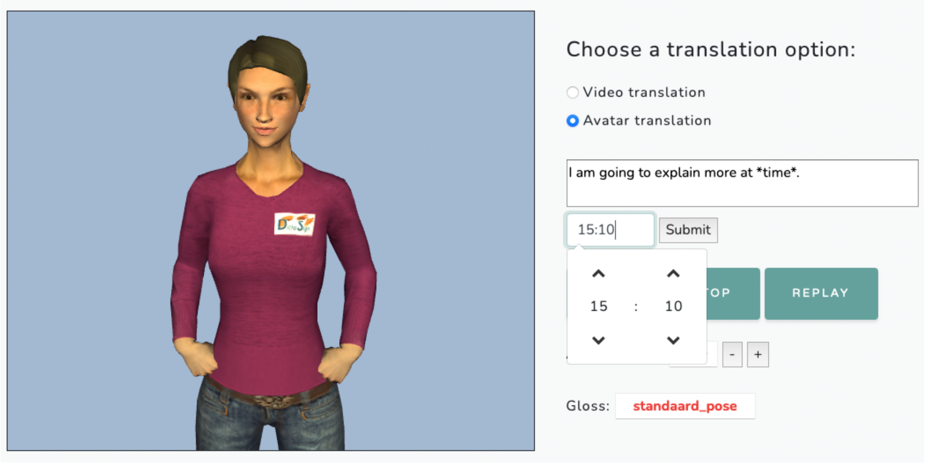
Choose a translation option:

● Video translation

○ Avatar translation

On a scale of 0-10, where 0 is 'no pain' and 10 is the worst pain you can imagine, how much pain are you feeling?

C UI example: translation of a sentence with a variable



Choose a translation option:

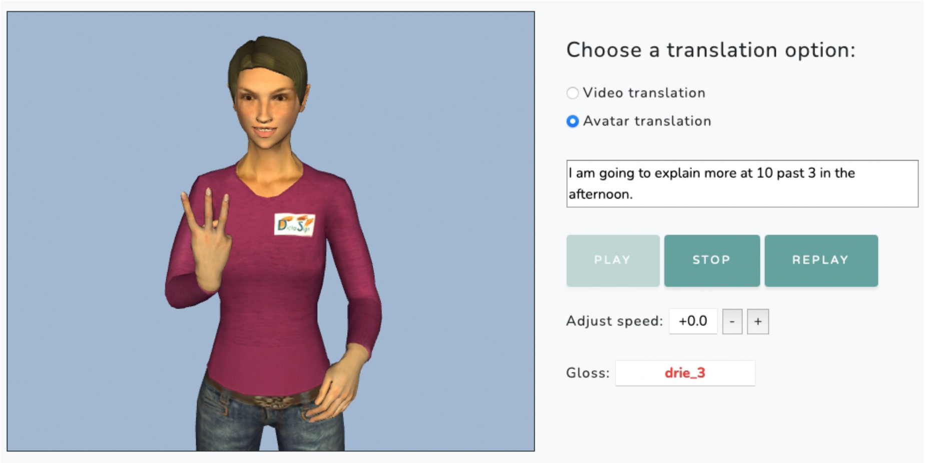
○ Video translation

● Avatar translation

I am going to explain more at *time*.

15:10

Gloss: standaard_pose



Choose a translation option:

○ Video translation

● Avatar translation

I am going to explain more at 10 past 3 in the afternoon.

Adjust speed: +0.0

Gloss: drie_3

A Comparison of the Word Similarity Measurement in English-Arabic Translation Memory Segment Retrieval including an Inflectional Affix Intervention

Khaled Mamer Ben Milad

[1] Swansea University, UK
882922@swansea.ac.uk

Abstract. The aim of this paper is to investigate the similarity measurement approach of translation memory (TM) in five representative computer-aided translation (CAT) tools when retrieving inflectional verb-variation sentences in Arabic to English translation. In English, inflectional affixes in verbs include suffixes only; unlike English, verbs in Arabic derive voice, mood, tense, number and person through various inflectional affixes e.g. pre or post a verb root. The research question focuses on how the TM matching metrics measure a combination of the inflectional affixes when retrieving a segment. If it is dealt with as a character intervention, are the types of intervention penalized equally or differently? This paper experimentally examines, through a black box testing methodology and a test suite instrument, the penalties that TM systems' current algorithms impose when input segments and retrieved TM sources are exactly the same, except for a difference in an inflectional affix. It would be expected that, if TM systems had some linguistic knowledge, the penalty would be very light, which would be useful to translators, since a high-scoring match would be presented near the top of the list of proposals. However, analysis of TM systems' output shows that inflectional affixes are penalized more heavily than expected, and in different ways. They may be treated as an intervention on the whole word, or as a single character change.

Keywords: Arabic inflectional affix, TM retrieval, TM metrics, penalty imposed

1 Introduction

A translation memory is a database that contains translation units which comprise source language segments aligned with their target language translations. When a new (input) text is uploaded for translating, the TM matching and retrieval mechanism computes the string similarity of the input in comparison with the source segments contained in the TM. Then, TM technology leverages translation candidates with the highest similarity to the input segment [1]. However, there is little detailed information on how the matching algorithms assign a score to the matching strings.

Translation and Interpreting Technology Online, pages 125–141,
July 05-07, 2021.
https://doi.org/10.26615/978-954-452-071-7_014

Most previous studies repeat the belief that the TM similarity measurement is based on the Levenshtein [2] distance algorithm (e.g. Simard and Fujita [3]). This similarity measurement uses three basic edit operations (insertion, deletion and substitution) to determine a distance between two strings, then the distance is normalized into a matching score. Hence, the question is how do TM systems measure the matching between two strings? Is the matching measurement based on a comparison word by word? Or, is the measurement computed character by character? For example, if two source segments are identical except for a difference in an inflectional affix, does the algorithm measure a combination of the inflectional affixes as a word intervention or a character intervention?

The researcher's hypothesis is that the TM metrics may compute inflectional verb variations is either as a word intervention, which means that the algorithm regards the inflected form as a totally different word, where the penalty would be expected to be very heavy, or as a character intervention, in which the penalty would be based on the edit type. Hence, we argue that TM similarity metrics could have difficulties detecting inflectional affixes, which would not result in seeing high-scoring TM proposals.

On the other hand, if the TM system were able to undertake a morphological analysis, it would treat the inflectional affix in a different way. However, Macklovitch and Russell [4] pointed out that one of the limitations of TM systems is the failure to recognize inflectional variants. They argue that despite any necessary minor adjustments, a segment that includes an inflected word is still potentially informative. Somers [5] highlighted that a high-matching technique is needed to use linguistic information such as inflection paradigms, synonyms and grammatical alternations in order to improve TM fuzziness. A fuzzy match means a percentage assigned by a TM metric occurs when the input is partially similar to TM source; if the difference is minor, the value is high. If, on the other hand, the difference is significant, the score is low.

In this paper, we aim to investigate the performance of TM similarity algorithms when retrieving inflectional verb-variation sentences in Arabic-to-English translations. To achieve the aims of the study, a special corpus of Arabic source segments and English target segments is provided, in which we apply a number of inflectional verb-variation transformation rules to the Arabic source segments. Test segments were extracted from the corpus and the edit distance metric was used as an analysis tool.

The paper is organised as follows: Section 2 reviews related works related to semantic matching in TMs. In section 3, we present a review of the verb inflectional affixes in Arabic. We describe the experimental methodology in section 4. We summarise the findings in section 5, and discussion of the results in section 6. Lost usability opportunity of highly similar TM proposals is analysed in section 7. Finally, the conclusions drawn from the research are in section 8.

2 Related studies

Due to the limitation of the TM algorithms, various researchers have focused on how to improve semantic matching in TMs. Gupta et al. [6, 7]; Gupta and Orasan [8] offer a semantically enhanced edit-distance method by introducing a paraphrase database into the edit-distance metric during the matching process. The extra paraphrase TM database contains semantic information such as lexical, phrasal and syntactic paraphrases. Paraphrases in the PPDB dataset are extracted using a statistical method. Both automatic and human evaluation have shown that paraphrasing improves TM matching and retrieval

In very recent research, Ranasinghe et al. [9] claim that most of the methods that try to capture semantic similarity in TM were trialled on small databases and are not appropriate for the large TMs normally employed by translators. These researchers, therefore, have introduced an approach that relies on encoding sentences into embedded vectors in order to improve the matching and retrieval process; this means that text similarity is calculated using deep learning (vector representation) rather than texts. The experiment employed the Universal Sentence Encoder for English released by Google [10]. A test was run on English ↔ Spanish languages pairs, using the DGT-TM of the European Commission's translation service. The results showed that universal sentence encoder architectures handle semantic textual similarity better than the edit distance metrics. The approach is language independence and could be employ to any language pair if there are embeddings available for the source language. It appears to be a promising method for the retrieval of a rich semantic similarity, like Arabic.

Further, Tezcan et al. [11] propose developing a "neural fuzzy repair" method by using sub-word-level segmentation in fuzzy match combination to maximise the coverage of source words. This method employs vector-based sentence similarity metrics for retrieving TM matches in combination with alignment-based features on overall translation quality. This method aims to maximise the added value of retrieved matches within the neural fuzzy repair paradigm. A test was run on eight language combinations: English ↔ Hungarian, English ↔ Dutch, English ↔ French, and English ↔ Polish using the DGT-TM. This study reaffirms the usefulness of fuzzy matching based on vector representations to capture semantic relationships between subwords.

3 Review of Arabic verb inflections

The Arabic language is a highly inflected language, and verb inflection (which is Known in Arabic as الأوزان, al-awzaan) is a conjugation process of creating new stems from the root using specific verbal templates. The verb conjugation involves the creation of new stems from the verb's root (the base of the verb form) using specific verbal templates. Neme [12] explains that the combination of a root with a pattern produces an inflected form in which the root signifies a morphemic abstraction for a verb, while the pattern is a template of characters (indices) surrounding the root consonants.

The verb's tense – and other aspects such as gender and number – are generally represented using the rules of inflectional verb morphemes. Tenses are used in either the perfect or imperfect form; the former indicates the past tense while the latter indicates the present or future tense. The language uses a unique inflection system: for example, verbs in the past tense are often designated by suffixes, whereas verbs in the present or future tense are often identified by a prefix. Numbers are classified as plural, dual or singular, with two gender categories, feminine and masculine. The number and gender features can be integrated with the verb's tense and expressed in single-word forms [13].

Another important characteristic of Arabic is that the overwhelming majority of verbs have roots consisting of three characters, in which the position of an inflectional affix (i.e. a character) that shapes the template is positioned either as a prefix or a suffix only, while the affix string may encompass one character or more. Habash [14] states in his book 'Introduction to Arabic Natural Language Processing' that verb inflections have a limited number of patterns: ten basic templates for a three-character root and two templates for a four-character root. This means that the triliteral (three-character-root) verb could be transformed from one template into another template just by attaching a prefix (an initial attachment) or a suffix (a final attachment), while the string of basic form stays as one chunk (no mid-form intervention). In Transformation sub-section (4.3) below, we describe a prefix and suffix combination with a three-character root in order to make different verbal templates.

4 Methodology and Experimental Setup

4.1 Evaluation method

The method of TM systems evaluation, which is further illustrated in the subsection on the experimental setup below, was based on the approach of considering the TM as a 'black-box' component advanced by Simard and Fujita [3]

The test segments were extracted from a corpus. The corpus, which was created by the researcher, was imported into the CAT applications as a TM. Then the test segments (i.e. the input) were uploaded as a document to be translated in the selected CAT tool. As a result, the matching scores of TM proposals offered a similarity measurement.

The goal of the study was to initially test then compare the five representative CAT tools in terms of retrieving inflectional verb-variation sentences in Arabic to English translation. Accordingly, the emphasis was on whether the TM could handle the intervention of inflectional affixes in a linguistic analysis or as an edit distance operation.

4.2 Preparing the experimental database

Finding a corpus including specific inflectional verb-variation sentences in Arabic proved difficult; thus, we created our own corpus in order to build more effective and robust results. The size of the corpus was 45 aligned sentences, with Arabic as the source language of the translation units and English as the target, while the segments' length ranged from 3 to 7 words. The procedure of making the Arabic source segment in the corpus was that the verb-stem was generated from a three-character root, combined with a single character as a prefix or suffix. We selected four templates (i.e., verb stems) to represent the inflectional verb variations. At least three samples were used in each event. We are aware that the corpus created was very small, therefore, we regard this work and the results as preliminary.

4.3 Transformation

For the purpose of the study, the four templates selected were transformed from perfective to imperfective or vice versa by changing their inflectional affix. The change of character led to a change in the verb tense only, while the aspects of the subject remained the same. We explain below the rules of transformation by using the canonical verb (فعل), (do), which is commonly used by Arabic grammarians in creating verb templates:

- Rule 1: The verb template (VT) of the source segment was changed from an imperfective (third person masculine) into a perfective pattern: يفعل (He does)> فعل or فعل (He did). The transformation was made by dropping an initial character ي (a single character prefix), or sometimes by adding a diacritic mark on the final-character لَ. However, the insertion of a diacritic mark is optional in Arabic, and it may be omitted from the text. For example, ‘يشرب الطفل الحليب الطازج صباحا.’ / yashrab altifl alhalib altaazij subahana / ‘The child drinks fresh milk in the morning’. In such example, if the prefix (يـ) is removed (deletion operation), the tense of the sentence changes into past ‘يشرب الطفل الحليب الطازج صباحا ا.’[1] / shrab altifl alhalib altaazij subahana / ‘The child drank fresh milk in the morning’. In the experiment, we removed such prefixes, so that the input string was different from the TM source by a single character. Table 1 below shows the verb template transformation process.
- Rule 2: In contrast to Rule 1, the verb template was changed from a perfective (third person masculine) into an imperfective pattern, فعل or فعلَ (He did)> يفعل (He does), by adding an initial-character ي (a single-character prefix). Table 1 below shows the verb template transformation process.
- Rule 3: The verb template of the source segment was changed from a perfective (third person feminine) into an imperfective pattern, فعلت (She did) > تفعل (She does), by changing a final character ت (a single-character suffix) into an initial character ت (a single-character prefix). Table 1 below shows the verb template transformation process.
- Rule 4: In contrast to Rule 3, the verb template was changed from an imperfective (third person feminine) into a perfective pattern: تفعل (She does) > فعلت (She did). The change was made by changing an initial character ت a (single-character prefix) into a final character ت (a single-character suffix). Table 1 below shows the verb template transformation process.

[1] Track Changes was used for the intervention.

Using an Arabic verb conjugator website,[2] the automated ACON application can conjugate the different templates of the Arabic verb by selecting the root and the type. Table 1 below shows the transformation of four templates in Arabic sentences using edit operations.

Table 1. Transformation of four verb templates in Arabic sentences using edit operations

Rule	Original VT	Morphological intervention	Edit distance	Transformed VT
1	يفعل [He does]	Dropping prefix	Deletion	فعل or [3]فعَل [He did]
2	فعل or [4]فعَل [He did]	Adding prefix	Insertion	يفعل [He does]
3	فعلت [She did]	Shifting suffix into prefix	Substitution	تفعل [She does]
4	تفعل [She does]	Shifting prefix into suffix	Substitution	فعلت [She did]

After applying the rules listed above each sentence of the test underwent a transformation, which converted linguistically the imperfective pattern of the verbs in the original sentences into the perfective patterns or vice versa using one type of edit operation. Then, the modified test segments, which were used as a document to be translated, were run against the TM corpus which included the original segments.

The verb templates in Table 1 above that represent the verb inflections in Arabic have the structure of the research query; the transformation of verb templates represents the rich morphology of the language; the edit operations potentially represent the similarity measurement used by translation memory systems.

4.4 Experiment with pre-translation

Having processed the test segments, they were then submitted to the CAT applications as files to be translated. If we had to translate again a segment from the source language, the match would obviously be 100%. The translation project in each CAT application was based on the corpus created as a TM file that included the original segments; to make the comparison as fair as possible, the same input text (test segments) was uploaded as a file for translation in the five CAT tools. Then, a pre-translation was processed to gain the TM matching scores.

[2] ACON, the Arabic Conjugator - conjugate Arabic verbs online (baykal.be)

[3] The diacritic mark of *fatha*

[4] The diacritic mark of *fatha*

The input text, which contained 45 segments, was translated by five CAT tools: Déjà Vu X3 (hereafter referred to as DVX3);[5] OmegaT;[6] memoQ 9.0;[7] Memsource Cloud;[8]and Trados Studio 2019.[9] These CAT tools, widely used by professional translators [15], produced fuzzy matches that were analysed according to their results. As the test segments and TM source were identical except for a difference in an inflectional affix, it was desirable for the TM similarity metrics to produce a very high score which could be presented at the top of the list of proposals presented to the translator.

5 Findings

This section displays the results obtained from the TM systems' attempts to retrieve matches for the test segments. We assumed that scores at the higher end are better, for example 95% is better than 80%.

5.1 Déjà Vu X3 Scoring

The matches retrieved by DVXwere found to occupy a consistent band according to the length of the test segments and whether they contained an inflectional affix intervention (deletion, insertion, or substitution). The matching scores decreased in a consistent way as the number of words in the segment decreased and ranged from 67% to 86%. Figure 1 (below) illustrates the fuzzy matching scores (three samples were used in each event) that each segment length (SL) supplied due to their inflectional affix combination (inserting a one-character prefix, deleting a one character prefix, and shifting one character into suffix or vice versa).

The figure below clearly shows that DVXtreated the test segments equally regardless of the type of inflectional affixes intervention. Further, the retrieved matches of three-to-seven-word segments were distributed among the different fuzzy bands. For example, 67% provided a low fuzzy score (i.e. a 33% penalty per one-edit operation), while for seven-word segments, 86% provided a high fuzzy score (i.e. a 16% penalty per one-edit operation, or approximately one word in seven). This means that TM users may not see proposals of high fuzzy matches for short sentences that have just a single character difference.

5.2 memoQ 9.0 scoring

The scores of memoQ were categorised in two phases. The matching scores of memoQ were derived from two different ranges: a low match range and a high match

[5] https://atril.com/

[6] https://omegat.org/

[7] https://www.memoq.com/memoq-versions/memoq-9-5

[8] https://www.memsource.com/

[9] https://www.trados.com/products/trados-studio/

range. The five-, six- and seven-word segment routines were in the low fuzzy range, while the three- and four-word segments were given a relatively high fuzzy range whether these segments contained an inflectional affix intervention (deletion, insertion, or substitution). The match scores ranged from 77% to 91%. Figure2 (below) illustrates the different range of matches for each segment length (SL) provided.

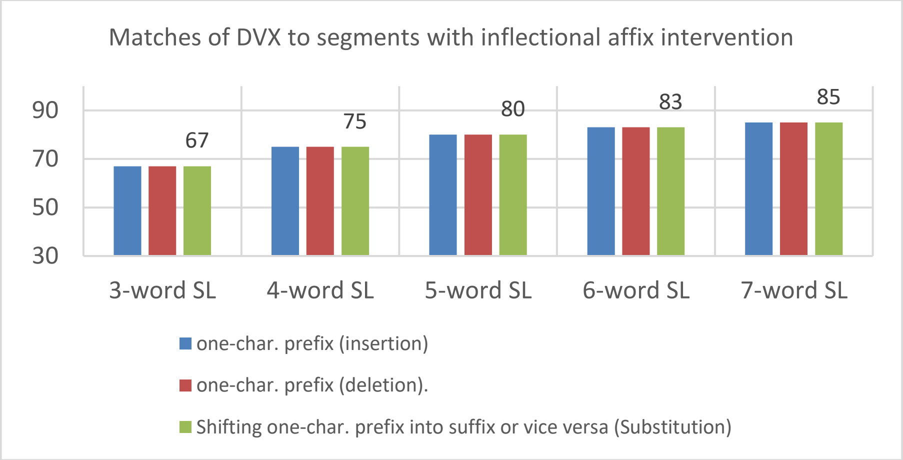


Fig.1. DVX matching scores for 3-to-7-word segment lengths (SL) with an inflectional affix intervention.

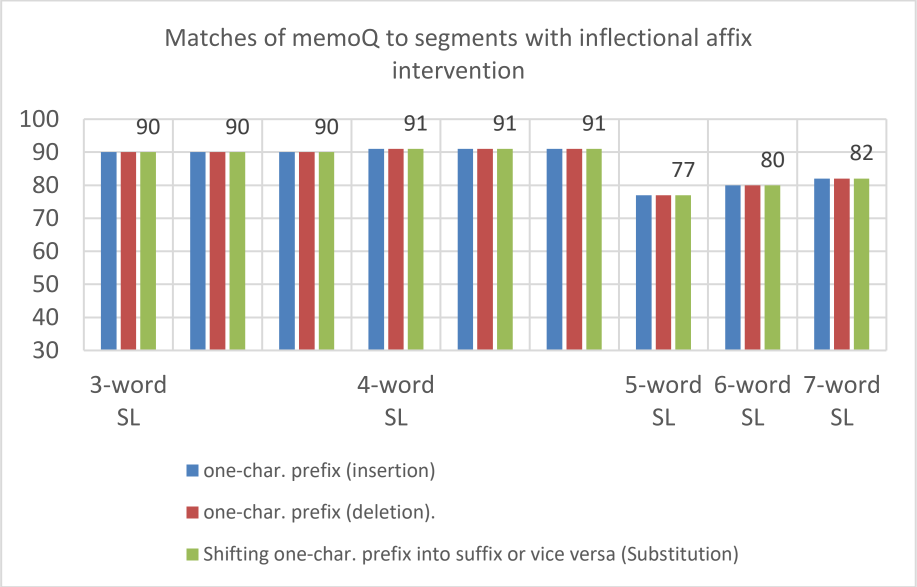


Fig. 2. memoQ matching scores for 3-to-7-word segment lengths (SL) with an inflectional affix intervention

As the figure above shows, the matches of three- and four-word segments with an inflectional affix were retrieved in a high fuzzy match. For example, the three-word and four-word segments were provided with a 90% and 91% match, respectively (i.e., a 10% and 9% penalty). In terms of the segments of five words and above, the scores unexpectedly matched lower regardless of the edit operation. For example, five-word segments provided a match of 77% (i.e., a 23% penalty).

This suggests that the retrieval of segments of five words or above was based on the number of words, while the retrieval of three- and four-word segments was not. It seems that the measurement was based on the total number of characters. This may explain the difference in the matching levels: the character-based measurement produced considerably better results. As a result, the short segments would be offered in a high fuzzy band, while longer segments would be scored lower, although in all cases the difference was just a single character.

5.3 Memsource Cloud scoring

The TM system of Memsource retrieved the test segments in an inconsistent range of scores. Thus, the experiment used the filter feature in the system's setting to sort the source's shortest segment first, which was based on the number of characters. When observing the fuzzy matches, the scores appeared to decrease as the total number of characters in the segment fell regardless of how many words a segment contained. Similarly, when the source was sorted according to the principle of the longest first, the matches appeared to increase as the total number of characters in the segment increased. As A result, the matches appeared to rely in the first place on the total number of segment characters, and in the second place on the position of the edit operation. Further, the match values decreased as the total number of characters decreased; the length of segments varied from 16 to 49 characters (i.e., both characters and whitespaces), while the match scores varied between 73% and 98%. Due to these scattered scores, the matches illustrated in Figure 3 are presented as a chart, using a line with markers: the markers represent the inconsistency of scores, while the lines represent the impact of the segment length.

As Figure 3 shows, it is obvious that the retrieval of segments with a one-character prefix were given high percentages, whereas the operation of shifting a one-character prefix into a suffix position, or vice versa, was assigned a lower fuzzy band.

For example, the matches of segments ranging from 49 to 16 characters, produced by inserting a one-character prefix, ranged from 98% to 94%, whereas segments ranging from 49 to 76 characters. produced by deleting a one-character prefix, also scored between 98% and 94%. Shifting a one-character prefix into a suffix position, or vice versa, produced match scores in the lower fuzzy band. For instance, segments ranging from 46 characters to 18 characters produced scores between 90% and 73% when a one-character prefix was changed into a suffix, whereas segments ranging from 46 characters to 19 characters produced scores between 91% and 74% when a one-character suffix was changed into a prefix.

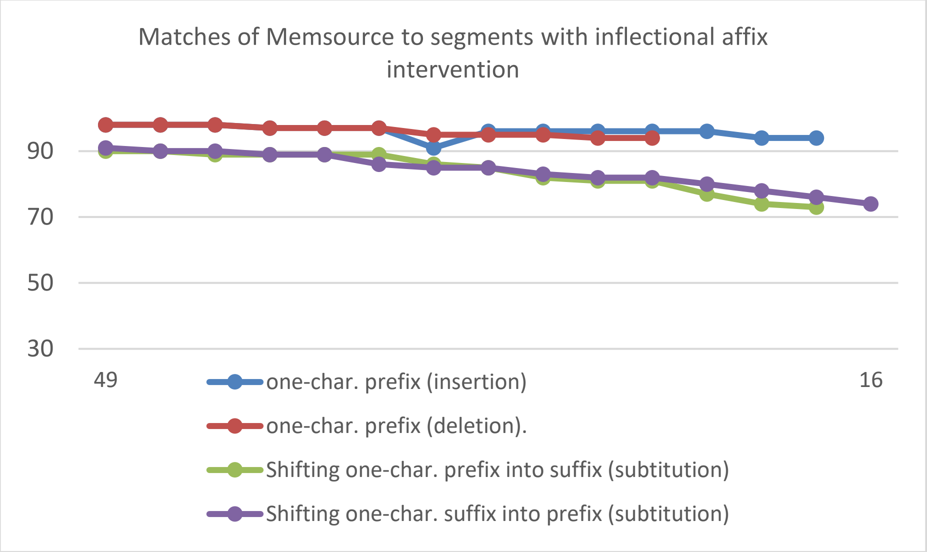


Fig. 3. Memsource matching scores for a segment 49-16 characters long due to changes to an inflectional affix.

The explanatory hypothesis is that, on the one hand, a one-character prefix was dealt with as a one-edit operation, while changing a one-character prefix into a suffix, or vice versa, was treated as a two-edit-operation. On the other hand, editing a one-character prefix occurred on the word-initial position, while changing a one-character prefix into a suffix, or vice versa, occurred on the word-initial and word-final positions. This suggests that the matching metrics dealt with the impact of a prefix combination in a different way to that of a suffix combination. As a result, the retrieval of segments with an inflection affix would be offered at a high fuzzy level under specific conditions. However, further research is needed to confirm this hypothesis since this study is based on the number of words in segments.

5.4 OmegaT scoring

The fuzzy matches provided by OmegaT were relatively high; however, they dropped gradually as the segment became shorter, whether it contained a deletion, insertion or substitution operation. The matching scores consistently related to the segments' word length – the scores ranged from 83% to 92%. Figure4 (below) shows the matching values for each segment length (SL) according to the editing of an inflectional affix.

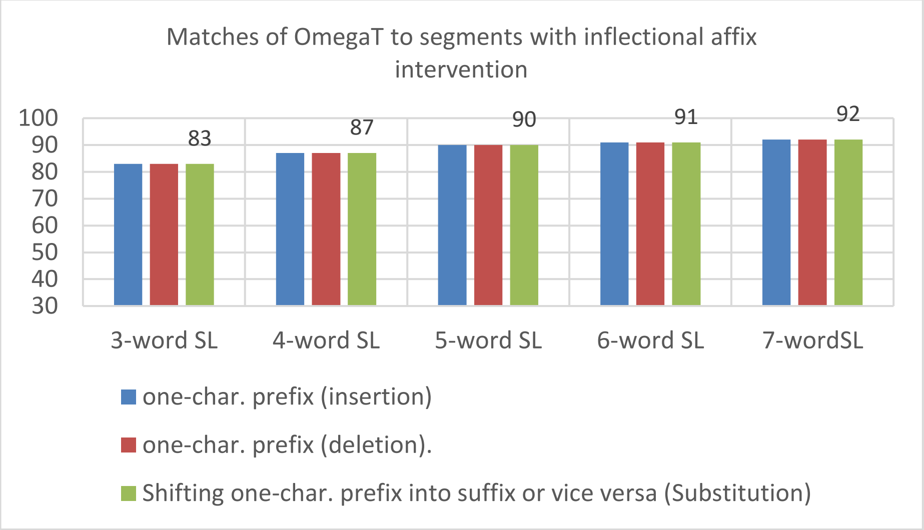


Fig. 4. OmegaT matching scores for 3-to-7-word segment lengths (SL) with an inflectional affix intervention.

As Figure 4 clearly shows, OmegaT's matching metrics dealt with the different ways of editing the inflectional affix in the same fashion, retrieving four- to seven-word segments in a high fuzzy band; only the three-word routine was placed in the middle fuzzy band. This means that OmegaT would retrieve segments with an inflectional affix – except for a three-word routine – in a high fuzzy band, which would be very useful from the perspective of translators.

5.5 Trados Studio 2019 scoring

The matching scores produced by Trados Studio also fell steadily as the segment length became shorter, whether these segments contained a deletion, insertion or substitution operation. The matching values were consistently related to the segment's word length. The match scores ranged from a 78% to 91%. Figure 5 (below) displays the matching values for the retrieval for each segment length (SL).

It can be seen that Trados Studio dealt with the retrieval of segments with an inflectional affix in the same way regardless of the type of character-edit operation involved. The matches were distributed between middle and high fuzzy bands, where the three- and four-word segments matched 78% and 83%, respectively (i.e. in the middle fuzzy band), and the five- six- and seven-word segments scored in a high fuzzy band. This means that TM users would not see three- and four-word segments with only a one-character difference in the high fuzzy band range.

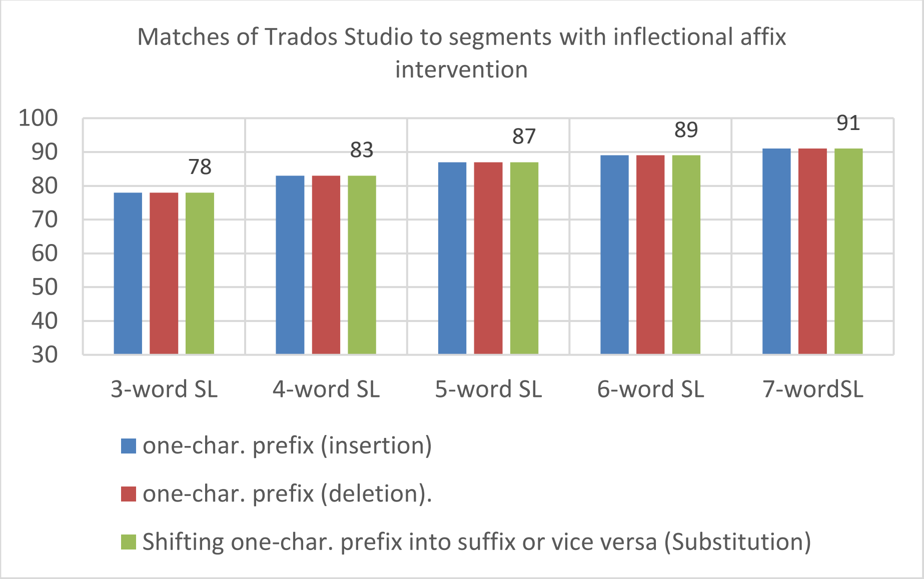


Fig. 5. Trados Studio matching scores for 3-to-7-word segment lengths (SL) with an inflectional affix intervention.

The results showed that the various TM systems differed in their handling of diacritic marks. First, the algorithm of DVX, OmegaT, Trados Studio systems and the scoring of five- to seven-word segments in memoQ, which produced consistent matches according to the segments' word length, did not appear to be influenced by the insertion or removal of diacritic marks – the matches retrieved were the same. Secondly, the metrics of Memsource and the scoring of three- or four-word segments in memoQ, whose character-based algorithm provided inconsistent values, were affected by a combination of diacritic markers. When calculating segments with and without a diacritic mark using a Levenshtein website,[10] the URL estimated a diacritic marker as a one-edit distance. Hence, a diacritic mark was treated as equal in weight to a one-character intervention in character-based metrics.

6 Discussion

The experiment's findings show that the TM systems treated a combination of inflectional affixes in different ways: the TM matching algorithms dealt with the morphological combination as an intervention on the whole word, as a single character change, or according to the position of the intervention. In all the systems, however, it appears that segment length had a bearing on the results.

[10] https://planetcalc.com/1721/

These findings prompted a comparative analysis of each TM's retrieval of fuzzy bands. This was accomplished by using the length of each segment and the affix position and type as independent variables.

Turning to the DVX results first, it seems that the TM system's algorithm dealt with the inflectional affix as an intervention on the whole word. To account for this, a procedure calculating the surface form of the strings was used. In five-word segments, for example, DVX provided an 80% match (i.e., a 20% penalty). This may be explained by the fact that the algorithm estimated that a four-word string was identical to a five-word string, while a one-word string was non-similar (i.e.$\frac{\mathbf{4}}{\boldsymbol{SL}\,\mathbf{5}}$ = $\frac{\mathbf{80\%}}{\mathbf{100}}\ \boldsymbol{identical\ vs.}\frac{\mathbf{1}}{\boldsymbol{SL}\,\mathbf{5}} = \frac{\mathbf{20\%}}{\mathbf{100}}$ non-similar). This implies that the DVX metrics recorded the edit operation (i.e., the inflectional affix) as an intervention on the whole word, resulting in low scores for segments that have a small number of words and an increase in scoring for longer segments.

The reason behind the OmegaT and Trados Studio results could be that their TM similarity algorithms are not only based on the number of words but also employ a specific mechanism for an individual edit operation (i.e., a single-character intervention) to measure the segments' similarity. In five-word segments, for example, any type of character editing (i.e., insertion, deletion, or substitution) was penalised 10% and 13% in OmegaT and Trados Studio, respectively; however, the matching scores provided were consistently in line with the segment's word length whatever the number of characters, which resulted in decreasing scores for short segments and increasing scores for longer ones. However, a comparison of the matching mechanisms of the two systems shows that OmegaT outperformed Trados Studio; the lowest match was scored 83% by OmegaT and 78% by Trados Studio, whereas the highest scores were 92% and 91% for OmegaT and Trados Studio, respectively.

As for the scores of memoQ, in terms of consistent scores, the system algorithm seems to use g an internal mechanism to compute a combination of inflectional affixes in segments of five words or above. The mechanism produced the lowest average scores for the five-, six- and seven-word routines compared with the other systems that provided consistent scores. With a five-word routine, for example, memoQ supplied a 77% match (a 23% penalty) whatever the type of character editing. The penalties imposed by DVX, Trados Studio and OmegaT were 20%, 13% and 10%, respectively. The penalty imposed by memoQ was the heaviest. This means that the similarity algorithms in memoQ, where the measurement was word-based, imposed the heaviest penalty due to the character combination. In terms of the inconsistent matches (i.e., the three- and four-word segments), the matches were retrieved with high percentages despite the short segment length. This may be explained by the fact that the recall was based on the number of characters.

Memsource's matches, which were apparently inconsistently produced according to the number of characters, showed that the retrieval of segments with the insertion or removal of a one-character prefix gave high percentage scores, while the operation of substituting one character produced a lower percentage. It seems that Memsource's retrieval mechanism penalised a prefix combination relatively lightly. This was calculated not according to a linguistic analysis but from the perspective that a prefix combination may cause less damage to the word form than a suffix combination. As a result, in some cases, the TM matching measurement performed well when a one-character prefix (i.e., inflectional affix) was inserted or removed, but not a one-character suffix.

Overall, the different tools appear to have different routines for handling such inflectional affix interventions. Although none of them is fully satisfactory, especially for short segments, Memsource outperformed the other systems when the intervention of an inflectional affix was a prefix only. The metrics of memoQ penalised the heaviest when the system provided consistent matches. In all the TM systems, the matching scores reduced as the length of the segments decreased but it was seen most clearly in the systems that produced consistent matches. To bear in mind, the study used a very short root – a three-character word including a single character combination, the retrieval of a longer base-form including a prefix or suffix combination may be scored differently by TM systems' algorithms.

To summarise, the TM matching measurements failed to recognise inflectional affixes. This outcome is in line with the results of the studies conducted by Macklovitch and Russell [4] and Planas and Furuse [16], which found that one of the limitations of TM systems is their inability to recognise inflectional variants when retrieving stored data. The current study has provided further experimental evidence, gathered from the scores supplied by five CAT applications, showing that TM matching metrics are not good at distinguishing morphological combinations.

7 Lost usability opportunity

From a usability point of view, the test results show that, although the translator would potentially spend less time and effort editing the inflectional verb-variation segments, they could miss out on seeing those TM proposals because of their low scores. What the users of TM would expect – from a translator's perspective – is that TM algorithms would retrieve inflectional verb-variation segments with a very high match score (i.e., a range of high fuzzy or 85%-94%) since these would need only one edit operation to be identical to the input text. The impact of high fuzzy matches appears in the translation cost. Contrary to this expectation, however, it appears that a translator working with short segments will not be shown a high but a low fuzzy proposal, which may result in the proposals being lost. Hence, the project manager, when preparing a report, may produce inappropriate fuzziness percentages for the translation of a text with a rich morphology including segments with inflectional verb variations, and the price they quote for the translation will consequently be higher than it should be. Table 2 shows the bands of fuzzy matches, according to Studio Trados,[11]produced for the test segments reported by each TM system.

[11] Fuzzy match grids in SDL Trados Studio | Signs & Symptoms of Translation (signsandsymptomsoftranslation.com)

Table 2. Fuzzy match bands as computed by each TM system

Fuzzy bands	**Range of scores**	**DVX**	**memoQ**	**Mem-source**	**OmegaT**	**Trados Studio**
Nearly exact match	95% - 99	0	0	20	0	0
High fuzzy band	85% - 94	12	24	26	48	36
Middle fuzzy band	75% - 84	36	36	10	12	24
Low fuzzy band	50% - 74	12	0	4	0	0
No match	0 - 49%	0	0	0	0	0
Total	Total	60	60	60	60	60

Table 2 displays the ways in which the TM systems differed in fuzzy-match distribution. OmegaT showed a significantly higher number of matches for the high fuzzy band (85-99%), followed by Memsource, while DVX ended up with a significantly smaller number than the other bands. The fuzzy matches varied in distribution according to the different TM systems:

- OmegaT retrieved only 12 out 60 segments, representing 20%, in a lower fuzzy band. These results appear to be the best.
- Memsource retrieved 14 out of 60 segments, representing 24%, in a lower fuzzy band; however, the high fuzzy scores were mainly produced when the intervention was a prefix.
- Trados Studio retrieved 24 out of 60 segments, representing 40%, in a lower fuzzy band.
- memoQ retrieved 36 out of 60 segments, representing 60%, in a lower fuzzy band.
- DVX retrieved 48 out of 60 segments, representing 80%, in a lower fuzzy band. These results are the worst.

As mentioned above, because the fuzzy match levels play a significant role in the calculation of translation costs, these results would have a definite impact on the discount applied to texts that are rich in morphological combinations. Preventing segments that include an inflection affix from ranking as a high fuzzy match would therefore impact the efficiency, consistency and cost of a translation.

8 Conclusion

The overall conclusion drawn from the results of testing the retrieval of TM sources for a text that is rich in morphological combinations is that all the selected systems revealed a deficiency when it came to identifying inflectional affixes, although OmegaT and Memsource returned more than three-quarters of segments in the high fuzzy band, and memoQ produced considerably better scores to short segments than longer segments. The overall matching scores appeared to be based purely on the string of surface forms and the internal machinery of each system's algorithm, without any linguistic analysis. Hence, the findings substantiate the proposals that implementation of deep learning and vector representations would help capture semantic textual similarity for TM matching. The outcome shows that an inflectional affix intervention was treated as either an intervention on a whole word or a single character change. Consequently, the high matching of retrieved inflectional verb-variation segments in an Arabic-to-English translation would depend on the segment length and the position of the intervention. Further work is needed to extend the investigation to other morphologically rich languages, different positional affixes and longer string formations such as a noun derivation. The findings substantiate the proposals that implementing of encoding sentences into embedded vector should be incorporated into similarity metrics of TM systems.

Acknowledgments

This work is a part of the Ph.D. research at Swansea University, UK. The author would like to extend special thanks to supervisors Prof. Andrew Rothwell and Dr. Maria Fernandez Parra for their helpful comments on the content of this paper. The PhD research programme is funded by High Education Ministry, Libyan Government.

References

1. Vázquez, L. M.: An empirical study on the influence of translation suggestions' provenance metadata. (2012).
2. Levenshtein, Vladimir I.: Binary codes capable of correcting deletions, insertions, and reversals. In Soviet physics doklady, vol. 10, no. 8, pp. 707-710. (1966)
3. Simard, Michel, and Fujita, A.: A poor man's translation memory using machine translation evaluation metrics. Proceedings of the 10th Conference of the Association for Machine Translation in the Americas: Research Papers. (2012)
4. Macklovitch, E., and Russell, G: What's been forgotten in translation memory. Conference of the Association for Machine Translation in the Americas. Springer, Berlin, Heidelberg, (2000)
5. Somers, H.: Translation memory systems. Benjamins Translation Library 35. (pp.31-48) (2003):
6. Gupta, R, Orăsan, C., Zampieri, M., Vela, M., and Van Genabith, J.: Can Translation Memories afford not to use paraphrasing?. In Proceedings of the 18th Annual Conference of the European Association for Machine Translation. (2015)
7. Gupta, Rohit, Orăsan, C., Liu, Q., and Mitkov, R.: A Dynamic Programming Approach to Improving Translation Memory Matching and Retrieval Using Paraphrases. In International Conference on Text, Speech, and Dialogue (pp. 259-269) Springer, Cham (2016)

8. Gupta, R., Orăsan, C.: Incorporating Paraphrasing in Translation Memory Matching and Retrieval. In: Proceedings of the 17th Annual Conference of the European Association for Machine Translation (pp. 3–10) EAMT-2014. Dubrovnik, Croatia: European Association for Machine Translation. (2014)
9. Ranasinghe, Tharindu, Orăsan, C., and Mitkov, R.: Intelligent Translation Memory Matching and Retrieval with Sentence Encoders. arXiv preprint arXiv:2004.12894 (2020).
10. Cer, Daniel, Yang, Y., Kong, S., Hua, N., Limtiaco, N., St John, R., Constant, N, Guajardo-Cespedes, M., Yuan, S., Tar, C. and Strope, B: Universal sentence encoder for English. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (pp. 169-174) (2018)
11. Tezcan, A., Bulté, B. and Vanroy, B.: Towards a Better Integration of Fuzzy Matches in Neural Machine Translation through Data Augmentation." In Informatics, vol. 8, no. 1, p. 7. Multidisciplinary Digital Publishing Institute (2021)
12. Neme, A. A.: A lexicon of Arabic verbs constructed on the basis of Semitic taxonomy and using finite-state transducers. In WoLeR 2011 at ESSLLI International Workshop on Lexical Resources. (2011)
13. Habash, N. and Rambow, O.: Morphophonemic and orthographic rules in a multi-dialectal morphological analyzer and generator for arabic verbs. In International symposium on computer and arabic language (iscal), riyadh, saudi arabia. (2007)
14. Habash, N. Y.: Introduction to Arabic natural language processing. Synthesis Lectures on Human Language Technologies 3, no. 1 (pp. 1-187) (2010)
15. Moorkens, J. and O’Brien, S.: Assessing user interface needs of post-editors of machine translation." In Human issues in translation technology (pp. 127-148) Routledge (2017)
16. Planas, E. and Furuse, O.: Formalizing translation memories. In Machine Translation Summit VII, vol. 1999 (pp. 331-339) (1999)

Audiovisual Translation through NMT and Subtitling in the Netflix Series *Cable Girls*

Lucía Bellés-Calvera[1][0000-0002-1329-6395] and Rocío Caro Quintana[2][0000-0003-2275-2679]

[1] Universitat Jaume I
lucia.belles@uji.es
[2] University of Wolverhampton
R.Caro@wlv.ac.uk

Abstract. In recent years, the emergence of streaming platforms such as Netflix, HBO or Amazon Prime Video has reshaped the field of entertainment [1], which increasingly relies on subtitling, dubbing or voice-over modes [2] [3]. However, little is known about audiovisual translation when dealing with Neural Machine Translation (NMT) engines. This work-in-progress paper seeks to examine the English subtitles of the first episode of the popular Spanish Netflix series *Cable Girls* and the translated version generated by Google Translate and DeepL. Such analysis will help us determine whether there are significant linguistic differences that could lead to miscomprehension or cultural shocks. To this end, the corpus compiled consists of the Spanish script, the English subtitles available on Netflix and the translated version of the script. As regards data analysis, errors have been classified following the DQF/MQM Error typology and have been evaluated with the automatic BLEU metric. Results show that NMT engines offer good-quality translations, which in turn may benefit translators working with audiovisual entertainment resources.

Keywords: Audiovisual translation, Neural Machine Translation (NMT), Errors.

1 Introduction

Over the past few years, the rise of Netflix, HBO, Amazon Prime Video and other streaming platforms has made it necessary to rethink entertainment media [1]. Accessibility to their catalogues not only offers the audience the opportunity to choose among a variety of films, series, documentaries and other audiovisual resources but also to make use of subtitling and dubbing options [2, 3]. However, even though these audiovisual translation practices are meant to meet the needs of different markets and users [4], the quality of the translation may be affected by errors when translators are given tight deadlines. Machine Translation (MT) is widely used in the translation industry, especially in technical fields because the texts tend to be repetitive, and studies have shown that it increases translators' productivity [5, 6] by post-editing the MT output. However, despite the fact that platforms like Netflix announced that they are using MTPE in their subtitling workflows three years ago, research on this topic is

Translation and Interpreting Technology Online, pages 142–148,
July 05-07, 2021.
https://doi.org/10.26615/978-954-452-071-7_015

still scarce in creative fields, such as literary or audiovisual texts. It might be assumed that NMT will not work when dealing with Audiovisual Translation due to its time and character constraints, and especially in media entertainment where cultural aspects are prevalent. Given that the quality of NMT, although still not at the human level [7, 8], is improving every day, some issues need to be considered. Would it be beneficial for audiovisual translators to use MT and post-edit the texts? Or is scratch translation still the best solution for this field?

In any commercial deployment of MT in a subtitling workflow, a bespoke engine would be used. In fact, there are already subtitling specialised MT systems available in the market like AppTek, Omniscien and XL8. However, growing volumes of audiovisual content, short turnaround times or lack of access to this type of engines are some of the challenges novice translators need to overcome. These issues have been addressed in previous studies where MT may serve as a possible solution [9] [10]. Numerous publications arised from the SUMAT project, a large-scale EU-funded project that inspected the creation of high-quality parallel corpora of subtitles through MT [10] [11]. Matusov et al. [12], for example, analysed improvement in productivity after integrating MT in audiovisual translation.

This ongoing project aims to ascertain the quality of Google Translate and DeepL translations (i.e. open MT resources) when compared to the subtitling of TV series in the source language. On this account, the current study draws from the following research questions: RQ(1) *How do English subtitling and translations from NMT differ from the source text in Spanish? What types of errors can be found?* and RQ(2) *Does the integration of MT on the audiovisual translation workflow benefit translators?*

The section below delves into the methodological procedure followed in this study. Later on, the discussion of the preliminary results as well as the conclusions and next steps of this ongoing project will be provided.

2 Methodology

The corpus under study revolves around the Spanish Netflix original series called Cable Girls. This drama, premiered in 2017 and set in the 1920s, tells the story of four women working as operators for the National Telephone Company at a time of social changes. The first season consists of eight episodes, with a length of 47 to 64 minutes.

For the compilation of this small corpus, the focus has been on the first 10 minutes of the first episode of the first season released in Spanish. The Spanish script and the official English subtitles incorporated in the streaming platform have been transcribed from Netflix [13] and examined for the purpose of this preliminary study. In addition, the Spanish transcript was translated with Google Translate [14] and DeepL [15] to analyse the quality of these NMT engines.

Google Translate is an MT engine that provides the translation of texts and files into more than 100 languages, including English, Spanish, Greek, Belarusian, Afrikáans or Chinese [14]. The fact that Google offers these services has caught the attention of scholars who have been concerned with error analysis on MT output.

Evidence may be found in Trzaskawka's study [16], which explored the accuracy of this tool in the translation of contracts in English and Polish. Issues related to the quality of the translation output have also been explored in specialised areas, such as literature [17] and scientific writing [18]. However, research on entertainment media seems to be scarce, with studies delving into the dubbing and subtitling of TV series [19] and documentaries [1].

DeepL Translator [15] is an NMT software developed in 2016 with the aim of producing high-quality translated texts. At the moment, DeepL works with more than 20 languages and also offers a formal/informal register for their translations. DeepL has also caught the attention of researchers and several studies compare its quality to other MT engines like Google Translate, Yandex or Microsoft Translator [20, 21, 22].

The quality of the machine-translated texts has been assessed manually following the DQF/MQM Error Typology [23] – the integration of DQF (Dynamic Quality Framework) [24] and MQM (Multidimensional Quality Metrics) [25] – paying attention to the categories labelled as Accuracy and Fluency. For this manual evaluation, 153 segments containing 7 words on average were examined by two annotators with experience in translation (i.e. post-editing) and linguistics. The translated texts were then analysed automatically with the BLEU metric [26], using the original subtitles as the human translation and the NMT output from Google Translate and DeepL.

3 Evaluation: Preliminary results

3.1 Manual evaluation

A total number of 153 segments were analysed manually following the DQF/MQM Error Typology. The most common errors were related to Fluency, Accuracy and Style. The distribution of errors in Google Translate and DeepL are presented in Table 1.

For this ongoing study on audiovisual translation, namely in subtitling, the character constraint – which entails 70 characters distributed in two lines and a maximum on-screen duration of 6 seconds, has not been analysed on the grounds that Google Translate and DeepL are not specialised systems in subtitling. Instead, the focus has been on the quality of the translation. Therefore, as noted in Table 1 above, the manual analysis of the output taken from both engines differs to a great extent. The findings reveal that only 15 errors have been identified in DeepL (10%), as opposed to Google Translate, where meaning was not properly conveyed in 41 segments (27%).

Table 1. Distribution of errors

Category	Number of errors		Sub-category
	Google Translate	DeepL	
Fluency	20	4	Grammar Grammatical register Inconsistency
Accuracy	14	10	Mistranslation Addition Over-translation
Style	5	1	Unidiomatic Awkward
Other	2	0	Culture-specific reference Tone
TOTAL	**41**	**15**	

Most errors in both engines have to do with Fluency and Accuracy. The number of fluency errors is higher in Google Translate, with a total of 20, and only 4 out of 15 in DeepL. Some examples of fluency errors can be seen in Table 2.

Regarding Accuracy errors, DeepL seems to perform better than Google Translate. Only 10 accuracy errors were spotted in DeepL, while these amount to 14 in Google Translate. Some accuracy errors are illustrated in Table 3.

Table 2. Fluency errors

Original	**English Translation**	**Error**
¡Corre!	Come on!	**Runs**! (Google Translate)
Como grites, te juro que te mato.	If you shout, I swear I'll kill you.	**As** you scream, I swear I will kill you (Google Translate)
Pues lo lamento, no se encuentra entre las preseleccionadas.	I'm sorry, you're not on the short list.	Well, I'm sorry, **she**'s not among the shortlisted. (DeepL)

Table 3. Accuracy errors

Original	**English Translation**	**Error**
Tú no te metas. ¡No te metas!	You stay out of this! Stay out of this!	**You do not mess. Do not mess!** (Google Translate)
600 km para poder estar aquí ahora. / -550.	Six hundred kilometers to get here. / Five hundred and fifty.	600 km to be here now. / 550. **550**. (DeepL)
A continuación, tenemos dos plantas para las salas de máquinas.	Next, two floors with the machine rooms.	Next we have two **plants** for the engine rooms. (Google Translate)

These findings suggest that efforts should still be devoted to refine Fluency and Accuracy in MT engines, as they still not work at the human level. In order to improve the quality of NMT outputs, more corpora should be processed.

3.2 Automatic evaluation

The quality of the texts was evaluated with the automatic metric BLEU [26] using the online BLEU score evaluator from Tilde [27]. Thus, the English subtitles employed in the Netflix platform were compared with the outputs generated by Google Translate and DeepL. The BLEU score for Google Translate is 36.44, in contrast to DeepL, which rises up to 40.79. Although these findings are not conclusive due to the size of the sample, DeepL appears to achieve better results than Google Translate when it comes to the translation of audiovisual resources.

4 Conclusions and further research

The research questions attempted to determine the quality of the translations provided by Google Translate and DeepL when dealing with audiovisual media. Hence, the Cable Girls series script in the source language was compared with the MT outputs from Google Translate and DeepL.

As to RQ(1), the findings suggest that the most common errors occur at Fluency and Accuracy levels. In addition, the results show that DeepL outperforms Google Translate in both manual and automatic evaluation.

With regard to RQ(2), the next steps of the project will delve into translators' post-editing efforts: is it useful to use MT for audiovisual texts? In this vein, technical, temporal, and cognitive variables will be considered to prove whether these efforts are higher or lower when integrating MT tools. Accordingly, an eye-tracking device and a keystroke logging tool will be employed.

Limitations in this study should be acknowledged. The small size of the corpus compiled for this preliminary study may affect the validity of the generalisations presented here. Nonetheless, it should be noted that the corpus will be expanded in the near future. Moreover, the MT engines that were used are not trained on subtitling and may contain an enormous amount of noise. DeepL and Google Translate were used to emulate the experience of freelance translators using general MT. Notwithstanding, the use of these MT engines could have a negative impact on translation quality as the length of the segments, a relevant feature in subtitling, is not taken into consideration.

Further research could also focus on other audiovisual resources, including documentaries or realities. Such examination would prove the efficiency of Google Translate in specialised and non-specialised contexts or the quality of other machine translation software like DeepL in audiovisual domains. Other lines of the proposal presented here could involve the role of MT in the translation of humour and cultural aspects, which are prolific in entertainment media.

References

1. Costan Davara, G.: Audiovisual Translation: Subtitling Netflix documentary â Black Hole Apocalypseâ. (PhD dissertation). Università degli Studi di Padova (2020).
2. Oh, K., Noh, Y.: The actual condition and improvement of audiovisual translation through analysis of subtitle in Netflix and YouTube: focusing on Korean translation. Journal of Digital Convergence 19(3), 25–35 (2021).
3. Díaz Cintas J.: Teaching and learning to subtitle in an academic environment. In Díaz Cintas J. (ed.), The Didactics of Audiovisual Translation (pp. 89-103). Amsterdam and Philadelphia: John Benjamins (2008).
4. Campbell V.: Science, Entertainment and Television Documentary. Palgrave Macmillan, London (2016).
5. Guerberof, A.: Productivity and quality in the post-editing of outputs from translation memories and machine translation. Localisation Focus. The International Journal of Localisation, 7(1), 11–21 (2009).
6. Sanchez-Torron M., Koehn P.: Machine translation quality and post-editor productivity. In: Proceedings of AMTA (p. 16-26). Association for Machine Translation in the Americas, AMTA, Austin, Texas (2016)
7. Läubli S, Sennrich R, Volk M.: Has machine translation achieved human parity? a case for document-level evaluation. arXiv preprint arXiv:1808.07048. 2018 Aug 21.
8. Toral, A., Castilho, S., Hu, K., Way, A.: Attaining the unattainable? reassessing claims of human parity in neural machine translation. arXiv preprint arXiv:1808.10432. (2018).
9. Volk, M.: The automatic translation of film subtitles: a machine translation success story? In: Nivre, J; Dahllöf, M; Megyesi, B. Resourceful Language Technology: Festschrift in Honor of Anna Sågvall Hein. Uppsala, Sweden, 202-214. (2008).
10. Bywood, L., Georgakopoulou, P., Etchegoyhen, T. Embracing the Threat: Machine Translation as a Solution for Subtitling. Perspectives 25(3), 492–508 (2017). doi:10.1080/0907676X.2017.1291695.
11. Bywood, L., Georgakopoulou, P., Volk, M., Fishel, M.: What is the productivity gain in machine translation of subtitles? Paper presented at Languages & The Media, Berlin. (2012).
12. Matusov, E., Wilken, P., Georgakopoulou, Y. Customizing Neural Machine Translation for Subtitling. Proceedings of the Fourth Conference on Machine Translation (WMT), 1, pp. 82–93. Florence, Italy. Association for Computational Linguistics. 2019
13. Campos, R., Neira, G.R.: Las chicas del cable (2017). https://www.netflix.com/es/title/80100929
14. Google Translate, https://translate.google.com/about/languages/
15. DeepL Translator, https://www.deepl.com/translator
16. Trzaskawka, P.: Selected Clauses of a Copyright Contract in Polish and English in Translation by Google Translate: A Tentative Assessment of Quality. International Journal for the Semiotics of Law-Revue internationale de Sémiotique juridique 33(3), 689–705 (2020).
17. King, K.M.: Can Google Translate be taught to translate Literature? A case for humanists to collaborate in the future of machine translation. Translation Review 105(1), 76–92 (2019).
18. Suhono, S., Zuniati, M., Pratiwi, W., Hasyim, U. A. A. Clarifying Google Translate Problems Of Indonesia-English Translation Of Abstract Scientific Writing. EAI (24-25), 1–13 (2020).

19. De Nardi, I. “La casa de papel”: comparación entre el doblaje y la subtitulación al italiano del primer capítulo. (PhD dissertation). Università degli Studi di Padova (2018).
20. Rescigno A.A., Vanmassenhove, E., Monti, J., Way, A.: A Case Study of Natural Gender Phenomena in Translation A Comparison of Google Translate, Bing Microsoft Translator and DeepL for English to Italian, French and Spanish. In: Association for Machine Translation in the Americas (AMTA): Workshop on the Impact of Machine Translation (iMpacT 2020) (pp. 62–90). Workshop on the Impact of Machine Translation at Association for Machine Translation in the Americas (AMTA).
21. Cambedda G.: A Study on Automatic Machine Translation Tools: A Comparative Error Analysis Between DeepL and Yandex for Russian-Italian Medical Translation.
22. Hidalgo-Ternero C.M.: Google Translate vs. DeepL: Analysing neural machine translation performance under the challenge of phraseological variation.
23. Harmonized DQF-MQM Error Typology https://www.taus.net/qt21-project#harmonized-error-typology
24. Dynamic Quality Framework https://www.taus.net/data-for-ai/dqf
25. Multidimensional Quality Metrics (MQM) http://www.qt21.eu/mqm-definition/definition-2015-12-30.html
26. Papineni, K., Roukos, S., Ward, T., Zhu, W. J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pp. 311–318. Association for Computational Linguistics, Philadeplhia, USA (2018).
27. Tilde BLEU Score Evaluator https://www.letsmt.eu/Bleu.aspx

Fake News Detection for Portuguese with Deep Learning

Lígia Iunes Venturott[1,2][0000−0002−6339−4700] and
Ruslan Mitkov[2][0000−0003−2522−066X]

[1] New Bulgarian University, Sofia, Bulgaria
[2] University of Wolverhampton, Wolverhampton, UK

Abstract. The exponential growth of the internet and social media in the past decade gave way to the increase in dissemination of false or misleading information. Since the 2016 US presidential election, the term "fake news" became increasingly popular and this phenomenon has received more attention. In the past years several fact-checking agencies were created, but due to the great number of daily posts on social media, manual checking is insufficient. Currently, there is a pressing need for automatic fake news detection tools, either to assist manual fact-checkers or to operate as standalone tools. There are several projects underway on this topic, but most of them focus on English. This research-in-progress paper discusses the employment of deep learning methods, and the development of a tool, for detecting false news in Portuguese. As a first step we shall compare well-established architectures that were tested in other languages and analyse their performance on our Portuguese data. Based on the preliminary results of these classifiers, we shall choose a deep learning model or combine several deep learning models which hold promise to enhance the performance of our fake news detection system.

Keywords: Fake news detection · Deep learning · Portuguese.

1 Introduction

The term "fake news" is relatively new, having emerged in the 19th century. Before the 18th century the word "fake" was seldom used as an adjective and the expression "false news" was more common. According to Google Trends, the search for the term has significantly increased since October/November 2016, coinciding with the 2016 US presidential elections.

Even though the term gained popularity only in the past few years, the phenomenon of fake news itself is not new. Misinformation, disinformation, propaganda, conspiracy theories and others have always existed. However, the spread of this kind of content has exponentially grown with the new communication technology.

As a result of the rapid growth of the internet in the past decades, as well as the growth of social media, false or misleading information have been spreading at an alarming rate. Social media introduced a new kind of public space, one

Translation and Interpreting Technology Online, pages 149–153,
July 05-07, 2021.
https://doi.org/10.26615/978-954-452-071-7_016

where every individual has the opportunity to voice their opinion and potentially be heard by any other individual with internet access. The growth of social media and the lack of control of online content are major contributing factors to the phenomenon of fake news. Nowadays any person or organisation can create a social media profile and disguise as a professional news outlet. These actors provide misleading information which pretends to be reliable news.

Fake news also tend to be more appealing than true stories. Tweets containing false information reach 1500 persons six times faster than tweets containing real information [13]. According to Garner's prediction, "By 2022 most people in mature economies will consume more false information than true information"[1].

Several fact-checking agencies have surged as a response to the growth in fake news dissemination. However, manual checking of news is clearly insufficient when we consider the volume of posts in any of the major social media platforms [9]. Only Twitter, for example, had 340 million users and 500 million tweets per day on December 2020 [8].

While most work on fake news detection has been done for English, the topic has been scarcely researched for Portuguese and this ongoing study seeks to fill in this gap. Our objective is not only to develop additional fake detection tools for Portuguese, but to employ latest Deep Learning techniques in order to establish whether they will enhance the state-of-the-art of fake detection for Portuguese.

The rest of the paper is structured as follows. Section 2 surveys the work on fake detection for Portuguese so far. Section 3 of this work-in-progress poster paper outlines the envisaged methodology and briefly touches on the planned evaluation.

2 Related Work

Reis et al.[9] analysed several classification algorithms that use supervised learning. They use a dataset consisting of 2282 BuzzFeed articles about the 2016 US presidential election labelled by journalists. The authors explain three types of information can be extracted from news: content features, source features and environmental features, and extract several features of each of these categories.

Several classification algorithms were selected: k-Nearest Neighbours (KNN), Naive Bayes (NB), Random Forests (RF), Support Vector Machine with RBF kernel (SVM), and XGBoost (XGB). These models were used to evaluate the discrimination power of the extracted features. While it would have been promising to experiment with deep learning techniques, the authors opted for hand crafted features. The manual extraction of these features is probably time consuming, even with the help of automatic tools.

Silva et al.[10] evaluated different detection methods on the Fake.BR dataset[7], a dataset of fake news in Portuguese. The authors go about their study in two different ways: by extracting linguistic features from the data and by employing vector representations. They chose three types of vector representations: bag-of-words with TF-IDF, Word2Vec and Fast Text, and used pre-trained Word2Vec and Fast Text vectors. The idea was to compare what would perform better:

hand-crafted features or automatically extracted features. Several classification algorithms were selected in order to evaluate the representations: logistic regression (LR), support vector machines (SVM), decision trees (DT), random forest (RF), bootstrap aggregating (bagging) and adaptive boosting (Ada-Boost). The experiments with linguistic based features showed that these features are good enough to detect more than 90% of the fake news. This result is very interesting, because hand-crafted features do not require complex classifiers, meaning that a detection algorithm could run on even the simplest of devices. In the experiments with vector representations, the authors conclude that the bag-of-words model delivered better results than Fast Text or Word2Vec. The best F-measure obtained with Word2Vec and FastText were 0.893 and 0.897, respectively, while the best F-measure with BoW was 0.971. The authors hypothesise that, since the pre-trained vectors were trained on well-written texts such as Wikipedia and Google News, these vectors were not the best fit to represent fake news, which contains noise, such as incorrect spelling and slang.

3 Methodology

3.1 Datasets

As supervised learning will be employed in this project, a dataset of fake and real news will be needed to train the algorithm. So far we have identified three datasets available online.

The first, Fake.BR [7], is a balanced dataset containing 7,200 news, of which 3,600 classified as fake and 3,600 classified as true. The second, FACTCK.BR, is unbalanced and contains 1,309 news in total, of which 943 true, 246 half-true and 120 false. The above dataset does not offer explicit texts, only the url from where the news was taken, which means that in order to use it we would need to add another step of web scraping. The third, Boatos.org is available on Kaggle[3] and contains 1,900 WhatsApp messages proved to be fake news by fact-checking agencies. Unfortunately, it does not contain messages without false information, so we cannot use this dataset alone to train the algorithms.

If the aforementioned datasets turn out not to be suitable for our study, we shall compile a small annotated corpus of fake news in Portuguese.

3.2 Preprocessing

A preprocessing pipeline will be also implemented in order to minimise the noise in the data. Normally the pipeline for text preprocessing consists of: (1) Tokenisation, (2) Normalisation (lower case, remove accents and special characters, convert to ascii), (3) Lemmatisation or Stemming, (4) Stop word removal and (5) Numeralisation. Other steps may be added if necessary.

[3] https://www.kaggle.com/rogeriochaves/boatos-de-whatsapp-boatosorg

3.3 Classifiers

In this project we shall use deep learning techniques to detect fake news. For this purpose, we will evaluate and compare different deep architectures and identify additional ones, if needed, that hold promise for high performance on this task.

LSTM There are several deep learning architectures available, but because we are dealing with text, which is a type of sequential data, we will start our experiments with the LSTM (Long Short-Term Memory)[6]. LSTMs were introduced as an improvement to regular Recurrent Neural Networks (RNN). An RNN analyses each word in a sentence at a time. At each time step the layer analyses one word and generates an output. This output is then used as an additional input for the next time step with the next word. By using this mechanism, when analysing a word in a sentence the layer has access to information about the previous word and, recursively, about every word that occurred before. Simple RNNs are not suited for real-world applications due to the vanishing/exploding gradient problem [6]. In the LSTM, the simple summation cell is substituted for a memory block with 3 gates. These gates allow the information to be stored for more time steps, what remedies the problem of vanishing gradients[5].

Attention Based on our literature review, we believe that the performance might be boosted by using attention mechanisms[2]. The attention mechanism was originally crated in the context of machine translation. It allows the decoder to focus on important areas of the source sentence during the generation of the target sentence. The attention mechanism can also be used in classification tasks, and it might help improve the performance of the LSTM.

BERT The Transformer architecture[12] was also created for machine translation. It uses layers of multi-head self-attention mechanism and simple feed-forward networks to build the encoder and decoder. BERT is a language representation model based on transformers[4]. It can be pre-trained on unlabelled data, and then fine-tuned for specific tasks. There is already a pre-trained BERT model for Brazilian Portuguese[11]. We plan to test this model with fine-tuning and compare it with the other approaches.

Based on the preliminary results of these classifiers, we shall choose a deep learning model or combine several deep learning models which hold promise to enhance the performance of our fake news detection system.

3.4 Evaluation

We will evaluate our fake news detection system using standard evaluation metrics such as Accuracy and F1-score. We will compare our system with other existing fake detection systems for Portuguese, such as the ones mentioned in Section 2. In order to report statistical significance, we plan to use the Wilcoxon signed-rank test[3]. In addition to this intrinsic evaluation, we also envisage extrinsic evaluation where users evaluate the efficiency of the tool to be developed.

References

1. Gartner reveals top predictions for it organizations and users in 2018 and beyond (Oct 2017), https://www.gartner.com/en/newsroom/press-releases/2017-10-03-gartner-reveals-top-predictions-for
-it-organizations-and-users-in-2018-and-beyond
2. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
3. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. **7**, 1–30 (Dec 2006)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2019)
5. Graves, A.: Long short-term memory. In: Supervised sequence labelling with recurrent neural networks, pp. 37–45. Springer (2012)
6. Hochreiter, S., Schmidhuber, J.: Long Short-Term Memory. Neural Computation **9**(8), 1735–1780 (11 1997). https://doi.org/10.1162/neco.1997.9.8.1735, https://doi.org/10.1162/neco.1997.9.8.1735
7. Monteiro, R.A., Santos, R.L.S., Pardo, T.A.S., de Almeida, T.A., Ruiz, E.E.S., Vale, O.A.: Contributions to the study of fake news in portuguese: New corpus and automatic detection results. In: Computational Processing of the Portuguese Language. pp. 324–334. Springer International Publishing (2018)
8. Omnicore: Omnicore. https://www.omnicoreagency.com/twitter-statistics/ (2020), accessed 05/05/2021
9. Reis, J.C.S., Correia, A., Murai, F., Veloso, A., Benevenuto, F.: Supervised learning for fake news detection. IEEE Intelligent Systems **34**(2), 76–81 (2019). https://doi.org/10.1109/MIS.2019.2899143
10. Silva, R.M., Santos, R.L., Almeida, T.A., Pardo, T.A.: Towards automatically filtering fake news in portuguese. Expert Systems with Applications **146**, 113199 (2020). https://doi.org/https://doi.org/10.1016/j.eswa.2020.113199, https://www.sciencedirect.com/science/article/pii/S0957417420300257
11. Souza, F., Nogueira, R., Lotufo, R.: BERTimbau: pretrained BERT models for Brazilian Portuguese. In: 9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear) (2020)
12. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. arXiv preprint arXiv:1706.03762 (2017)
13. Vosoughi, S., Roy, D., Aral, S.: The spread of true and false news online. Science **359**(6380), 1146–1151 (2018). https://doi.org/10.1126/science.aap9559, https://science.sciencemag.org/content/359/6380/1146

What Makes a Concept Complex? Measuring Conceptual Complexity as a Precursor for Text Simplification

Anne Eschenbrücher$^{[0000-0001-7024-6100]}$

Research Group in Computational Linguistics
University of Wolverhampton, UK
a.eschenbruecher@wlv.ac.uk

Abstract. Advancements within the field of Text Simplification (TS) have primarily been within syntactic or lexical simplification. However, conceptual simplification has previously been identified as another field of TS that has the potential to significantly improve reading comprehension. The first step to measuring conceptual simplification is the classification of concepts as either complex or simple. This research-in-progress paper proposes a new definition of conceptual complexity alongside a simple machine-learning approach that performs a binary classification task to distinguish between simple and complex concepts. It is proposed that this be a first step when developing new text simplification models that operate on a conceptual level.

Keywords: Text Complexity · Text Simplification · NLP in Educational Applications.

1 Introduction

Text is a constant companion in our daily lives. Without the ability to write down and record knowledge, the world and its people could arguably never have evolved to its current state. Despite the big impact of text on our daily lives, around one in six adults in the United Kingdom has poor literacy skills [1]. In order to combat this problem, texts have long been simplified to improve understanding for all readers [2]. With the advancements in the field of Natural Language Processing (NLP), or more specifically, Text Simplification (TS), true progress has been made especially within the subfields of lexical and syntactic simplifications [2–5]. A text can, however, be difficult to understand not only lexically or syntactically, but also conceptually. As of yet, conceptual text simplification has been paid significantly less attention than its two related subfields, although it has the potential to significantly improve reading comprehension. Conceptual complexity is in itself a concept hard to grasp. Definitions vary from "the importance of distinguishing between ideas in a text and fact burden of a text" [6] to the amount of background knowledge a reader needs to in order to understand the concepts mentioned in the text and how they relate to each other

Translation and Interpreting Technology Online, pages 154–160,
July 05-07, 2021.
https://doi.org/10.26615/978-954-452-071-7_017

[7, 8]. Yet another definition refers to a "need for algorithms that can understand what a user needs to know before they can understand a second concept" [9]. According to all these definitions, the amount of background knowledge is what essentially makes a text conceptually complex. However, they all fail to take into account reading processes that occur in the brain.

While reading a text, a reader constantly has to chunk concepts together into larger order units in order to free up working memory space to avoid displacement [10]. In order for this to happen, all concepts and their relations that are involved need to be present simultaneously in short-term memory (STM) [11]. This is where differences between skilled and poor readers become obvious as skilled readers are able to suppress irrelevant information more easily and only store what is truly important, while also being generally better at chunking the relevant information which enables them to comprehend a text more fully than poor readers would [12].

When reading a text, ideally a reader would chunk as much as possible to avoid displacement as this would mean that important concepts are lost from STM [10]. It can therefore be argued that a text is conceptually complex if the working memory's capacity is frequently at its limit. We assume that closely related concepts are probably easier to chunk together, which is why relatedness should also play an important role. Hence, our working definition of conceptual complexity is as follows: Conceptual complexity is the complexity of a concept by itself, referring to level of abstractness and frequency of use in everyday language, as well as the complexity of a concept's context, meaning how complex its relations are to other concepts in its direct vicinity.

2 Related Work

One particular challenge within the field of conceptual TS is the estimation of conceptual text complexity. While many studies focus on automatic assessment of lexical and syntactic complexity [13–15], automatic approaches for measuring conceptual complexity are rare. One of the first automated tools that took conceptual complexity into account was the Coh-Metrix tool which uses latent semantic analysis semantic features and coreference indices alongside measures of lexical and syntactic complexity [16]. More recently, the estimation of conceptual complexity has been attempted using knowledge graphs [7], as well as a mix of surface-based and shallow semantic features [8].

Recently, the focus of text simplification systems has shifted from simplifying a text to its easiest possible version to simplifications that are tailored to a specific reader's needs [5]. So far, text simplification systems that take the concept level into account use mainly extractive approaches that eliminate irrelevant information [17, 18]. Some concepts, however, might be essential to a text and cannot be eliminated which is why an abstractive approach needs to be used that generates new and simpler output. To identify which concepts are complex and need to be replaced by simpler substitutes, it is not sufficient to estimate the complexity of an entire text. Rather, each individual concept needs to be assigned a complexity score that takes into account the concept's individual complexity, as well as its context complexity.

3 Methodology

The proposed methodology contains several features that take into account a concept's relations with other concepts that occur in the same sentence, as well as how complex the concept is by itself. Those features, as well as the model and annotation process are briefly described below.

3.1 Pairwise Features

Features that take into account a concept's relations to other concepts are computed using WordNet and ConceptNet databases [19, 20]. In order to determine how complex a concept is in relation to other concepts it is being mentioned with, we compute two types of pairwise features.

Firstly, we compute the WordNet path similarity between a concept and all other concepts in a sentence. The mean for each concept is taken as that concept's score for the respectable context. Secondly, ConceptNet relatedness is computed using the ConceptNet API. Here, we also take the mean for each concept's score together with all other concepts in that sentence as feature.

In the sentence "Such were Elizabeth Elliot's sentiments and sensations; such the cares to alloy, the agitations to vary", we first extract and lemmatise the concepts "sentiment", "sensations", "cares" and "agitations". We then take the first concept "sentiment" and compute e.g. its path similarity with all other concepts in this sentence, finally taking the mean of all those as a feature. We repeat this process for each concept in a given sentence before moving on to the next. The same is repeated to compute ConceptNet relatedness.

3.2 Single Features

A total of eight features is computed for each single concept. Word length is taken into account as more complex meanings tend to be encoded in longer structures [21]. Furthermore, type-token ratio is computed for each concept per sentence, as well as for the entire text. As complex concepts tend to be more specific, the number of WordNet hypernyms and hyponyms is also taken into account. The WordNet number of senses for each concept is computed to account for polysemy and abstractness. Some concepts such as e.g. "love", may be complex, yet they are mentioned so frequently that their complexity is not perceived anymore. To account for this, the frequency per one million words for each mention of a concept is taken from the British National Corpus [22]. Finally, for each concept, the number of ConceptNet relatives is counted if its weight, that is its confidence score, is greater than or equal to 0.5.

3.3 Human Annotation

For this task, a total of 247 sentences are taken from “The Great Gatsby” by F. Scott Fitzgerald, “Moby Dick” by Herman Melville, as well as “Persuasion” by Jane Austen. Those 247 sentences contain a total of 7978 tokens, 1687 of which are concepts. All of these are manually annotated by one human annotator according to a set of annotation guidelines. All concepts are judged in terms of their conceptual complexity where “0” denotes a conceptually simple concept, while “1” means that the concept is conceptually complex.

3.4 Model

Using h2o’s AutoML [23], classification is attempted on all 1687 concepts. We run three subsets of the data, the first containing all features, while the second and third subsets each exclude either those features derived from WordNet or those derived from ConceptNet. The leading model in all three cases is a stacked model comprised of a deep learning model stacked on top of a gradient boosting machine, a random forest and decision trees.

4 Results

The stacked model was built using all 1687 concepts of which 1373 were used for trainig while 157 were used respectively for valiation and testing. At this point in time, the model built on the subset of the data without the WordNet features (hypernyms, hyponyms, number of senses, path similarity) performs best, achieving a mean squared error of 0.062 and a f1-score of 0.7083. When training the model with a subset not containing the ConceptNet features (number of relatives, relatedness), the results achieved were the worst among the three subsets.

Table 1. Results for stacked machine learning model.

Metric	All Features	w/o WordNet	w/o ConceptNet
MSE	0.0653	**0.062**	0.0648
RMSE	0.2555	**0.249**	0.2546
Accuracy	**0.931**	0.9141	0.8865
Precision	**0.75**	0.6071	0.5224
Recall	0.5	0.85	**0.875**
F1 Score	0.6	**0.7083**	0.6542

Out of the 157 concepts in the test set, 15 were judged to be conceptually complex by the human annotator. While both the model built on all features, as well as the model built on the subset without ConceptNet features found around the same amount of complex concepts, 16 and 14, they included many false positives. In contrast, the model without WordNet features classified a total of 22 concepts as complex, yet included more actually complex concepts.

5 Discussion

As this is a work-in-progress paper and, as of yet, there are no finalised results, the discussion will be primarily focused on the weaknesses of this methodology rather than relating the findings to previous works in the field.

So far, due to computational restraints, the dataset used was fairly small. In the future, we plan on increasing the size of the dataset. Furthermore, the increased dataset would need to be annotated by more than one annotator in order to decrease the possibility of the one annotator having a bias that is being transferred onto the data.

In terms of feature engineering, other knowledge graphs and ontologies should be explored to possibly replace those features based on the WordNet ontology. A big shortcoming of WordNet is that its depth varies in different parts. This leads to some concepts having less hyper -or hyponyms than others and, hence, the possibility of these being falsely classified as simple when they are in fact complex.

Furthermore, the predictions on the test set show that each of the three models has a slight tendency to falsely classify longer words as more complex. Future experiments should experiment with excluding the word-length feature to see whether this can improve results.

6 Conclusion

In this paper, we proposed a new definition for conceptual complexity. We extracted concepts from a text and use a variety of pairwise, as well as single features, to classify them as either simple or complex. It is proposed that our research forms part of text simplification systems in order to improve output.

While the estimation of conceptual complexity can be a good precursor for text simplification systems, it could also be useful for various other natural language processing tasks. Taking Machine Translation (MT) as an example, conceptual estimation could be performed as a step in the MT pipeline for under-resourced language.

For future work, it would be interesting to experiment with more features involving different types of knowledge graphs as well as including more pair-based features.

References

1. Adult Literacy, https://www.literacytrust.org.uk/parents-and-families/adult-literacy/. Last accessed 14 May 2021
2. Siddharthan, A.: A survey of research on text simplification. ITL - International Journal of Applied Linguistics **165**(2), 159–298 (2014)
3. Paetzold, G. H., Specia, L.: A survey on lexical simplification. Journal of Artificial Intelligence Research **60**, 549–593 (2017)
4. Saggion, H.: Automatic Text Simplification. Sunthesis Lectures on Human Language Technologies **10**(1), 1–137 (2017)

5. Sikka, P., Singh, M., Pink, A., Mago, V.: A survey on text simplification. Journal of the Association for Computing Machinery **37**(4), (2020)
6. Dolch, E. W.: Fact burden and reading difficulty. The Elementary English Review **16**(4), 135–138 (1939)
7. Štajner, S., Hulpuş, I.: Automatic assessment of conceptual text complexity using knowledge graphs. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 318–330. ACL, Santa Fe, New Mexico, USA (2018)
8. Štajner, S., Hulpuş, I.: When shallow is good enough: Automatic assessment of conceptual text complexity using shallow semantic features. In: Proceedings of the 12th Language Resources and Evaluation Conference, pp. 1414–1422. ACL, Marseille, France (2020)
9. Collins-Thompson, K.: Computational assessment of text readability: A survey of current and future research. ITL - International Journal of Applied Linguistics **165**(2), 97–135 (2014)
10. Daneman, M., Carpenter, P. A.: Individual differences in working memory and reading. Journal of verbal learning and verbal behaviour **19**(4), 450–466 (1980)
11. Shiffrin, R. M., Schneider, W.: Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. Psychological review **84**(2), 127–190 (1977)
12. Palladino, P., Cornoldi, C., De Beni, R., Pazzaglia, F.: Working memory and updating processes in reading comprehension. Memory and cognition **29**(2), 344–354 (2001)
13. Vajjala, S., Meurers, D.: Assessing the relative reading level of sentence pairs for text simplification. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pp. 288–297. ACL, Gothenburg, Sweden (2014)
14. Vajjala, S., Meurers, D.: Readability-based sentence ranking for evaluating text simplification.In: arXiv preprint, arXiv:1603.06009 (2015)
15. Ambati, B. R., Reddy, S., Steedman, M.: Assessing relative sentence complexity using an incremental CCG parser. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1051–1057. ACL, San Diego, California (2016)
16. Graesser, A. C., McNamara, D. S., Louwerse, M. M., Cai, Z.: Coh-Metrix: Analysis of text on cohesion and language. Behaviour research methods, instruments, & computers **36**(2), 193–202 (2004)
17. Narayan, S., Gardent, C.: Hybrid simplification using deep semantics and machine translation. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL), pp. 435–445. ACL, Baltimore, Maryland (2014)
18. Štajner, S., Glavaš, G.: Leveraging event-based semantics for automated text simplification. Expert systems with applications **82**, 383–395 (2017)
19. Oram, P.: WordNet: An Electronic Lexical Database. MIT Press, Cambridge, Massachusetts (1999)
20. Speer, R., Chin, J., Havasi, C.: ConceptNet5.5: An Open Multilingual Graph of General Knowledge. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 4444–4451. AAAI, San Francisco, California (2017)
21. Lewis, M. L., Frank, M. C.: The length of words reflects their conceptual complexity. Cognition **153**, 182–195 (2016)
22. The British National Corpus, version 3. Distributed by Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium. http://www.natcorp.ox.ac.uk/ (2007)

23. LeDell, E., Poirier, S.: H2O AutoML: Scalable Automatic Machine Learning. In: 7th ICML Workshop on Automated Machine Learning. AutoML (2020)

Integration of Machine Translation and Translation Memory: Post-editing efforts

Rocío Caro Quintana

University of Wolverhampton, UK
R.Caro@wlv.ac.uk

Abstract. The development of Translation Technologies, like Translation Memory and Machine Translation, has completely changed the translation industry and translator's workflow in the last decades. Nevertheless, TM and MT have been developed separately until very recently. This ongoing project will study the external integration of TM and MT, examining if the productivity and post-editing efforts of translators are higher or lower than using only TM. To this end, we will conduct an experiment where translation students and professional translators will be asked to translate three short texts; then we will check the post-editing efforts (temporal, technical and cognitive efforts) and the quality of the translated texts.

Keywords: Translation Memory, Machine Translation, Integration, Post-editing efforts, Eye-tracking

1 Introduction

The way translators work has changed considerably in the last decades: with a world that is more globalised each day, the creation of texts has increased, and this has also affected and transformed the translation industry. The development of Translation Memory (TM) tools and Machine Translation (MT) systems is the result of these changes.

Translation Memory and Machine Translation as applications have been used for many years by academics and professional translators, but they have been developed and studied in isolation [1]. Nonetheless, recent research suggests there is an increase in the interest of integrating these two applications.

The integration of TM and MT creates new questions for all the parties involved in the translation process (translators, language service providers, clients) and academia. Will the integration save time? Therefore, will it also save money? Will the quality of the texts be as good as a translation from scratch?

Even if the answer to these questions turns out to be positive (e.g., the time will be reduced and the quality maintained), another question remains: Will the cognitive efforts of translators decrease? With the integration of TM and MT, more options (or segments) are presented to the translator and the translator's cognitive efforts may vary as the translation process would have one more step (explained in five simple steps):

Translation and Interpreting Technology Online, pages 161–166,
July 05-07, 2021.
https://doi.org/10.26615/978-954-452-071-7_018

1) Read source segment
2) Read TM output
3) Read MT output
4) Compare and choose which option is better
5) Translate or post-edit

The aim of this work-in-progress paper is to compare the different types of post-editing efforts (temporal, technical and cognitive) of translating with only TM or MT and translating with the combination of TM and MT. That is, calculating if the time it takes to finish a translation task is higher or lower using both technologies, checking if the technical efforts (studying the keystrokes and edits) are the same, lower, or higher; and examining if the cognitive efforts are also higher or lower using TM and MT. Our initial hypothesis is that the three efforts will decrease with the combination of TM and MT, however, while there are studies about the differences between translating with TM and MT [2, 3], empirical research about translating with a Computer-Assisted Translation (CAT) tool that integrates TM and MT is still limited.

Therefore, this ongoing project will try to answer the following research questions**:**
RQ1: Will the translator save time using a system that integrates TM and MT?
RQ2: Will the translator invest less cognitive efforts when using TM and MT?
RQ3: Will the quality of the translation not be deteriorated when using TM systems with integrated MT?

2 Related work

The integration of TM and MT was first envisaged by Peter Arthern [4] and nowadays most CAT tools have an option to integrate both technologies.

According to Zaretskaya et al. [5] the integration of TM and MT can be separated in two main categories: internal and external. The internal integration aims at improving the quality of the TM systems by using MT techniques, or vice versa, thus providing only one option to the translator. The external integration provides another suggestion, the MT output, in addition to the matches from the TM.

One of the first studies on the internal processing integration of TM and MT is from authors Lange and Bennet [6], in which they integrated MT into the translation process by treating the MT segments as fuzzy matches. Several papers working on internal integration were published the following years [7, 8, 9].

Concerning the external integration, Kanavos and Kartsaklis [10] is one of the first studies about this topic, however, the integration is not straightforward as the MT is not suggested with every segment, but only with the segments with a low fuzzy match. Other relevant studies are The MateCat project [11], and Eriguchi et al. [12] where TM segments were combined with Neural Machine Translation.

Herbig et al. [13] briefly discuss that “machines can generate a variety of probable translations from MT and TM instead of a single one” and that this could bias or confuse the human translator. The translators could, for example, see an output from the TM or the MT and consider it accurate since the other aspects of the sentence (grammar, punctuation, coherence, etc.) may be correct, whereas the translation is not. They claimed, therefore, that future research should focus on estimating the cognitive efforts of translators in these types of scenarios.

3 Methodology

3.1 Experiments

The experiments we will carry out are, according to Zaretskaya et al. [5], within the external integration category. We will check if the translators' post-editing efforts increase or decrease in a setting where both segments from TM and MT are suggested.

As the main objective of the project is to examine the differences between translating with the help of only TM or MT and translating with the combination of both technologies, professional translators and Translation students will be asked to translate three short texts (around 300 words) with a CAT tool from English into Spanish.

The first text will be translated using TM, the second text will be translated with MT, and the third text will be translated using TM and MT. As the texts are different the results may not be conclusive, therefore the same texts should be translated again changing the order.

Table 1. Distribution of the texts

1st text	2nd text	3rd text
TM	MT	TM + MT

The study will be divided in two stages: the pilot project and the main project. For the pilot project and, due to the COVID-19 pandemic, participants can carry out the task remotely (in their homes, university, library) as long as they have an internet connection; in the main project, the translation task will be carried out in the presence of the researcher using an eye-tracking device that records the eye movements and the size of the pupil of the translators.

3.2 Post-editing efforts

The post-editing efforts, first described by Krings [14], will be studied as follows:

Temporal effort: The total time it takes for each translator to complete the task will be recorded, as well as the seconds per segments and the seconds per word.

Technical effort: An analysis of the keystrokes will be made with a key-logging tool, as well as the edit distance, which is the number of changes that were made from the source text to the translated text. and the post-edit distance. Post-edit distance is calculated by dividing the number of words or characters in the original text between the number of words or characters in the post-edited text.

Cognitive effort: To study the cognitive efforts, we will use an eye-tracking device. The measures that will be studied with the eye-tracker are the following: pauses or fixations (fixation duration, fixation count, first and second fixation duration), eye-movement, pupil dilation and regression behaviour. The eye-tracking software generates a video recording of the screen and the face/eyes of the participant, as well as an Excel file with the data collected. It also includes heat maps and gaze plots.

3.3 Qualitative data

In order to collect qualitative data, this project will also include two questionnaires: one before the translation task and one after the translation task. The first questionnaire will contain questions about the background of the participants (education, job experience on translation), their knowledge of CAT tools and their general opinion of working with translation technologies. The questionnaire after the translation task will include questions regarding their opinion of the experiment: if they think using TM and MT is beneficial, if they prefer working only with TM, etc. Previous studies have also conducted similar surveys [2, 15, 3].

The quantitative results (gathered from the translation task) may differ from the qualitative results. For instance, the eye-tracking results could show that the translator invests less cognitive efforts during the translation of the texts using TM and MT combined, but the experience of the participants could be the opposite, i.e., they find it more difficult to translate with TM and MT, rather than with only TM. Therefore, we also plan to analyze the times each translator chooses each option (translate from scratch, edit TM or post-edit MT), and if the option they choose is the most beneficial in terms of cognitive and temporal efforts.

3.4 Quality Evaluation Methods

The results will be evaluated manually and automatically. For the manual evaluation, the TAUS DQF/MQM Error Typology [16] will be used.

Regarding the automatic evaluation, all the translations made by the participants will be assessed with the automatic metrics BLEU [17], TER [18] and METEOR [19] comparing the results of the participants with a ‘gold standard’ to check if they correspond to the manual evaluation.

4 Conclusion

To sum up, we will carry out an experiment to check whether the external integration of TM and MT, where translators are able to see segments proposed by both technologies, is beneficial for translators. To do that, we will conduct a translation task, and we will assess if the temporal, technical and cognitive efforts increase or decrease. We will also check the quality of the translations, and we will collect qualitative data about the opinion of the participants with two questionnaires.

References

1. Koehn, Philipp, and Jean Senellart. "Convergence of translation memory and statistical machine translation." Proceedings of AMTA Workshop on MT Research and the Translation Industry. (2010)
2. Guerberof Arenas, Ana. "Productivity and quality in the post-editing of outputs from translation memories and machine translation." Localisation Focus The International Journal of Localisation 7.1 (2008): 11-21.
3. Sánchez-Gijón, Pilar, Joss Moorkens, and Andy Way. "Post-editing neural machine translation versus translation memory segments." Machine Translation 33.1 (2019): 31-59.
4. Arthern, Peter J. "Machine translation and computerized terminology systems: a translator's viewpoint." Translating and the Computer: Proceedings of a Seminar, London. Vol. 14. (1978)
5. Zaretskaya, Anna, Gloria Corpas Pastor, and Miriam Seghiri. "Integration of Machine Translation in CAT tools: State of the art, evaluation and user attitudes." Skase Journal of Translation and Interpretation 8.1 (2015): 76-89.
6. Lange, Carmen Andres, and Winfield Scott Bennett. "14. Combining Machine Translation with Translation Memory at Baan." Translating into success. John Benjamins, (2000) 203-218.
7. Marcu, Daniel. "Towards a unified approach to memory-and statistical-based machine translation." Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, (2001)
8. Langlais, Philippe, and Michel Simard. "Merging example-based and statistical machine translation: an experiment." Conference of the Association for Machine Translation in the Americas. Springer, Berlin, Heidelberg, (2002)
9. Groves, Declan, and Andy Way. "Hybrid data-driven models of machine translation." Machine Translation 19.3-4 (2005): 301-323.
10. Kanavos, Panagiotis, and Dimitri Kartsaklis. "Integrating machine translation with translation memory: A practical approach." (2010).
11. Federico, Marcello, Alessandro Cattelan, and Marco Trombetti. "Measuring user productivity in machine translation enhanced computer assisted translation." Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas (AMTA). (2012)
12. Eriguchi, Akiko, Spencer Rarrick, and Hitokazu Matsushita. "Combining translation memory with neural machine translation." Proceedings of the 6th Workshop on Asian Translation. (2019)
13. Herbig, Nico, et al. "Integrating Artificial and Human Intelligence for Efficient Translation." arXiv preprint arXiv:1903.02978 (2019).
14. Krings, Hans P. Repairing texts: Empirical investigations of machine translation post-editing processes. Vol. 5. Kent State University Press, (2001)
15. Bundgaard, Kristine. "Translator Attitudes towards Translator-Computer Interaction–Findings from a Workplace Study." Hermes 56 (2017): 125-144.
16. Harmonized DQF-MQM Error Typology https://www.taus.net/qt21-project#harmonized-error-typology
17. Papineni, Kishore, et al. "Bleu: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting of the Association for Computational Linguistics. (2002)

18. Snover, Matthew, et al. "A study of translation edit rate with targeted human annotation." Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers. (2006)

19. Banerjee, Satanjeev, and Alon Lavie. "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments." Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. (2005)

Interactive Models for Post-Editing

Marie Escribe, Ruslan Mitkov

Research Group in Computational Linguistics
University of Wolverhampton
{m.c.escribe,r.mitkov}@wlv.ac.uk

Abstract. Despite the increasingly good quality of Machine Translation (MT) systems, MT outputs require corrections. Automatic Post-Editing (APE) models have been introduced to perform these corrections without human intervention. However, no system has been able to fully automate the Post-Editing (PE) process. Moreover, while numerous translation tools, such as Translation Memories (TMs), largely benefit from translators' input, Human-Computer Interaction (HCI) remains limited when it comes to PE. This research-in-progress paper discusses APE models and suggests that they could be improved in more interactive scenarios, as previously done in MT with the creation of Interactive MT (IMT) systems. Based on the hypothesis that PE would benefit from HCI, two methodologies are proposed. Both suggest that traditional batch learning settings are not optimal for PE. Instead, online techniques are recommended to train and update PE models on the fly, via either real or simulated interactions with the translator.

Keywords: Automatic Post-Editing, Machine Translation, Human-Computer Interaction.

1 Introduction

The emergence of automatic translation dates back to 1949, when Warren Weaver, a researcher at the Rockefeller Foundation, presented a set of proposals for MT solutions which were based on information theory and successes in code breaking during the Second World War. However, MT research faced several challenges over the years, which led to what was almost a standstill in the field for many years to come. In 1966, the Automatic Language Processing Advisory Committee (ALPAC) report concluded that MT outputs were too disappointing to continue investigating such systems, especially since there were enough human translators to complete translation projects [1]. This conclusion no longer stands. Translation plays a crucial role in communication in today's globalised world, and the number of texts requiring translation continues to rise. With such high volumes and tight deadlines, translators now heavily rely on technological assistance. Most projects are indeed undertaken in a Computer-Assisted Translation (CAT) environment, in which professionals use TMs and/or post-edit MT outputs. As the quality of MT outputs continues to increase, PE has become a common step in translation workflows nowadays. Nevertheless, since transla-

Translation and Interpreting Technology Online, pages 167–173,
July 05-07, 2021.
https://doi.org/10.26615/978-954-452-071-7_019

tions generated by MT systems need editing, methods have been proposed to automate this process using APE models.

2 Automatic Post-Editing

APE was first suggested by Knight and Chander [2] as a module complementing MT systems. Allen and Hogan [3] also observed that recurring errors are often found in MT outputs and introduced an APE module based on a controlled language to address this issue. This module was trained using triplets (source, MT output, post-edited version) to extract PE rules and apply these to unseen texts. This is still a typical approach for most APE models today, but more sophisticated methods have been suggested to improve the training phase. In fact, the evolution of APE has tended to follow that of MT, as they both benefitted from very similar technological improvements. Just like MT systems, APE was first based on rules and then adopted statistical methods before utilising machine learning and neural networks. Since 2006, when Llitjós and Carbonell [4] raised the issue of a lack of fully automatic solutions to PE, APE has regained popularity with the first APE shared task at the WMT conference series [5]. These have been running regularly ever since, providing datasets and a forum for discussing the latest advances in this field. The first round of the APE shared task was not very successful, as no system could beat the baseline [6]. However, significant improvements were later achieved, when APE systems began to use neural approaches [7]. Despite this, it is common for APE systems to produce overcorrections and to fail to detect certain errors [5].

Moreover, APE became even more challenging, as Neural MT (NMT) provides translations of a higher quality compared to previous MT systems, thus making automatic correction a more complex task [5,8]. In fact, the usefulness of APE models resides in the improvement margin observed in MT outputs [5].

Nevertheless, APE models still have clear advantages. While it could be argued that retraining an MT system would yield similar results in terms of corrections applied to the target text, this might not always be feasible, in particular because it would require having access to the MT system parameters [6].

3 Human-Computer Interaction in MT & PE

Although APE might contribute to improving MT outputs, the target texts generated in this way require proofreading by human experts. It could even be argued that working with an APE output necessitates more effort than PE alone, as extra attention should be paid to overcorrections and unspotted errors. This creates a significant discrepancy between APE systems and the real needs of MT users. Moreover, professional post-editors benefit from translation technologies (e.g. CAT), but these are not optimal for PE. TMs are valuable in assisting professionals and benefit from interaction with the translator: they are populated with approved translations which can easily be reused, via either automatic propagation of fuzzy matches or a manual concordance search. TMs are therefore constantly adapted to the style of the post-editor and to the register and terminology of the document being translated. When it comes to PE, however, HCI is very limited, as corrections made by post-editors are not exploited.

Several interactive models have been introduced for MT, which use an interface to collect and reuse corrections made by post-editors. Well-known Interactive MT (IMT) models include Transtype [9], Transtype 2 [10], MIPRCV (Multimodal Interaction in Pattern Recognition and Computer Vision) [11] and CASMACAT (Cognitive Analysis and Statistical Methods for Advanced Computer-Aided Translation) [12]. These models offer a form of autocompletion in which suffixes are suggested based on translation prefixes validated by the post-editor, who might select to accept or modify the predictions. More recently, several studies have suggested implementing IMT models based on NMT systems [13, 14, 15, 16]. This typically involves adapting the search algorithm to make predictions which are constrained to a given prefix. IMT is therefore a human-centred approach [17], as IMT models evolve based on interactions with the translator. Despite this, certain studies comparing the PE effort in a traditional setting versus IMT yielded mixed results. For instance, in the paper by Underwood et al. [18], certain participants did not find the interactive setting helpful, while others reported positive experiences. Alves et al. [19] found that IMT did not yield improvements in efficiency (time and number of keystrokes) but could contribute to reduce the cognitive effort (shorter fixation durations were observed using eye-tracking equipment in the case of IMT). Several explanations to these results can be put forward, such as the lack of familiarity of the participants with such interfaces and the time and effort required to engage with predictions which are constantly being updated.

While HCI has been explored using glass-box approaches in IMT settings, several studies have also introduced ‘interactive’ APE models. APE is a black-box approach, as it does not require access to the MT system parameters but only to an MT output. Augmenting such systems by adding an interactive component enables APE models to learn from corrections during the PE process and to make informed correction predictions. Consequently, such models do not fall under the category of APE, as this process would not be fully automatic. Instead, they would be Interactive PE (IPE) systems. The study by Knight and Chander [2] was the first to briefly formulate this concept, which they called an ‘adaptive post-editor’. This suggestion is in line with the findings of the user survey conducted by Lagoudaki [20], which revealed that translators found the concept of an adaptive PE system highly relevant.

With the objective to implement this theory, Simard and Foster [21] introduced PEPr (Post-Edit Propagation), a model inspired by TMs which takes a phrase-based MT output and uses an APE module based on an online method to learn from human corrections on the fly. The assumption is that propagating corrections automatically can be beneficial when the number of repetitions in a text is high. Building on this model, Lagarda et al. [22] proposed an online APE system specifically designed to improve domain adaptation.

More recently, Chatterjee et al. [23] emphasised the need for APE systems capable of handling continuous streams of data to adapt to evolving settings and to the variety of domains present in real-world translation workflows, and presented a statistical online APE model designed to address this challenge. Building on [23], Negri et al. [24]

developed an online APE system which can learn from simulated interaction with the human post-editor. However, these interactive APE models all simulate interaction with the translator, which has several implications. The most salient limitation is that corrections are constrained to the pre-existing references, which allows for one possible correction only. This is a rather unrealistic scenario, as translation is an open-ended problem. Consequently, retrieving human input in an interactive environment would provide a unique opportunity to improve the PE process.

4 Suggested Methodologies

Two methodologies are proposed to address this issue. In both cases, the underlying assumption is that batch training is not optimal for PE tasks. Most machine learning algorithms behind MT and APE are trained offline, which impedes any form of interaction, as adding new attributes involves retraining the model from scratch. Online learning, on the other hand, allows for handling continuous streams of data and updating the model parameters on the fly. This approach thus appears more appropriate for implementing or simulating interactive PE settings. Furthermore, it should be noted that the objective of the approaches presented here is not to improve APE but rather to enhance the PE experience.

4.1 Fully Interactive PE Model

The first option is to design an IPE environment. As it seems very unlikely to produce fully automatic translations, an IPE setting would enable a PE system to learn from human corrections in real time. To the best of our knowledge, such a setting has not yet been proposed in research, as HCI is typically simulated in APE (e.g. [21, 24]). This approach would require implementing an interface to collect human corrections. Such a system would be similar to an IMT model (e.g. CASMACAT), but it would not require continuously updating the MT system parameters to make translation predictions. Instead, it would incrementally learn real-time PE actions to suggest corrections.

Investigating the benefits of such an interactive scenario would be insightful to design more effective IPE environments. Several alternatives are conceivable for IPE, as the translator can be shown either autocompletion suggestions which are updated when new prefixes are entered (this would be similar to IMT) or an entire MT correction suggestion which can be accepted or edited.

Consequently, it appears relevant to study the number of edits necessary to make the system responsive enough as well as to examine translators' experiences. More specifically, comparing different settings, such as a single correction suggestion versus a list of n-best suggestions, and other adjustments (e.g. the length of predictions), as suggested in by Barrachina et al. [25], would be beneficial. Measuring the effort (e.g. time and keystroke logging as in [26]) in each case would also help to formulate IPE models which would suit user needs. Nevertheless, this would require building a rather complex system, which might be difficult due to time and cost constraints.

4.2 Online APE Model

The second option is to simulate human corrections by adopting an online learning method using pre-existing post-edited texts (such as the datasets made available for the WMT APE shared tasks). While this has been done in previous work, only Negri et al. [24] have implemented this method on NMT outputs. This therefore leaves room for further experiments. It should be noted that Negri et al. worked with eSCAPE (Synthetic Corpus for Automatic Post-Editing [27]), a synthetic corpus designed to train APE systems. While eSCAPE is a valuable resource (it is freely available and contains over 7 million triplets), the post-edited segments are artificial (they are pre-existing translations). This is understandable due to the scarcity of training data for APE. However, since the MT output and its corresponding 'post-edited' version are likely to be very different (or at least more distant than in a real PE scenario), a system created using this might be prone to overcorrections. Therefore, it seems relevant to examine the performance of a similar system using training data in which the PE side is comprised of corrections performed by human translators. As pointed out by Ortiz-Martínez [28], further research on online APE would also benefit from corpora containing documents with a high rate of repetitions, which could serve to examine the performance of online APE models in technical translation.

5 Conclusion

This paper discussed how translation technologies could benefit from the creation of more interactive PE environments. Two methods were proposed, both differing from traditional batch training approaches. The first method consists of creating an interactive environment for PE, and the second entails the use of already-available post-edited translations to simulate human interaction. In both cases, the systems would be based on incremental learning and would learn from corrections in a continuous feedback loop, thus suggesting corrections based on the PE actions observed.

References

1. Hutchins, J.: The first public demonstration of machine translation: the Georgetown-IBM system, 7th January 1954 (2005).
2. Knight, K., Chander, I.: Automated postediting of documents. In: Proceedings of AAAI 1994, pp. 779-784. AAAI Press, Seattle, Washington, USA (1994).
3. Allen, J., Hogan, C.: Toward the development of a post editing module for raw machine translation output: a controlled language perspective. In: Kuhn, T., Fuch, N. E. (eds.) Proceedings of the third International Controlled Language Applications Workshop (CLAW), pp. 62-71. Springer, Zurich, Switzerland (2000).
4. Llitjós, A. F., Carbonell, J. G.: Automating post-editing to improve MT systems. Institute for Software Research (2006).

5. do Carmo, F., Shterionov, D., Moorkens, J., Wagner, J., Hossari, M., Paquin, E., Schmidtke, D., Groves, D., Way, A.: A review of the state-of-the-art in automatic post-editing. Machine Translation, pp. 1-43 (2020).
6. Junczys-Dowmunt, M.: Are we experiencing the golden age of automatic post-editing? In: Proceedings of the AMTA 2018: Workshop on Translation Quality Estimation and Automatic Post-Editing, pp. 144-206. Publisher, Boston, Massachusetts, USA (2018).
7. Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huang, S., Huck, M., Koehn, P., Liu, Q., Logacheva, V., Monz, C., Negri, M., Post, M., Rubino, R., Specia, L., Turchi, M.: Findings of the 2017 conference on machine translation (WMT17). In: Proceedings of the Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pp. 169–214. Association for Computational Linguistics, Copenhagen, Denmark (2010).
8. Chatterjee, R., Negri, M., Rubino, R., Turchi, M.: Findings of the WMT 2019 shared task on automatic post-editing. In: Proceedings of the Third Conference on Machine Translation, Volume 3: Shared Task Papers, pp. 710–725. Association for Computational Linguistics, Belgium, Brussel (2018).
9. Foster, G., Langlais, P., Lapalme, G.: User-friendly text prediction for translators. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 148-155. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA (2002).
10. Esteban, J., Lorenzo, J., Valderrábanos, A. S., Lapalme, G.: Transtype2 – an innovative computer-assisted translation system. In: Proceedings of the ACL Interactive Poster and Demonstration Sessions, pp. 94-97. Association for Computational Linguistics, Barcelona, Spain (2004).
11. Toselli, A. H., Vidal, E., Casacuberta, F.: Multimodal interactive pattern recognition and applications. Springer Science & Business Media (2011).
12. Alabau, V., Bonk, R., Buck, C., Carl, M., Casacuberta, F., García-Martínez, M., González, J., Koehn, P., Leiva, L., Mesa-Lao, B., Ortiz, D., Saint-Amand, H., Sanchis, G., Tsoukala, C.: CASMACAT: An open-source workbench for advanced computer aided translation. The Prague Bulletin of Mathematical Linguistics, 100, pp. 101-112 (2013).
13. Wuebker, J., Green, S., DeNero, J., Hasan, S., Luong, M. T.: Models and inference for prefix-constrained machine translation. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp. 66-75. Association for Computational Linguistics, Berlin, Germany (2016).
14. Knowles, R., Koehn, P.: Neural interactive translation prediction. In: Green, S., Schwartz, L. (eds.) Proceedings of AMTA 2016, vol. 1: Researchers' track, pp. 107-120. Association for Machine Translation in the Americas, Austin, Texas, USA (2016).
15. Santy, S., Dandapat, S., Choudhury, M., Bali, K.: INMT: Interactive neural machine translation prediction. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations, pp. 103-108. Association for Computational Linguistics, Hong Kong, China (2019).
16. Peris, Á., Casacuberta, F.: Online learning for effort reduction in interactive neural machine translation. Computer Speech & Language, 58, pp. 98-126 (2019).
17. Casacuberta, F., Civera, J., Cubel, E., Lagarda, A. L., Lapalme, G., Macklovitch, E., Vidal, E. Human interaction for high-quality machine translation. Communications of the ACM, 52(10), pp. 135-138 (2009).
18. Underwood, N. L., Mesa-Lao, B., García-Martínez, M., Carl, M., Alabau, V., González-Rubio, J., Leiva, J. A., Sanchis-Trilles, G., Ortíz-Martínez, D, Casacuberta, F. Evaluating

the effects of interactivity in a post-editing workbench. In: Proceeding of LREC, pp. 553-559. European Language Resources Association, Reykjavik, Iceland (2014).

19. Alves, F., Koglin, A., Mesa-Lao, B., García Martínez, M., de Lima Fonseca, N. B., de Melo Sá, A., Gonçalves, J. L., Sarto Szpak, K., Sekino, K., Aquino, M. Analysing the impact of interactive machine translation on post-editing effort. New directions in empirical translation process research, pp. 77-94. Springer, Cham (2016).
20. Lagoudaki, E.: The value of machine translation for the professional translator. In: Proceedings of the 8th Conference of the Association for Machine Translation in the Americas, pp. 262-269. Association for Machine Translation in the Americas, Waikiki, Hawaii, USA (2008).
21. Simard, M., Foster, G.: Pepr: Post-edit propagation using phrase-based statistical machine translation. In: Sima'an, K., Forcada, M.L., Grasmick, D., Depraetere, H., Way, A. (eds.) Proceedings of the XIV Machine Translation Summit, pp. 191-198. International Association for Machine Translation and the European Association for Machine Translation, Nice, France (2013).
22. Lagarda, A. L., Ortiz-Martínez, D., Alabau, V., Casacuberta, F.: Translating without in-domain corpus: Machine translation post-editing with online learning techniques. Computer Speech & Language, 32(1), pp. 109-134 (2015).
23. Chatterjee, R., Gebremelak, G., Negri, M., Turchi, M. Online automatic post-editing for MT in a multi-domain translation environment. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pp. 525-535. Association for Computational Linguistics, Valencia, Spain (2017).
24. Negri, M., Turchi, M., Bertoldi, N., Federico, M.: Online neural automatic post-editing for neural machine translation. In: Cabrio, E., Mazzei, A., Tamburini, F. (eds.) In: Proceedings of the 5th Italian Conference on Computational Linguistics. CEUR-WS, Torino, Italy (2018).
25. Barrachina, S., Bender, O., Casacuberta, F., Civera, J., Cubel, E., Khadivi, S., Lagarda, A., Ney, H., Tomás, J., Vidal, E., Vilar, J. M. Statistical approaches to computer-assisted translation. Computational Linguistics, 35(1), pp. 3-28 (2009).
26. Aziz, W., de Sousa, S. C. M., Specia, L.: PET: a tool for post-editing and assessing machine translation. In: Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC), pp. 3982–3987. European Language Resources Association, Istanbul, Turkey (2012).
27. Negri, M., Turchi, M., Chatterjee, R., Bertoldi, N.: eSCAPE: a Large-scale Synthetic Corpus for Automatic Post-Editing. In: Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC). European Language Resources Association, Miyazaki, Japan (2018).
28. Ortiz-Martínez, D. Online learning for statistical machine translation. Computational Linguistics, 42(1), pp. 121-161 (2016).

Using CollGram to Compare Formulaic Language in Human and Machine Translation[⋆]

Yves Bestgen[1][0000−0001−7407−7797]

Université catholique de Louvain, 10 place Cardinal Mercier, Louvain-la-Neuve, 1348, Belgium yves.bestgen@uclouvain.be

Abstract. A comparison of formulaic sequences in human and neural machine translation of quality newspaper articles shows that neural machine translations contain less lower-frequency, but strongly-associated formulaic sequences, and more high-frequency formulaic sequences. These differences were statistically significant and the effect sizes were almost always medium or large. These observations can be related to the differences between second language learners of various levels and between translated and untranslated texts. The comparison between the neural machine translation systems indicates that some systems produce more formulaic sequences of both types than other systems.

Keywords: Neural machine translation · Multiword unit · Lexical association indice.

1 Introduction

Neural machine translation (NMT) systems are currently considered to bridge the gap between human and machine translation [22, 26]. However, little research has been done to determine whether NMT systems are also very effective in processing multiword units [20, 27], whereas the importance of preformed units in language use is now well established, including in foreign language learning and translation [1, 21, 24]. The present study addresses this issue by comparing formulaic language in human and neural machine translation. It focuses on a specific category of multiword units, the "habitually occurring lexical combinations" [17], such as *dramatic increase*, *depend on*, *out of*, which are not necessarily semantically non-compositional, but are considered statistically typical of the language because they occur "with markedly high frequency, relative to the component words or alternative phrasings of the same expression" [2]. These formulaic sequences (FSs) are analyzed by means of a technique proposed by [4], improved by [13] and automated by [10] under the name *CollGram*[1].

[⋆] The author is a Research Associate of the Fonds de la Recherche Scientifique - FNRS (Fédération Wallonie Bruxelles de Belgique). He would like to warmly thank Sylviane Granger and Maïté Dupont for access to the PLECI corpus.

[1] As one reviewer pointed out, this automation is not an "easily available plug-and-play implementation". However, there is a freely available system that implements it

Translation and Interpreting Technology Online, pages 174–180,
July 05-07, 2021.
https://doi.org/10.26615/978-954-452-071-7_020

CollGram rests on two lexical association indices that measure the strength of attraction between the words that compose a bigram, mutual information (MI) and t-score [14], calculated on the basis of the frequencies of occurrences in a reference corpus [10, 13]. These two indices are complementary, MI favoring lower-frequency, but strongly-associated, FSs such as *self-fulfilling prophecy*, *sparsely populated* or *sunnier climes* while the t-score favors high-frequency bigrams such as *you know*, *out of* or *more than*. A series of studies have shown that, compared to native speakers, English as a foreign language learners tend to underuse collocations with high MI scores, while overusing those with high t-scores and that exactly the same differences are observed between advanced learners and intermediate learners [10, 13]. These observations are in agreement with usage-based models of language learning which ”hold that a major determining force in the acquisition of formulas is the frequency of occurrence and co-occurrence of linguistic forms in the input” [13]. It is worth noting that the same differences were observed between translated and untranslated texts, but the proposed explanation relies on a tendency towards normalization in translation [5, 8]. Since neural models also seem to be affected by frequency of use [15, 19], the hypothesis tested in the present study is that the same effects could be observed when comparing human translations (HTs) and NMTs, namely that NMTs will underuse high MI FSs and overuse high t-score FSs.

2 Method

2.1 Translation Corpus

The texts used are taken from the journalistic section of the PLECI corpus (uclouvain.be/en/research-institutes/ilc/cecl/pleci.html). It is a sentence-aligned translation corpus of quality newspaper articles written in French and published in *Le Monde diplomatique* and in English in one of the international editions of this same newspaper. Two hundred and seventy-nine texts, published between 2005 and 2012, were used for a total of 570,000 words in the original version and of 500,000 words in the translation.

All original texts were translated into English by three well-known NMT systems: DeepL (deepl.com/translator), Google Translate (translate.google.com) and Microsoft Translator (microsoft.com/translator). Online translators were used for the first two, while the version available in *Office 365* was used for the third. All these translations were performed between March 24 and April 6, 2021.

(http://collgram.pja.edu.pl) [18, 25]. Some of the indices used can also be easily obtained with the TAALES software [16] which allows the automatic analysis of many other lexical indices. TAALES presents however an important limitation because it only takes into account bigrams that occur at least 51 times in the reference corpus [9], a value much too high for the MI at the heart of the CollGram approach.

2.2 Procedure

Each translated text was tokenized and POS-tagged by CLAWS7 [23] and all bigrams were extracted. Punctuation marks and any character sequences that did not correspond to a word interrupted the bigram extraction. Each bigram, which did not include any proper name or number according to CLAWS, was then searched for in the 100 million word British National Corpus (BNC[2], www.natcorp.ox.ac.uk). When it is present, the corresponding MI and t-score were used to decide whether it is highly collocational or not. Based on [8] and [13], bigrams with a score greater than or equal to 5 for the MI and 6 for the t-score were considered highly collocational. The last step consisted in calculating, for each text and for each association index, the percentage that the bigrams considered as highly collocational represent compared to the total number of bigrams present in the text.

3 Analyses and Results

Table 1 shows the average percentages of highly collocational bigrams for the MI and t-score in the four type of translations. The four means for each measure of association were analyzed using the Student's test for repeated measures since the same texts, which are the unit of analysis, were translated by the four translators. All these comparisons were statistically significant ($p < 0.0001$).

Table 1. Average percentages of highly collocational bigrams for the two indices in the four translation types.

Measure	Human	DeepL	Google	Microsoft
High MI	11.21	10.48	10.07	10.27
High t-score	58.76	60.60	59.49	59.89

Table 2 presents the differences between the means as well as two effect sizes. The first is Cohen's d, which expresses the size of the difference between the two means as a function of the score variability. According to [11], a d of 0.50 indicates a medium effect and that a d of 0.80 a large effect. The second effect size indicates the percentage of texts for which the difference between the two translations has the same sign as the mean difference. A value of 100 means that all texts produced by a given translator have a higher score than those translated by the other one and a value of 50 means that there is no difference.

As shown in these tables, both hypotheses are verified. Compared to HTs, texts translated by the three neural systems contain a significantly smaller percentage of highly collocational bigrams for the MI and a larger percentage

[2] [6] showed that CollGram produces the same results if another reference corpus, such as COCA (corpus.byu.edu/coca) or WaCKy [3], is used.

of highly collocational bigrams for the t-score. Cohen's *d*s are almost always medium or large and the percentages of texts for which differences are observed are greater than 70% except in one case.

Table 2. Differences (row translator minus column translator) and effect sizes for the two indices in the four translation types.

	Human			DeepL			Google		
	D	Es	%	D	Es	%	D	Es	%
				High MI					
DeepL	-0.72	0.59	73.48						
Google	-1.14	1.00	84.33	-0.41	0.65	74.91			
Microsoft	-0.94	0.77	81.00	-0.21	0.35	62.72	0.20	0.36	64.52
				High t-score					
DeepL	1.83	0.84	80.65						
Google	0.72	0.32	62.37	-1.11	0.98	84.59			
Microsoft	1.13	0.51	71.33	-0.70	0.62	70.97	-0.41	0.42	69.18

An analysis of the passages in which the differences between HT and NMT are the largest suggests that the origin lies at least partially in the less literal nature of human translations (see Table 3 for an example).

Table 3. Example of the four translation types and percentages of highly collocational bigrams for the two indices.

Type	Phrase	%High MI	%High t-score
Original	A raison de huit heures par jour		
Human	In an eight-hour day	67	33
DeepL	At eight hours a day	25	100
Google & Microsoft	At the rate of eight hours a day	14	100

The differences between the three NMT systems are smaller, but still statistically significant. However, they require a different interpretation. When NMTs are compared to HTs, the patterns of differences are reversed according to the MI or the t-score, as expected. For the NMT systems, these patterns are identical for both types of collocation. The average percentages of highly collocational bigrams (see Table 1) are always higher in texts translated by DeepL than in those translated by Microsoft and also higher in the latter than in those translated by Google. Only a detailed qualitative analysis could determine whether these results indicate a difference in effectiveness.

4 Discussion and Conclusion

The reported analyses confirm the hypotheses and thus suggest that, compared to HTs, NMTs more closely resemble texts written by intermediate learners than by advanced learners of English as a foreign language, a result that could be interpreted in the context of a usage-based model of language learning [13]. The NMTs also resemble translated texts more than untranslated texts, but it is not clear that this can be explained by a normalization process. Statistically significant differences, but smaller in terms of effect size, were also observed between the three NMT systems.

It is important to keep in mind that the present study only considers global quantitative properties of MWUs. At no point is the appropriateness in context of the MWUs assessed. It is therefore a very partial approach. However, it has the advantage of not requiring a human qualitative evaluation that is often complicated and cumbersome to set up. Moreover, it is likely that the appropriateness of a MWU is much more important for non-compositional expressions than for the habitually occurring lexical combinations studied here [12].

Another important feature of the approach is that it relies on a native reference corpus to identify highly collocational bigrams for both indices. As already mentioned, research on foreign language learning, but also on the comparison of translated and untranslated texts, has shown that the use of other large reference corpora such as COCA or WaCKy [3] did not change the results [5, 6]. One can also wonder whether the use of a comparable reference corpus, rather than a generic one, would have returned different results. In the case of the comparison of translated and untranslated texts, [5] observed that the use of a journalistic corpus, the *Corpus Est Républicain* (115 million words) made available by the Centre National de Ressources Textuelles et Lexicales, produced differences similar to those obtained with the WaCKy corpus.

Before considering taking advantage of these observations to try to improve NMT systems, a series of complementary analyses must be conducted. Indeed, this study has many limitations, such as focusing only on a subcategory of MWUs [20], on a single language pair, and on a single genre of texts. Moreover, a thorough qualitative analysis is essential to better understand the results and evaluate the proposed explanations. As it has been shown in foreign language learning [7], it would also be interesting to verify that the observed effects are not explained by differences in single-word lexical richness. Finally, the differences between the three NMT systems also require further analysis.

References

1. Baker, M.: Patterns of idiomaticity in translated vs. non-translated text. Belgian Journal of Linguistics **21**, 11–21 (2007)
2. Baldwin, T., Kim, S.N.: Multiword expressions. In: Indurkhya, N., Damerau, F.J. (eds.) Handbook of Natural Language Processing, pp. 267–292. CRC Press (2010)
3. Baroni, M., Bernardini, S., Ferraresi, A., Zanchetta, E.: The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. Language Resources and Evaluation **43**, 209–226 (2009)

4. Bernardini, S.: Collocations in translated language. combining parallel, comparable and reference corpora. In: Proceedings of the Corpus Linguistics Conference. pp. 1–16. Lancaster University (2007)
5. Bestgen, Y., Granger, S.: Collocation et traduction. analyse automatique au moyen d'indices d'association. In: Kauffer, M., Keromnes, Y. (eds.) Theorie und Empirie in der Phraseologie / Approches théoriques et empiriques en phraséologie, pp. 101–113. Stauffenburg Verlag (2019)
6. Bestgen, Y.: Evaluation automatique de textes. Validation interne et externe d'indices phraséologiques pour l'évaluation automatique de textes rédigés en anglais langue étrangère. Traitement automatique des langues **57**(3), 91–115 (2016)
7. Bestgen, Y.: Beyond single-word measures: L2 writing assessment, lexical richness and formulaic competence. System **69**, 65–78 (2017)
8. Bestgen, Y.: Normalisation en traduction : analyse automatique des collocations dans des corpus. Des mots aux actes **7**, 459–469 (2018)
9. Bestgen, Y.: Evaluation de textes en anglais langue ètrangère et séries phraséologiques : comparaison de deux procédures automatiques librement accessibles. Revue française de linguistique appliquée **24**, 81–94 (2019)
10. Bestgen, Y., Granger, S.: Quantifying the development of phraseological competence in L2 English writing: An automated approach. Journal of Second Language Writing **26**, 28–41 (2014)
11. Cohen, J.: Statistical power analysis for the behavioral sciences. Erlbaum (1988)
12. Constant, M., Eryiğit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M., Todirascu, A.: Survey: Multiword expression processing: A Survey. Computational Linguistics **43**(4), 837–892 (dec 2017)
13. Durrant, P., Schmitt, N.: To what extent do native and non-native writers make use of collocations? International Review of Applied Linguistics in Language Teaching **47**, 157–177 (2009)
14. Evert, S.: Corpora and collocations. In: Lüdeling, A., Kytö, M. (eds.) Corpus Linguistics. An International Handbook, pp. 1211–1248. Mouton de Gruyter (2009)
15. Koehn, P., Knowles, R.: Six challenges for neural machine translation (2017)
16. Kyle, K., Crossley, S., Berger, C.: The tool for the automatic analysis of lexical sophistication (TAALES): version 2.0. Behavior Research Methods **50**, 1030–1046 (2018). https://doi.org/10.3758/s13428-017-0924-4
17. Laufer, B., Waldman, T.: Verb-noun collocations in second language writing: A corpus analysis of learners' English. Language Learning **61**, 647–672 (2011)
18. Lenko-Szymanska, A., Wolk, A.: A corpus-based analysis of the development of phraseological competence in EFL learners using the collgram profile. In: Paper presented at the 7th Conference of the Formulaic Language Research Network (FLaRN) (2016)
19. Li, M., Roller, S., Kulikov, I., Welleck, S., Boureau, Y.L., Cho, K., Weston, J.: Don't say that! making inconsistent dialogue unlikely with unlikelihood training. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 4715–4728. Association for Computational Linguistics (2020)
20. Monti, J., Seretan, V., Pastor, G.C., Mitkov, R.: Multiword units in machine translation and translation technology. In: Mitkov, R., Monti, J., Pastor, G.C., Seretan, V. (eds.) Multiword Units in Machine Translation and Translation Technology, pp. 2–37. John Benjamins (2018)
21. Pawley, A., Syder, F.H.: Two puzzles for linguistic theory: nativelike selection and nativelike fluency. In: Richards, J.C., Schmidt, R.W. (eds.) Language and Communication. Longman (1983)

22. Popel, M., Tomkova, M., Tomek, J., Kaiser, L., Uszkoreit, J., Bojar, O., Zabokrtsky, Z.: Transforming machine translation: A deep learning system reaches news translation quality comparable to human professionals. Nature Communications **11**, 1–15 (2020). https://doi.org/10.1038/s41467-020-18073-9
23. Rayson, P.: Matrix: A statistical method and software tool for linguistic analysis through corpus comparison. Ph.D. thesis, Lancaster University (1991)
24. Sinclair, J.: Corpus, Concordance, Collocation. Oxford University Press (1991)
25. Wołk, K., Wołk, A., Marasek, K.: Unsupervised tool for quantification of progress in L2 english phraseological. In: Proceedings of the 2017 Federated Conference on Computer Science and Information Systems. pp. 383–388 (2017)
26. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Łukasz Kaiser, Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., Dean, J.: Google's neural machine translation system: Bridging the gap between human and machine translation (2016)
27. Zaninello, A., Birch, A.: Multiword expression aware neural machine translation. In: Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020). pp. 3816–3825. European Language Resources Association (ELRA) (2020)

MultiTraiNMT: Training Materials to Approach Neural Machine Translation from Scratch*

Gema Ramírez-Sánchez[1], Juan Antonio Pérez-Ortiz[2], Felipe Sánchez-Martínez[2], Caroline Rossi[3], Dorothy Kenny[4], Riccardo Superbo[5], Pilar Sánchez-Gijón[6], and Olga Torres-Hostench[6]

[1] Prompsit Language Engineering, gema@prompsit.com
[2] Universitat d'Alacant, {japerez,fsanchez}@ua.es
[3] Université Grenoble-Alpes, caroline.rossi@univ-grenoble-alpes.fr
[4] Dublin City University, dorothy.kenny@dcu.ie
[5] KantanMT.com, riccardos@kantanmt.com
[6] Universitat Autònoma de Barcelona,
{pilar.sanchez.gijon,olga.torres.hostench}@uab.cat

Abstract. The aim of the MultiTraiNMT Erasmus+ project is to develop an open innovative syllabus in neural machine translation (NMT) for language learners and translators as multilingual citizens. Machine translation is seen as a resource to provide support to citizens when trying to acquire and develop language skills, provided they are given informed and critical training. Machine translation would thus help tackle the mismatch between the EU aim of having multilingual citizens who speak at least two foreign languages and the current situation in which they generally fall far short of this objective. The training materials consist of an open-access coursebook, an open-source NMT web application (MutNMT) for training purposes and corresponding activities.

Keywords: machine translation · neural machine translation · training · multilingual citizens · project description.

1 Introduction

The aim of the Erasmus+ strategic partnership "MultiTraiNMT – Machine Translation Training for Multilingual Citizens"[7] (2019–2022) is to develop, evaluate and disseminate open-access materials and open-source applications that will lead to the enhancement of teaching and learning about machine translation [3] among language learners, language teachers, trainee translators, translation teachers and professional translators across Europe.

MultiTraiNMT brings together experts at four European universities — Universitat Autònoma de Barcelona, Université Grenoble–Alpes, Dublin City University and Universitat d'Alacant, and two enterprises — Prompsit Language

[*] With the support of the Erasmus+ programme of the European Union.
[7] Project website: https://multitrainmt.eu

Translation and Interpreting Technology Online, pages 181–186,
July 05-07, 2021.
https://doi.org/10.26615/978-954-452-071-7_021

Engineering and Xcelerator Machine Translations, and is supported by numerous associate partners in education and the translation industry, all of whom are interested in teaching and learning about the use of machine translation. The partnership aims specifically at developing an innovative syllabus in machine translation (MT) in general, and neural machine translation (NMT) in particular [3]. On completion we will provide the following components which will be described in the following sections:

– An **open-access coursebook** that addresses both the technical foundations of machine learning—and especially deep learning—as used in MT, and the ethical, societal and professional implications of this technology.
– MutNMT, a **pedagogical NMT web application** that allows users to learn how NMT works, and gain insight into the internal workings of NMT systems.
– **Learning activities** related to the coursebook and MutNMT that allow language learners and translators to co-construct knowledge on NMT.

The training is designed to be followed in both asynchronous and synchronous forms. On the one hand, self-learners will be able to follow the coursebook and perform the corresponding activities. On the other hand, any interested teacher will be able to use the course and the activities in synchronous form with students. The coursebook and learning activities are designed taking into account different progress levels to approach different student profiles. Measurability, quality and progress of the project can be followed in the project website.

In short, we are developing an up-to-date syllabus on MT for use in European Higher Education and elsewhere; one that will allow students to acquire the technical and ethical skills and competences required to become informed, critical users of contemporary MT in their own language learning and translation practice. In so doing, we open up the world of machine learning to language and translation students, their teachers and others, enhancing their ability to function as technologically competent, informed citizens in a multilingual Europe.

2 The coursebook

The Creative Commons-licensed[8] open-access coursebook is organised in eight chapters. Instructors may conveniently arrange them in a different order for their courses. The modules are:

1. Multilingualism
2. Introduction to machine translation

[8] https://creativecommons.org

3. How to choose a suitable MT system and evaluation of machine translation quality
4. How to prepare and select texts for machine translation
5. How to deal with machine translation mistakes, post-editing and error fixing
6. Ethical aspects of machine translation
7. How neural machine translation works
8. Custom neural machine translation

3 The web application

MutNMT is an open-source web application[9] to train NMT for didactic purposes. It lets users train, inspect, evaluate and translate using neural engines. Contributions to other open-source projects have also been made, namely to JoeyNMT [4], a command-line tool to train NMT engines. Technical documentation is provided along with the code. Manuals for both users and instructors are available. A production installation of MutNMT is currently under evaluation.[10]

There are three profiles of users: *beginners* with default access to all basic features; *experts* with access to basic and advanced features without administration rights; and *admins* with full access to all features. Admins are able to upgrade beginners (default profile) to experts or admins by request at any time. A brief description of the main features and windows of MutNMT follows:

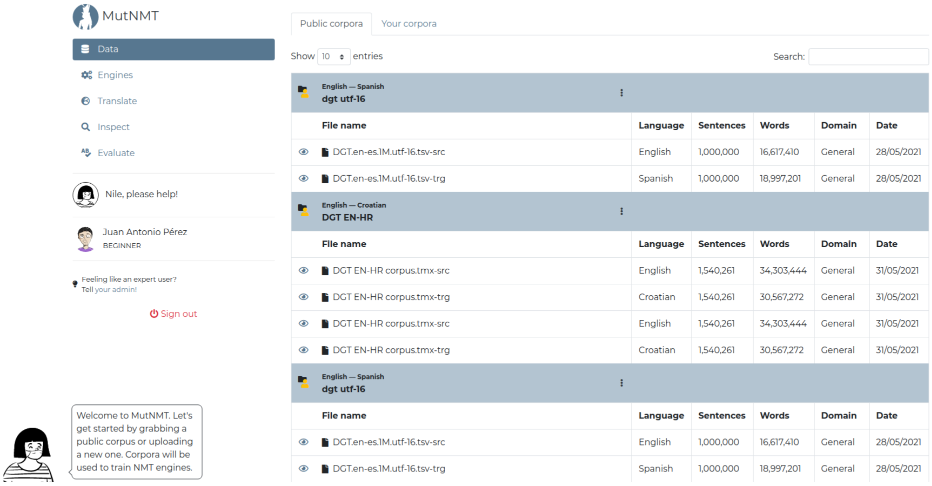
Data. MutNMT, as every other NMT system, needs corpora in the form of parallel data to learn from. Previewing, downloading and grabbing corpora is possible. Corpora uploaded to MutNMT retain their original licences. While free/open source corpora are recommended in MutNMT, users are allowed to upload proprietary corpora and keep them private. These corpora are shown in the application as a collection of resources as shown in figure 1.

Engines. There is also a library of engines in MutNMT, that is, already available MT systems that have been trained and shared. Of special interest are also the actions allowed: seeing the full training log of an engine, downloading the model or downloading the corpora it was trained with. Models created with MutNMT come wiht a GPL-v3 free/open-source license. While beginners can only see training reports, experts and admins are able to resume the training of an engine.

Training. This is an advanced feature for experts and admins that allows them to train NMT engines using MutNMT. Users will need to set up engine details, configuration parameters and select corpora for training a particular system.

[9] Code available on `https://github.com/Prompsit/mutnmt`

[10] `https://ntradumatica.uab.cat`



Welcome to MutNMT. Let's get started by grabbing a public corpus or uploading a new one. Corpora will be used to train NMT engines.

English — Spanish
dgt utf-16

File name	Language	Sentences	Words	Domain	Date
DGT.en-es.1M.utf-16.tsv-src	English	1,000,000	16,617,410	General	28/05/2021
DGT.en-es.1M.utf-16.tsv-trg	Spanish	1,000,000	18,997,201	General	28/05/2021

English — Croatian
DGT EN-HR

File name	Language	Sentences	Words	Domain	Date
DGT EN-HR corpus.tmx-src	English	1,540,261	34,303,444	General	31/05/2021
DGT EN-HR corpus.tmx-trg	Croatian	1,540,261	30,567,272	General	31/05/2021
DGT EN-HR corpus.tmx-src	English	1,540,261	34,303,444	General	31/05/2021
DGT EN-HR corpus.tmx-trg	Croatian	1,540,261	30,567,272	General	31/05/2021

English — Spanish
dgt utf-16

File name	Language	Sentences	Words	Domain	Date
DGT.en-es.1M.utf-16.tsv-src	English	1,000,000	16,617,410	General	28/05/2021
DGT.en-es.1M.utf-16.tsv-trg	Spanish	1,000,000	18,997,201	General	28/05/2021

Fig. 1. A screen capture of the MutNMT's window that allows users to preview, add, and download parallel corpora.

Translate. By using the available engines, all users will be able to copy and paste a series of sentences and translate them. They will get the resulting translation in a text box and be able to export a TMX [1] out of the whole translation, thus saving pairs of source and target sentences in a standard format. Document translation is also supported and will give as a result the translation in either the original document format or in a TMX file.

Inspect. There are several options in this window, all aimed at seeing the internals of the translation engines at work such as allowing users to input a sentence and see it at different steps of processing by a particular engine: pre-processed input, hypothesis generation (n-best), pre-final output (still to be post-processed) and final output.

Evaluate. As a final step, users will be able to evaluate the output of MT comparing it to other machine translated texts or to professional human translations. MutNMT provides several automatic evaluation metrics, such as BLEU [5] and ChrF3 [6], at both document and sentence levels. All these results can also be downloaded in spreadsheet format.

4 The learning activities

Two types of learning activities are being created. On the one hand, self-learning questions are aimed at students working at their own pace; these are short-answer questions with immediate automatic feedback. On the other hand, open-answer teacher-guided activities can be customised and adapted to different contexts. After exploring different formats and repositories of learning objects, we have opted for the open-source H5P platform,[11] as it allows each of our activities to be self-contained and easily embeddable by instructors in learning management systems such as Moodle[12] or more general environments such as Wordpress[13]. Each exercise includes metadata such as difficulty, estimated answering time, comments for instructors and considerations when adapting the text of the question to other language combinations.

5 The MultitraiNMT associate partner network

MultiTraiNMT invites higher education institutions and teachers of translation and foreign languages to join the project as associate partners/members. In order to become an associate partner, interested parties may visit the website section *Join us / Become a member* in order to download the *Associate Partners Agreement* and adapt it to their needs and interests. The aim of the network is not only to share the aforementioned coursebook, MutNMT and activities but also to create a working group to share activities, experience and best practices so that the project becomes collaborative. The partners may:

- Evaluate the use of the project coursebook in their classes.
- Test the MutNMT educational system for managing NMT engines for didactic purposes.
- Participate with project partners in the piloting of project activities on MT training or voluntary sharing of MT training activities in the project.
- Arrange with the MultitraiNMT project the certification of participants.
- Participate actively in the project multiplier events or other dissemination events.
- Participate in any other training or research activity which fosters the development of MT skills in general among multilingual citizens.

6 Conclusions

Despite recent advances in freely available NMT engines, machine translation is still often considered too complex to be understood by a non-specialist audience. The materials developed within the MultiTraiNMT project are intended to show that MT literacy can be developed across various target audiences, in line with recent proposals [2]. We have designed a course that includes activities and a platform (MutNMT) for learning NMT by doing, and we invite collaborations within our network of associate partners.

[11] `https://h5p.org/`

[12] `https://moodle.org/`

[13] `https://wordpress.org/`

References

1. Localisation industry standards: Translation Memory eXchange (TMX). `https://www.etsi.org/deliver/etsi_gs/LIS/001_099/002/01.04.02_60/gs_LIS002v010402p.pdf` (2013), accessed: 2021-06-24.
2. Bowker, L., Buitrago, J.: Machine translation and global research: Towards improved machine translation literacy in the scholarly community. Emerald Group Publishing (2019)
3. Koehn, P.: Neural machine translation. Cambridge University Press (2020)
4. Kreutzer, J., Bastings, J., Riezler, S.: Joey NMT: A minimalist NMT toolkit for novices. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations. pp. 109–114 (2019)
5. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318. Association for Computational Linguistics (2002)
6. Popović, M.: chrF: character n-gram F-score for automatic MT evaluation. In: Proceedings of the Tenth Workshop on Statistical Machine Translation. pp. 392–395. Association for Computational Linguistics, Lisbon, Portugal (Sep 2015)

The Post-Editing Workflow: Training Challenges for LSPs, Post-Editors and Academia

Viveta Gene[0000-0002-4374-8305]

Ionian University, Corfu, Greece

viveta.gene@gmail.com

Abstract. Language technology is already largely adopted by most Language Service Providers (LSPs) and integrated into their traditional translation processes. In this context, there are many different approaches to applying Post-Editing (PE) of a machine translated text, involving different workflow processes and steps that can be more or less effective and favorable. In the present paper, we propose a 3-step Post-Editing Workflow (PEW). Drawing from industry insight, this paper aims to provide a basic framework for LSPs and Post-Editors on how to streamline Post-Editing workflows in order to improve quality, achieve higher profitability and better return on investment and standardize and facilitate internal processes in terms of management and linguist effort when it comes to PE services. We argue that a comprehensive PEW consists in three essential tasks: Pre-Editing, Post-Editing and Annotation/Machine Translation (MT) evaluation processes (Guerrero, 2018) supported by three essential roles: Pre-Editor, Post-Editor and Annotator (Gene, 2020). Furthermore, the present paper demonstrates the training challenges arising from this PEW, supported by empirical research results, as reflected in a digital survey among language industry professionals (Gene, 2020), which was conducted in the context of a Post-Editing Webinar. Its sample comprised 51 representatives of LSPs and 12 representatives of SLVs (Single Language Vendors) representatives.

Keywords: Post-editing workflow; training challenges; pre-editing; error annotation.

1 Introduction

The role of the post-editor was first mentioned almost thirty years ago. However, the skills, competences, tasks and processes related to this role need to be revised in light of the rising significance of MT (Rico, 2017) and the translation workflow needs to be human-centered to offer advantages for all stakeholders (Guerrero, 2018).

The use of Translation Memories (TMs) and MT in the localization workflow paved the way for the exploration of new methods of producing higher-volume translations at lower costs while maintaining quality. This resulted in new translation workflows including pre-editing and PE of raw output and the creation of new guidelines to support the work in this environment and the training of translators (Guerberof, 2017).

According to Guerberof (2017), even if the pre-editing and post-editing of raw output have been implemented in some organizations since the 1980s, it is only in the last ten years that Machine Translation Post-Editing (MTPE) has been introduced as a

Translation and Interpreting Technology Online, pages 187–198,
July 05-07, 2021.
https://doi.org/10.26615/978-954-452-071-7_022

part of the standard translation workflow in most localization agencies worldwide (Lommel & DePalma 2016a). As a result, the need of training in the ways that technology affects the standard translation workflow and the agents involved in the process becomes a priority.

2 The Post-Editing Workflow

In a digital survey among language industry professionals (Gene, 2020), which was conducted in the context of a Post-Editing Webinar, one of the questions was about the workflow in use by LSPs for Post-Editing. The results hereby presented reflect the answers of 51 LSPs (see Fig. 1):

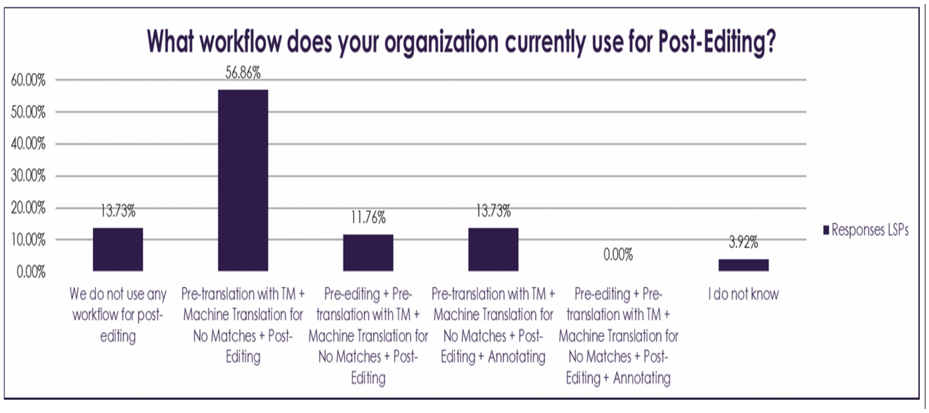


Fig. 1. GALA Survey, Question on the LSPs PEW

The responses to this question reveal the training challenges related to the PEW. They indicate that TM and MT technology are now well established in the workflows of LSPs, although LSPs and all other stakeholders (Post-Editors, Academia and Clients) either lack the training and budget needed or the training and adequate volumes/relevant nature of projects to be able to implement or make the most of a post-editing workflow.

In this paper, we argue that a comprehensive PEW consists in three human-centered tasks (see Fig. 3): Pre-Editing, Post-Editing and Quality Checks/Annotation/MT evaluation processes (Gene, 2020).

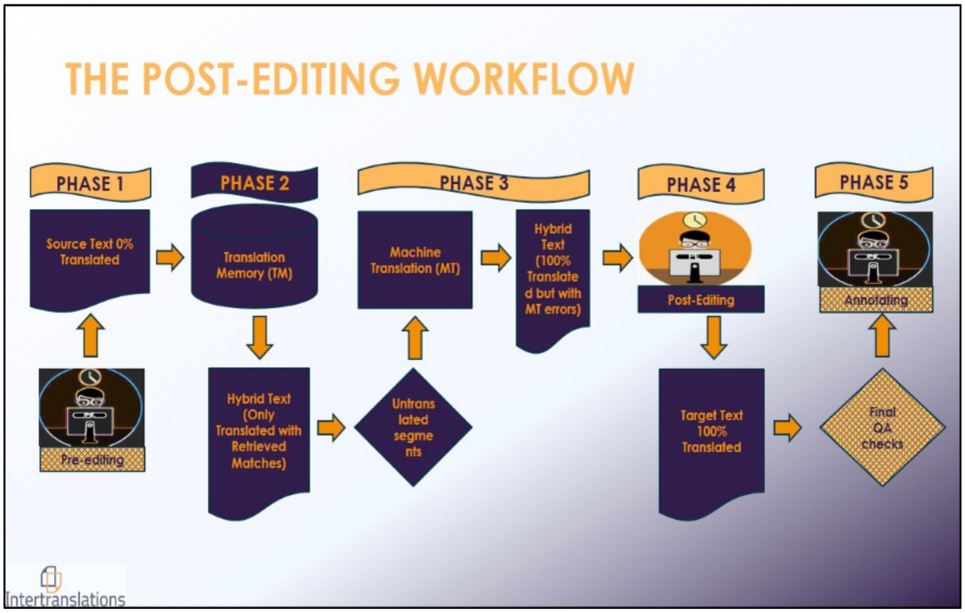

THE POST-EDITING WORKFLOW
PHASE 1
PHASE 2
PHASE 3
PHASE 4
PHASE 5
Source Text 0% Translated
Translation Memory (TM)
Machine Translation (MT)
Hybrid Text (100% Translate d but with MT errors)
Post-Editing
Annotating
Pre-editing
Hybrid Text (Only Translated with Retrieved Matches)
Untrans lated segme nts
Target Text 100% Translated
Final QA checks
Intertranslations

Fig. 1. The Post-Editing Workflow

It should be noted that not all translation projects are good candidates for a comprehensive 3-step PEW (O'Brien, 2003), as this depends on the language combination, specialization field, use-case scenario, nature, volume, budget, deadline and a collaborative post-editing protocol established between the LSP and the Client. This section provides the definitions of the three aforementioned human-centered tasks from an industrial as well as from an academic point of view as the starting point for research in the language industry.

2.1 Pre-editing

Pre-editing is performed using a set of terminological and stylistic guidelines or rules for the pre-processing of the source text before any translation automation in the scope of improving the raw output quality, therefore reducing the effort required on the part of the linguists to post-edit said output and increasing their productivity (Sanchez-Martinez, 2012). Besides, errors in the source text may prevent the translation system from finding the best matches for each segment (Guerrero, 2018).

According to O'Brien (2003), pre-editing is intended for very specific domains and companies. When combined with controlled language rules which aim to simplify and polish the source text, pre-editing ensures that the material to be processed is in the optimal condition to allow for the optimal quality of MT output production and minimal effort in the PE phase, providing lexical clarity and simplification of complex grammatical structures (Sanchez-Martinez, 2012). The goal is, in other words, to end up with a source that will be well "understood" by both the reader and the MT engine.

In this aspect, we may consider “controlled language” as a way of expression that is more compatible with the MT engine’s own “language”.

For optimal results, these methods should be combined with other technology tools: Computer-Assisted Translation (CAT) tools should be integrated in the post-editing workflow in a way that complements machine-generated results and ensures a superior level of accuracy and consistency in terms of terminology and domain-specific jargon.

After the aforementioned pre-editing steps are completed, the untranslated text is connected to an existing Translation Memory (TM) to leverage matches from previously translated content. This process results in a text partially translated only by TM-retrieved segments. At this point, MT comes into play, with lowest scoring untranslated strings (0-50% matches) going through the engine to be translated.

The final product is an entirely translated text, which, however, contains MT errors that are going to be processed during the actual Post-Editing stage of the workflow. This hybrid text, combining matches retrieved from human-generated TM and (preferably customized) machine-translated output, is what should ideally reach the Post-Editor, in order to ensure minimal editing effort and superior linguistic quality.

2.2 Post-editing

Post-editing is to edit, modify and/or correct pre-translated text that has been processed by an MT system from a source language into (a) target language(s) (Allen, 2003). In the comprehensive PEW described here (see Fig. 2), Post-Editing follows the pre-editing step.

The focus here is on the linguist’s tasks. These involve being able to quickly evaluate the text’s elements (whether deriving from TM suggestions or MT output) and correct or eliminate errors, add any missing elements or remove redundant ones, while paying attention to terminology and to the text’s overall fluency and style. This requires a rather clear understanding of the way the specific type of MT engine at hand functions and which kinds of errors it is most likely to generate. It is worth noting that in recent times, the focus of linguists and post-editing experts (and, on a broader level, the focus of Post-Editing training) is practically on NMT systems, which are all the more widely used and have replaced SMT or rule-based systems almost entirely (Blagodarna, 2018).

To summarize the points made so far, the key to this stage of the PEW is in the way MT errors are handled by the linguists. The goal is to adopt a balanced approach when selecting between available MT and TM suggestions and identifying the cases where translating from scratch would be required. Every option should be weighed against not only the quality of the final product but also the amount of effort required to process or generate it in order to increase productivity and to accelerate the translation process (Guerberof, 2018).

2.3 Quality Checks - MT Evaluation – Error Annotation

A comprehensive PEW is finalized by performing QA checks for the project in order to determine whether the desired quality level has been attained, followed by MT-evaluation processes and error annotation. Some industry professionals choose to add an additional human revision step to their workflow, which makes the result equal to Translation, Editing and Proofreading (TEP), instead of comparable.

MT evaluation aims at providing quality data for the MT system used, identifying edit patterns and determining whether the engine at hand is appropriate for a specific type of content. This process involves assessing meaning preservation, fluency and PE effort based on some pre-established industry methods, known as automated evaluation models. These models provide highly valuable information to LSPs and MT researchers alike, allowing them to monitor the engine's performance and improve their systems over time, but also constitute a reliable point of reference for linguists themselves, in order to set up appropriate pricing models.

These human-executed processes constitute the final stage of the optimal PE workflow, known as annotation. Annotation consists in the classification and analysis of errors identified by a linguist in the machine translated text in order to not only provide quality data for the specific task and additional information on the MT system's functionality in general but also to keep track of the changes and corrections applied. This usually involves tagging errors according to an industry-standard typology and providing information related to the nature of the edits applied by post-editors depending on the type and frequency of the errors.

There are many ways to collect feedback from the post-editors about any given system. It can be gathered using exhaustive reporting systems and tools[1]. However, feedback may also be collected in a plain Excel file, an email or even a call. What is important is that we allow the post-editors who have worked using the system to give their own opinion about the errors found and use this to improve it (Guerrero, 2018).

3 The Training Challenges of the PEW

We may envision the translators at the very center of the post-editing process, using the computer and deciding how to best combine the materials they have on hand at each part of the process, either glossaries, translation memories or machine translation engines (Rico, 2017). The Training Challenges of the PEW, including Pre-Editing, Post-Editing and Error Annotation, are examined in relation to three Groups: Post-Editors, LSPs and Universities based on a digital survey among language industry professionals (Gene, 2020).

[1] DQF4, which allow for severity and error classification of the output, measurement of MT productivity, ranking of MT engines, evaluation of adequacy and fluency (Sanchez-Martinez, 2018). Recently, a new metric for error typology has been developed based on the harmonization of the MQM and DQF frameworks, and available through an open TAUS DQF API. This harmonization allows errors to be classified, firstly, according to a broader DQF error typology and, subsequently, by the subcategories as defined in MQM.

3.1 The Training Challenges of the PEW for the LSPs

For LSPs, it seems that all of the challenges mentioned in Fig. 3 are equally important, which reflects the work that needs to be done and the investment in time, effort, communication, research and training on the LSPs' end.

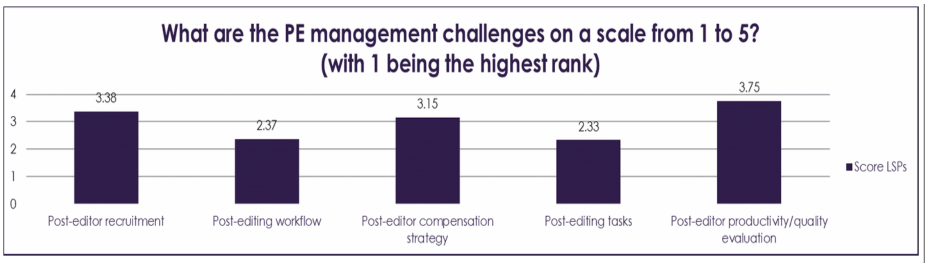
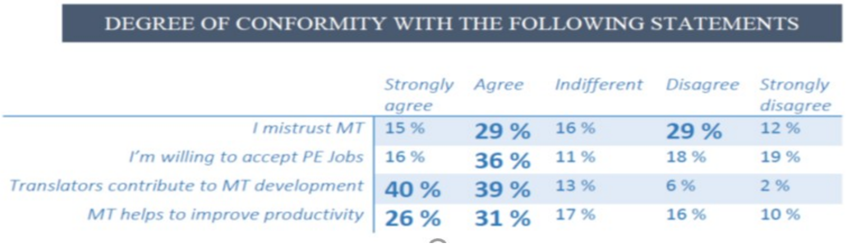


Fig. 1. What are the PE management challenges on a scale from 1 to 5? (Gene, 2020)

Training challenges affect all departments of LSPs, from the recruiters and vendor managers who draft the job descriptions and post-editors' profile(s) to the production manager who integrates the post-editing workflow, measures the productivity and applies the post-editor compensation strategy, and the quality manager who is responsible for setting a quality evaluation procedure to measure the quality results.

In a survey conducted as part of the EAMT 2018 21st Annual Conference (Pérez-Macías, Rico and Forcada, 2018), 52% of the translators were willing to accept post-editing jobs and 79% considered that translators contribute to MT development:



DEGREE OF CONFORMITY WITH THE FOLLOWING STATEMENTS

	Strongly agree	Agree	Indifferent	Disagree	Strongly disagree
I mistrust MT	15 %	29 %	16 %	29 %	12 %
I'm willing to accept PE Jobs	16 %	36 %	11 %	18 %	19 %
Translators contribute to MT development	40 %	39 %	13 %	6 %	2 %
MT helps to improve productivity	26 %	31 %	17 %	16 %	10 %

Fig. 1. Degree of conformity with several statements about MT

However, it is particularly worrying that, when answering the question 'How often the translators' needs about MT are heard', a total of 40% answered 'never' or 'almost never'. As part of this question the translators added their suggestions on ways to contribute to the MT process:

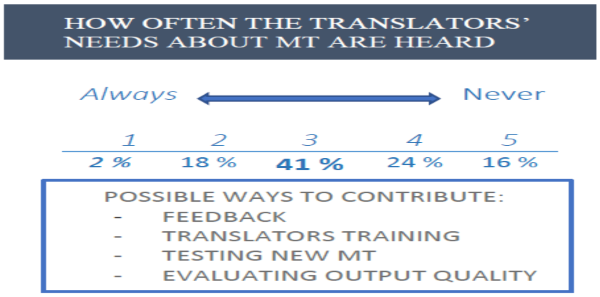

HOW OFTEN THE TRANSLATORS' NEEDS ABOUT MT ARE HEARD
Always
Never
1
2
3
4
5
2 %
18 %
41 %
24 %
16 %
POSSIBLE WAYS TO CONTRIBUTE:
- FEEDBACK
- TRANSLATORS TRAINING
- TESTING NEW MT
- EVALUATING OUTPUT QUALITY

Fig. 1. How often the translators' needs about MT are heard and possible ways to contribute

Based on the above, one of the main challenges for LSPs is mutual collaboration. According to Guerrero (2018) if all parties involved in machine translation processes acknowledge that mutual collaboration is not only possible but also desirable, then the challenge for LSPs and machine translation buyers is to take up the torch from academic research and establish new relationships with post-editors, moving towards a more translator-centered process.

3.2 The Training Challenges of the PEW for the Post-Editors

Even if productivity increases while quality is maintained, actual experience shows that PE is a tiring task for translators (Guerberof, 2018), which pushes boundaries and has no clear boundaries as a service.

In the context of Gene's survey (2020), SLVs responses (see Fig. 4) validate the feedback of LSPs (see Fig. 3) prioritizing the lack of training, the lack of post-editing skills with most important the ambiguity of post-editing tasks guidelines, which highlights the importance of mutual collaboration between LSPs and SLVs in terms of training.

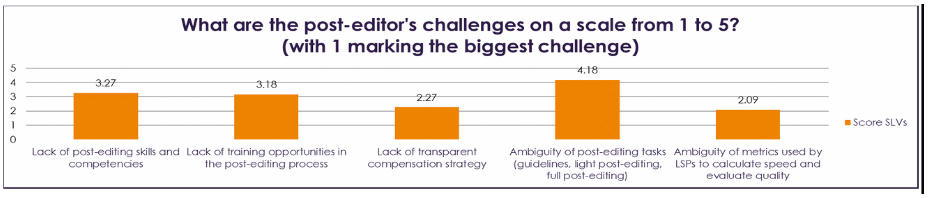


Fig. 1. What are the post-editor's challenges on a scale from 1 to 5? (Gene, 2020)

3.3 The Training Challenges of the PEW for Academia

Machine translation and post-editing are little by little finding their place as independent subjects in translation graduate and post-graduate programmes (Guerrero, 2018)[2].

[2] Regarding MT and PE training for translators, the skills needed have been described by O'Brien (2002), Rico and Torrejón (2012) and Pym (2013), while syllabi have been de-

Based on Gene’s survey (2020), LSPs and Universities seem to be isolated with no strong connection link between them. In their majority, SLVs find that the syllabus offered by Universities is not adapted to the translation industry needs, while 30% of them believe that it does meet the industry needs.

The main challenge for Academia right now is the lack of trainers. As MTPE-related courses have only recently been added to the curricula of some Universities, time will be needed for new students to evolve into future MTPE Trainers who will inspire the translation industry and bridge the training gap between LSPs and Post-Editors, balancing the demand and the offer respectively and meeting the quality standards needed.

4 Conclusions

As we have attempted to highlight throughout this paper, Post-Editing is actually but one step in the optimal workflow. Industry experience shows that, in practice, Language Service Providers rarely stick to this 3-step comprehensive workflow, with pre-editing and annotation steps often skipped, and consequently fail to profit from a real return on investment. However, unless pre-editing and annotation procedures are applied we could argue that the whole process is at best ineffective or even futile; without these a Post-Editing project is merely a one-off that simply saves the linguist and client some time through the use of MT technology, but bears no real value for the company supporting it. For this reason, LSPs should revise their standard practices and establish a workflow that is both profitable and sustainable in the long term by making the most of available technologies.

To sum up, on the one hand, LSPs should bear in mind that Post-Editing is best suited to larger volumes (over 100k TUs). On the other hand, they should not forget that human processes such as pre-editing and annotation are the keys to achieving a better return on investment in the long term. Workflows limited to MT and post-editing processes do not constitute an effective strategy with long-term profitable results, but only temporary discounts. In order to enjoy the real benefits of integrating post-editing to their list of offered services, an investment should be made in the complete, optimal post-editing workflow.

Regarding the training challenges arising as a result of the PEW, the suggestion is to abandon the current machine-centered paradigm and work with all stakeholders towards a translator-centered process, in which the post-editor is transformed in a translation expert applying critical thinking, problem-solving and decision-making skills not only being involved to correct the errors stubbornly produced by the MT system

signed, and courses explained and described, by Doherty et al. (2012), Doherty and Moorkens (2013), Doherty and Kenny (2014), Koponen (2015) and Mellinger (2017). The training suggestions made by Guerberof (2018) include teaching basic MT technology concepts, MT evaluation techniques, Statistical MT (SMT) training, pre-editing and controlled language, monolingual PE, understanding various levels of PE (light and full), creating guidelines, MT evaluation, output error identification, when to discard unusable segments and continuous PE practice.

but having a vital role in each and every step of the workflow (Guerrero, 2018). The derived data on training challenges among the stakeholder will serve for further analysis and speculations for future problem-solving and more specific training pathways. The closer interaction between the Academia and the LSPs during the traineeships with a common training model agreed between them would be interesting to be examined along with the evaluations of LSPs and the queries of Post-Editors, which could feed the PE teaching models of the Universities.

Bridging the distance between the different stakeholders (LSPs, Post-Editors and Academia) through training is the key to the enlargement of the translation industry. Translators should be encouraged to embrace MT processes and companies to integrate post-editors into their processes, investing in the linguist-focused and not the machine-centered processes for higher quality and productivity.

References

1. Aranberri, N., G. Labaka, A. Diaz de Ilarraza and K. Sarasola (2014), ‘Comparison of Post-editing Productivity Between Professional Translators and Lay Users’, in S. O’Brien, M. Simard and L. Specia (eds), *Proceedings of the 3rd workshop on post-editing technology and practice*, 20–33, Vancouver: AMTA.
2. Arthern, P. J. (1979), ‘Machine Translation and Computerized Terminology Aystems: A Translator's Viewpoint’, B. M. Snell (ed.), *Translating and the Computer: Proceedings of a seminar,* 77–108, London: North-Holland.
3. Aziz, W., S. Castilho and L. Specia (2012), ‘PET: A Tool for Post-editing and Assessing Machine Translation’, in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC’12)*, 3982–3987, Istanbul: European Language Resources Association (ELRA).
4. Bartolomé Mesa-Lao, bm.ibc@cbs.dk, Introduction to post-editing, Center for Research and Innovation in Translation and Translation Technology, Copenhagen Business School, Denmark
5. Zainurrahman Sehan, STKIP Kie Raha Ternate, Five Translation Competencies
6. Celia Rico, Senior Lecturer, Facultad de Artes y Comunicación, Universidad Europea de Madridcelia.rico@uem.es, Enrique Torrejón, Language Technologies Consultantenrique.-torrejon@gmail.com, Skills and Profile of the New Role of the Translator as MT Post-editor
7. Allen, J. (2003). Post-Editing. In H. Somers (Ed.), *Computers and Translation: a Translator's Guide* (pp. 297-317). Amsterdam: John Benjamins.
8. Aranberri, N. (2017). What Do Professional Translators Do When Post-Editing for the First Time? First Insight Into the Spanish-Basque Lanuage Pair. *HERMES - Journal of Language and Communication in Business*(56), 89-110.
9. Bentivogli, L., Bisazza, A., Cettolo, M., & Federico, M. (2016). Neural Versus Phrase-Based Machine Translation Quality: a Case Study. *Conference on Empirical Methods in Natural Language Processing*, (pp. 257-267). Austin, Texas.
10. Blagodarna, O. (2018). ‘Enhancement of Post-Editing Performance: Introducing Machine Translation Post-Editing in Translator Training’, Universidad Autonoma de Barcelona.
11. Blanchon, H., & Besacier, L. (2017, September 21). *Comparing Statistical Machine Translation (SMT) and Neural Machine Translation (NMT) Performance.* Retrieved from http://lig-membres.imag.fr/blanchon/SitesEns/NLSP/resources/SMT-vs-NMT.pdf

12. Daems, J., Vandepitte, S., Hartsuiker, R. J., & Macken, L. (2017). Identifying the Machine Translation Error Types with the Greatest Impact on Post-editing Effort. *Frontiers in Psychology, 8, 1282*. doi:https://doi.org/10.3389/fpsyg.2017.01282
13. de Almeida, G. (2013). Translating the post-editor: an investigation of post-editing changes and correlations with professional experience across two Romance languages. Dublin: Dublin City University.
14. Dino, G. (2017, December 18). *3 Reasons Why Neural Machine Translation is a Breakthrough.* Retrieved from Slator: Language Industry Intelligence:
15. Gene, V. (2020, May 21). *The Management and Training Challenges of Post-Editing (Part 2).* Retrieved from GALA: Globalization and Localization Association: https://www.gala-global.org/ondemand/management-and-training-challenges-post-editing-part-2
16. Gene, V. (2020, May 7). *The Management and Training Challenges of Post-Editing (Part 1).* Retrieved from GALA: Globalization and Localization Association: https://www.gala-global.org/ondemand/management-and-training-challenges-post-editing-part-2
17. Gene, V. (2020, November). The Role and Perspective of the Post-Editor: What Are the Training Challenges?, *Translating and the Computer - TC42.* London.
18. Ginovart Cid, C., Colominas, C., & Oliver, A. (2020). Language Industry Views on the Profile of the Post-editor. *Translation Spaces*. doi:https://doi.org/10.1075/ts.19010.cid
19. Guerberof Arenas, A., & Moorkens, J. (2019, January). Machine Translation and Post-editing Training as Part of a Master's Programme. *The Journal of Specialised Translation*(31), 217-238.
20. Guerrero, L. (2018). From a Discreet Role to a Co-Star: The Post-Editor Profile Becomes Key to the Post-Editing Workflow. *Translating and the Computer - TC40.* London.
21. Hu, K., & Cadwell, P. (2016). A Comparative Study of Post-Editing Guidelines. *Baltic Journal of Modern Computing, 4*(2), 346-353.
22. Koponen, M. (2015), 'How to Teach Machine Translation Post-editing? Experiences from a Post-editing Course', in S. O'Brien, M. Simard and L. Specia (eds), Proceedings of the 4th Workshop on Post-editing Technology and Practice (WPTP 4), 2–15, Miami: AMTA.
23. Koponen, M., W. Aziz, L. Ramos and L. Specia (2012), 'Post-editing Time as a Measure of Cognitive Effort', in S. O'Brien, M. Simard and L. Specia (eds), Proceedings AMTA 2012 Workshop on Post-editing Technology and Practice, 11–20, San Diego: AMTA.
24. Krings, H. P. (2001). Repairing texts: Empirical investigations of machine translation post-editing processes. Kent, Ohio: Kent State University Press.
25. Lommel, A. R. (2017), 'Neural MT: Sorting Fact from Fiction', Common Sense Advisory. Available online: http://www.commonsenseadvisory.com/AbstractView/tabid/74/ArticleID/37893/Title/NeuralMTSortingFactfromFiction/Default.aspx (accessed 21 September 2018).
26. Lommel, A. R. and D. A. De Palma (2016a), 'Post-editing Goes Mainstream', Common Sense Advisory. Available online: (accessed 21 September 2018).
27. Lommel, A. R. and D. A. De Palma (2016b), 'TechStack: Machine Translation', Common Sense Advisory. Available online: (accessed 21 September 2018).
28. Mellinger, C. D. (2017), 'Translators and Machine translation: Knowledge and Skills Gaps in Translator Pedagogy', The Interpreter and Translator Trainer, 1–14, Taylor & Francis Online.
29. Moorkens, J. and S. O'Brien (2013), '*User Attitudes to the Post-editing Interface*', in S. O'Brien, M. Simard and L. Specia (eds), Proceedings of 2nd Workshop on Post-editing Technology and Practice, 19–25, Nice: EAMT.
30. Moorkens, J. and S. O'Brien (2015), '*Post-editing Evaluations: Trade-offs Between Novice and Professional Participants*', in Proceedings of EAMT, 75–81, Antalya: EAMT.

31. Nitzke, Jean. 2019. Problem solving activities in post-editing and translation from scratch: A multi-method study (Translation and Multilingual Natural Language Processing 12). Berlin: Language Science Press.
32. Nunes Vieira, L., Alonso, E., & Bywood, L. (2019). Introduction: Post-editing in practice – Process, product and networks. *The Journal of Specialised Translation*, 2-13.
33. O'Brien, S., (2002). *Teaching Post-editing: A Proposal for Course Content,* 6th EAMT Workshop Teaching Machine Translation.
34. O'Brien, S., Roturier, J., & de Almeida, G. (2009). *Post-Editing MT Output; Views from the researcher, trainer, publisher and practitioner*. Retrieved from http://www.mt-archive.info/MTS-2009-OBrien-ppt.pdf
35. O'Brien, S., & Moorkens, J. (2014). Towards Intelligent Post-Editing Interfaces. In W. Baur, B. Eichner, S. Kalina, N. Kessler, F. Mayer, & J. Orsted (Ed.), *20th FIT World Congress*, (pp. 131-137). Berlin.
36. O'Brien, S., Winther Balling, L., Carl, M., Simard, M., & Specia, L. (Eds.). (2014). *Post-editing of Machine Translation: Processes and Applications*. Newcastle: Cambridge Scholars Publishing
37. O'Brien, S. and S. Castilho (2016), 'Evaluating the Impact of Light Post-editing on Usability', in Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 310–6, Portorož: European Language Resources Association (ELRA).
38. O'Brien, S. (2017), 'Machine Translation and Cognition', The Handbook of Translation and Cognition, J. W. Schwieter and A. Ferreira (eds), 311–31, Wiley-Blackwell.
39. Pym, A. (2013), 'Translation Skill-sets in a Machine-translation Age', Meta, 58 (3), 487–503.
40. Rico, C., & Torrejon, E. (2012). Skills and Profile of the New Role of the Translator as MT Post-Editor. *Tradumàtica: tecnologies de la traducció*, 166-178.
41. Saillard, J.-L. (2019, January 28). *Transitioning to a post-editing machine translation business model.* Retrieved from Smartcat: https://www.smartcat.com/blog/transitioning-to-a-post-editing-machine-translation-business-model/
42. Sanchez-Martinez, F. (2012). "Motivos del creciente uso de la traduccion automatica seguida de posedicion". Tradumatica, 10, 150-156 [Download on 8 March 2015].
43. TAUS. (2011a). Adequacy/Fluency Evaluation [Download on 4 March 2015].
44. TAUS. (2011b). Adequacy/Fluency Process Guide [Download on 4 March 2015].
45. TAUS. (2011c). Dynamic Quality Framework [Download on 4 March 2015].
46. TAUS. (2011d). Error Typology [Download on 4 March 2015].
47. TAUS. (2014, January 27). *Evaluating Post-Editor Performance Guidelines.* Retrieved from TAUS: the language data network: https://www.taus.net/academy/best-practices/postedit-best-practices/evaluating-post-editor-performance-guidelines
48. Tyler, S., Hertel, P., & McCallum, M. H. (1979). Cognitive Effort and Memory. *Journal of Experimental Psychology: Human Learning and Memory, 5*(6), 607-617.
49. Vashee, K., & Wiggins, D. (2014). *A 6 Step Plan to Create an Optimal MT Post-Editing Strategy.* Retrieved from https://omniscien.com/wp-content/uploads/2016/02-new/Asia-Online-Webinar-A-6-Step-Plan-to-Create-an-Optimal-MT-Post-Editing-Strategy2.pdf
50. Vasconcellos, M. and M. León (1985): "SPANAM and ENGSPAN: Machines Translation at the Pan American Health Organization", Computational Linguistics 11, 122-136. [online] Available. http://acl.ldc.upenn.edu/J/J85/J85-2003.pdf [accessed 13 September 2012]
51. Zainurrahman, S. (2010). *Five Translation Competencies.* Retrieved from Zainurrahman's Personal Journal: https://zainurrahmans.wordpress.com/2010/06/06/five-translation-competencies/

52. Zaretskaya, A. (2019). Optimising the Machine Translation Post-editing Workflow. In I. Temnikova, C. Orasan, G. Corpas Pastor, & R. Mitkov (Ed.), *2nd Workshop on Human-Informed Translation and Interpreting Technology*, (pp. 136-139). Varna.
53. Zaretskaya, A., Vela, M., Corpas Pastor, G., & Seghiri, M. (2016). Measuring Post-editing Time and Effort for Different Types of Machine Translation Errors. *New Voices in Translation Studies, 15*, 63-91.
54. Καλαντζή, Δ. (2016). Αυτόματη Μετάφραση και Post-Editing: Εισαγωγικά Στοιχεία και Σκέψεις. *5th Meeting of Greek Translatologists.*

Benchmarking ASR Systems Based on Post-Editing Effort and Error Analysis

Martha Maria Papadopoulou[1], Anna Zaretskaya[2], and Ruslan Mitkov[1]

[1] University of Wolverhampton, UK {m.m.papadopoulou,R.Mitkov}@wlv.ac.uk
[2] TransPerfect azaretskaya@translations.com

Abstract. This paper presents a comparative evaluation of four commercial ASR systems which are evaluated according to the post-editing effort required to reach "publishable" quality and according to the number of errors they produce. For the error annotation task, an original error typology for transcription errors is proposed. This study also seeks to examine whether there is a difference in the performance of these systems between native and non-native English speakers. The experimental results suggest that among the four systems, Trint and Microsoft obtain the best scores. It is also observed that most systems perform noticeably better with native speakers and that all systems are most prone to fluency errors.

Keywords: Automatic Speech Recognition· speech-to-text· post-editing · error annotation.

1 Introduction

The rapid technological progress in the field of Automatic Speech Recognition (ASR) has lead to claims that speech-to-text systems can achieve up to 90% accuracy [9,15]. The aim of this paper is to shed some light on the impact that this progress has on the productivity of end users. Until now, the evaluation of ASR systems relied exclusively on Word Error Rate (WER) and similar metrics. Calculating these metrics is usually an expensive and time-consuming task as manual transcriptions are used for reference. In addition, these traditional approaches do not provide information on the cognitive effort required to reach "publishable" quality. In this paper, the aim is to address the aforementioned issues by proposing a way to depart from the traditional methods of ASR evaluation. The key idea is to deploy the post-editing (PE) method in the evaluation process. To bridge the gap of the underrepresented aspect of cognitive effort, four ASR systems: Amazon[3], Microsoft[4], Trint[5] and Otter[6] were evaluated based on

[3] https://aws.amazon.com/transcribe/

[4] https://azure.microsoft.com/en-us/services/cognitive-services/speech-to-text/

[5] https://trint.com/

[6] https://otter.ai/

Translation and Interpreting Technology Online, pages 199–207,
July 05-07, 2021.
https://doi.org/10.26615/978-954-452-071-7_023

post-editing effort. To this end, the PET tool [1] was employed to compute the post-editing (PE) effort in terms of PE time and PE distance at a sentence level. The objective of the PE process was to rank all systems based on their overall score.

In the attempt to provide a qualitative analysis, a secondary objective seeks to investigate the types of errors that these systems produce. To accomplish this, a new error typology for transcription errors was developed following the TAUS DQF-MQM [17] main error categories. In this novel typology, which is essential for this study, the subcategories are tailored to suit transcription errors. To the best of our knowledge, this is the first study that seeks to investigate which types of errors the ASR systems are most prone to. The comparison between the error annotation results and the post-editing results will lead to new insights of their correlation.

Another goal of this study is to examine the role of the speaker's accent. It investigates whether the performance of the systems is affected by the speaker's accent. To address this question, results from native and non-native English speakers are compared.

The rest of the paper is structured as follows: Section 2 contextualises the current study by discussing related work. Section 3 outlines the data used, Section 4 presents the experimental setup. Section 5 discusses the results of the experiments conducted. Finally, Section 6 provides the conclusions of this study.

2 Related Work

The rapid development of state-of-the-art Automatic Speech Recognition systems led to the need for these systems to be evaluated. A recent study [2] benchmarked commercial ASR systems by comparing their results against human quality and evaluating them using the WER metric. This research pays specific attention to named entity recognition. Related research includes [6], where a tool was designed to perform comparisons between commercial and open-source ASR systems using the WER metric. A recent systematic review [14] discusses the problems of benchmarking ASR systems, which were presented in various studies and expresses skepticism for the very low WER results reported. They demonstrated that the WER rate was considerably higher than the best results reported in those studies. A further study [4] also benchmarked three commercial ASR systems, but they reported results using three metrics: WER, Hper and Rper. A qualitative analysis [10] on ASR systems was performed aiming to evaluate the accuracy of the Language Model adaptation; in order to do so, the WER metric was applied only to relevant words.

It is worth noting that none of the existing approaches appears to overcome the limitations of the WER metric. There is therefore a need for a new evaluation approach. Two new performance metrics: MER (match error rate) and WIL (word information lost) were proposed in [12]. Furthermore, with the aim to represent human perception of ASR accuracy, HPA (Human Perceived Accuracy) was developed [11]. Another metric was introduced in [3] seeking to achieve a better correlation to human evaluation. Finally, an extension of the WER metric was proposed in [8], where weighted penalties were applied by implementing word embeddings.

To the best of our knowledge, studies that employ post-editing in order to evaluate ASR performance are scarce. Post-editing was explored in [7], where users browsed and corrected automatic transcriptions of lectures in a web-based interactive interface. This study aimed to compare WER rates with comprehensibility improvements after transcripts were post-edited. As detailed in [16], an ASR system was developed for Polish, which introduced the novel idea of applying automatic post-editing in the ASR output. Finally, two crowdsourcing studies were compared in [5] with the objective to investigate whether it is preferable to transcribe from scratch or to perform post-editing on ASR output. They concluded that post-editing is preferable only when WER accuracy is lower than 30%. However, effort indicators of the post-editing task were not examined in this study.

The above discussion provides compelling evidence that there is a pressing need for an alternative approach to account for the cognitive effort required to post-edit raw ASR outputs. To the best of our knowledge, this study constitutes the first analysis of the evaluation of post-editing effort in this field. The added value of this paper is also highlighted by the qualitative analysis on the transcription errors, which remains unexplored in the literature. With this aim in mind, this paper puts forward an error typology for ASR transcription errors. The suggested error typology is the first of its kind to be specifically designed for the use case where the ASR output is post-edited by humans to reach “publishable quality”.

3 Data Description

For the purpose of this study, the video data were obtained from the research seminar series “Specialised Seminar: Technologies for Translation and Interpreting: Challenges and Latest Developments”[18], hosted by Prof R Mitkov at the University of Wolverhampton. More specifically, the videos were recordings of talks given by invited speakers on topics related to Translation and Interpreting Technologies, which were held online via Zoom. Thus, all data have the same register and belong to the same domain. It should be mentioned that for Microsoft and Amazon ASR systems the video files were converted to audio files (.wav), as these systems operate exclusively on this file format. Two videos were used as input data: one with a native American English speaker and one with a non-native American English speaker. The mother tongue of the non-native English speaker is Russian. The videos were trimmed in order to have the same length—approximately 15 minutes per video. Each ASR system produced a transcription of approximately 2,000 words per video, thus the size of the post-editing and error annotation tasks for all four systems consisted of approximately 16,000 words.

4 Experimental Setup

4.1 Post-editing

The transcriptions produced by the ASR systems were exported in simple text format and tokenised into sentences in order to be imported into the post-editing tool. The tokenisation task was performed using the Punkt Sentence Tokenizer module from the NLTK Python library. The post-editing process was carried out using the PET tool [1], an open-source post-editing tool, which served a double purpose both to facilitate the post-editing task and to collect sentence-level information. Along with the post-editing process, this tool gathered information related to the post-editing effort such as editing time and number of edits per segment. These results were exported to calculate the post-editing effort.

The character-based Levenshtein distance was used in this experiment. It was calculated on the basis of the number of characters that were changed (insertions, deletions and substitutions) out of the total number of characters in the segment.

The PE task was performed by a single post-editor with intermediate experience in the field. As the desired outcome was a verbatim transcription, the post-editor was instructed to perform light post-editing. For this reason, speech disfluencies and repetitions were not corrected.

4.2 Error Annotation

For the error annotation task, the BLAST tool [13] was used, which is an open-source tool. For the purposes of this task an error typology was designed following the DQF-MQM TAUS Error Typology format, which was customised to correspond to transcription-related errors only (see Table 1). The DQF-MQM TAUS Error Typology was selected as a basis since its main error categories correspond to transcription errors and customisation was only required for the sub-categories. The error annotation task was performed by the post-editor. The results of each annotation task were automatically generated by the BLAST tool.

5 Results

As seen in Table 2, Microsoft obtained the best score for total PE time. It is also worth noting that all systems required more PE time for the non-native speaker, with the exception of Otter. It was also noted that Otter required the most PE time for the native speaker transcription. This is mainly caused by the increased average segment length of 144.32 characters compared to the rest of the systems, whose average segment length range between 78.13–97.88 characters. In particular, Otter's average segment length reached a peak of 5,082 characters in a single segment. It is worth highlighting the significant difference in PE time between the native and non-native Amazon transcriptions. This is also represented in the PE distance and will be discussed further as part of the error analysis.

Table 1. Error Typology

Accuracy	Omission	Prefix
		Suffix
		Article
		Preposition
	Addition	Prefix
		Suffix
		Article
		Preposition
	Mistranscription	Proper noun
		Number
		Single to multiple words
		Multiple to single word
		Single to single word
		Multiple to multiple words
	Homophone	
Fluency	Segmentation	
	Punctuation	Additional punctuation mark
		Missing punctuation mark
		Wrong punctuation mark
	Spacing Capitalisation Filler word	
Grammar	Grammatical number Grammatical tense	
Style	Inconsistent style Abbreviated form Spelled out form	
Terminology	Term Abbreviation	

Table 2. Total PE time(s)

	Native Speaker	Non-Native Speaker	Total
Microsoft	2,087.93	2,426.79	4,514.72
Trint	2,165.87	2,442.37	4,608.25
Amazon	1,520.41	4,855.12	6,375.53
Otter	5,550.37	3,039.74	8,590.11

In terms of average and overall PE distance, Trint produced the best score (see Table 3). The aforementioned differences between native and non-native speakers for Amazon and Otter are also reflected in the PE distance results.

Table 3. Average PE distance per segment

	Native Speaker	Non-Native Speaker	Overall Average
Trint	4.14%	7.41%	5.69%
Otter	8.95%	4.50%	6.43%
Microsoft	5.34%	8.65%	7.10%
Amazon	4.37%	15.66%	9.08%

According to the PE results, Trint performed best in terms of post-editing effort, taking into consideration both PE distance, where it scored first, and in terms of PE time, where it delivered the second best results but with small differences from the first system.

According to the error annotation results, Trint performed the best with the lowest total number of errors for both speakers (see Table 4).

Table 4. Total number of errors

	Native Speaker	Non-Native Speaker	Total
Trint	109	185	294
Microsoft	141	210	351
Otter	213	250	463
Amazon	163	464	627

As for the results related to the different error categories, a general tendency towards fluency errors was observed (see Table 5). The percentage of fluency errors ranges between 48.55%–71.12% of the total errors. The tendencies towards the second and the third most frequent error categories are also consistent through all systems, with accuracy ranging between 21.57%–37.08% in second place, and terminology ranging between 4.20%–7.86% in third place.

Table 5. Percent of errors per error category

	Fluency	Accuracy	Terminology	Grammar	Style
Trint	48.55%	37.08%	7.86%	2.72%	3.80%
Microsoft	56.20%	32.04%	5.95%	3.80%	2.01%
Otter	71.12%	21.57%	5.20%	1.03%	1.07%
Amazon	60.81%	32.48%	4.20%	2.19%	0.21%

As seen in Table 6, the systems are ranked in terms of PE time, PE distance and number of errors. It is evident that the number of errors does not always correlate with the PE effort. The results support the conclusion that systems with lower number of errors do not necessarily have the best score in terms of PE time and PE distance.

Table 6. Ranking systems based on PE time, PE distance and number of errors

PE time	PE distance	Number of errors
Microsoft	Trint	Trint
Trint	Otter	Microsoft
Amazon	Microsoft	Otter
Otter	Amazon	Amazon

A closer look at the error annotation results suggests further observations regarding the correlation of PE time and error categories. Firstly, there is a strong correlation between fluency errors and PE time: the higher the rate of fluency errors the more PE time is required. For example, Otter has the highest fluency rate and is the system that required almost twice as much PE time as the systems that ranked first and second (see Table 3). The most frequent fluency errors in this case were punctuation and segmentation. These two categories also justify the longer segment rate for Otter and the correlation with the increased PE time.

Secondly, a weak correlation between accuracy errors and PE time is noted. A high rate of accuracy errors, contrary to the popular belief, does not require extra PE time. For instance, Trint reported the highest accuracy rate; however, it was ranked second based on PE time. In this case, the low correlation could be justified by the high number of omission and addition errors, which are easily detectable and require less cognitive effort, combined with the low number of mistranscription errors, which require more cognitive effort.

Finally, it should be highlighted that there is a big performance difference in PE time between native and non-native speakers for Amazon. This difference can be explained by the high number of filler word, mistranscription, segmentation and terminology errors of the non-native speaker transcription.

6 Conclusions

In this study, outputs from commercial ASR systems were post-edited and then the errors were annotated. The ASR systems were ranked based on the post-editing effort required to reach “publishable” quality and the number of errors they produced. In accordance with the PE and error results presented above, it can be concluded that with the data used in this experiment, Trint is the best performing system in terms of PE distance and total number of errors, while Microsoft is the best performing system in terms of PE time. Moreover, the number of errors does not always correlate with the PE effort. It is also evident that there is a general tendency towards fluency errors, which are assumed to be the most time-consuming errors. The experiments point to the conclusion that most ASR systems perform better with a native speaker.

The constraints of this study include its limited scope and the involvement of only one post-editor and annotator; larger-scale study results may be different. While the size of the data was another constraint, the results reported remain insightful. In particular, this study will pave the way for further research in the field of ASR evaluation, post-editing and error analysis. Future work could explore the correlation between the suggested approach and the traditional WER metric.

7 Acknowledgements

I would like to thank TransPerfect for offering me a placement opportunity to conduct this research. In particular, I am very grateful to my placement supervisor Anna Zaretskaya for her guidance and support. I would also like to thank my supervisor Prof R Mitkov for his constant support and encouragement. Some special thanks go to my classmate Rea Bartlett Tandon for proofreading this paper.

References

1. Aziz W, Castilho S, Specia L.: PET: a Tool for Post-editing and Assessing Machine Translation. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC), pp. 3982-3987. European Language Resources Association (ELRA) (2012)
2. Del Rio M, Delworth N, Westerman R, Huang M, Bhandari N, Palakapilly J, McNamara Q, Dong J, Zelasko P, Jette M.: Earnings-21: A Practical Benchmark for ASR in the Wild. arXiv preprint arXiv:2104.11348 (2021).
3. Favre B, Cheung K, Kazemian S, Lee A, Liu Y, Munteanu C, Nenkova A, Ochei D, Penn G, Tratz S.: Automatic human utility evaluation of ASR systems: Does WER really predict performance? In: INTERSPEECH-2013, pp 3463–3467 (2013).
4. Filippidou, F., Moussiades, L.: A Benchmarking of IBM, Google and Wit Automatic Speech Recognition Systems. In: International Conference on Artificial Intelligence Applications and Innovations IFIP, pp.73-82. Springer International Publishing (2020).
5. GaurY, Lasecki WS, Metze F, Bigham JP.: The effects of automatic speech recognition quality on human transcription latency. In: 13th Web for All Conference, pp.1–8. Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2899475.2899478
6. Këpuska, V., Bohouta, G.: Comparing Speech Recognition Systems (Microsoft API, Google API And CMU Sphinx). Int. Journal of Engineering Research and Application 7(3), 20–24 (2017).
7. KolkhorstH, Kilgour K, Stüker S, Waibel A.: Evaluation of interactive user corrections for lecture transcription. In: International Workshop on Spoken Language Translation (IWSLT) 2012, pp. 1-8 (2012).

8. Le N-T, Servan C, Lecouteux B, Besacier L.: Better evaluation of ASR in speech translation context using word embeddings. In: Interspeech 2016, pp. 1-6 (2016).
9. LevitM, Chang S, Buntschuh B, Kibre N.: End-to-end speech recognition accuracy metric for voice-search tasks. In: International Conference on Acoustics (IEEE), Speech and Signal Processing (ICASSP), pp. 5141-5144 (2012). https://doi.org/10.1109/ICASSP.2012.6289078
10. MdhaffarS, Estève Y, Hernandez N, Laurent A, Dufour R, Quiniou S: Qualitative Evaluation of ASR Adaptation in a Lecture Context: Application to the PASTEL Corpus. In: INTERSPEECH-2019, pp.569–573 (2019).
11. MishraT, Ljolje A, Gilbert M.: Predicting human perceived accuracy of ASR systems. In: 12th Annual Conference of the International Speech Communication Association, pp.1945-1948 (2011).
12. MorrisAC, Maier V, Green P.: From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. In: 8th International Conference on Spoken Language Processing, pp. 2765-2768 (2004).
13. Stymne, S.: Blast: A tool for error analysis of machine translation output. In: Proceedings of the ACL-HLT 2011 System Demonstrations, pp. 56-61 (2011).
14. SzymańskiP, Żelasko P, Morzy M, Szymczak A, Żyła-Hoppe M, Banaszczak J, Augustyniak L, Mizgajski J, Carmiel Y: WER we are and WER we think we are. arXiv preprint arXiv:2010.03432 (2020).
15. WilliamsJD, Melamed ID, Alonso T, Hollister B, Wilpon J.: Crowdsourcing for difficult transcription of speech. In: IEEE Workshop on Automatic Speech Recognition & Understanding, pp.535-540 (2011). https://doi.org//10.1109/ASRU.2011.6163988.
16. Wnuk D, Wołk K.: Post-editing and Rescoring of Automatic Speech Recognition Results with OpenNMT-APE, In: Proceedings of the PolEval 2020 Workshop, pp. 33-37 (2020).
17. Harmonized DQF-MQM Error Typology, `https://www.taus.net/qt21-project#harmonized-error-typology`. Last accessed 22 May 2021
18. Technologies for Translation and Interpreting: Challenges and Latest Developments 2020/21, `https://em-tti.eu/em-tti-seminar-series/`. Last accessed 22 May 2021

Feedback in Online Translation Courses and the Covid Era

Miguel A. Jiménez-Crespo[1][0000-0002-4938-3095]

[1] Rutgers University NJ 08901 USA
[2] 15th Seminary Pl., 5th Floor, New Brunswick, NJ, 08901, USA
jimenez.miguel@rutgers.edu

Abstract. The Covid pandemic upended translation teaching globally. The forced move to online teaching represented a gargantuan challenge for anyone only experienced in face-to-face teaching. Online translation teaching requires distinct approaches to guarantee that students can reach the targeted learning goals. This paper presents a literature review on the provision of effective feedback in the light of these drastic changes in translation teaching as well as a description as how existing research on online feedback for translation training has been applied to the design of online courses at the translation program at Rutgers University.

Keywords: Translation training, feedback, online feedback

1 Introduction

The Covid pandemic upended translation teaching globally and the forced move to online teaching represented a gargantuan challenge for anyone only experienced in face-to-face teaching [8]. Online translation teaching requires distinct approaches to guarantee that students can reach the targeted learning goals [4, 10]. This paper presents a literature review on the provision of effective feedback in the light of these drastic changes in translation teaching. This is, without any doubt, one of the most important issues in online courses to successfully engage learners and to improve their translation skills. This is supported by research both in face-to-face and online courses, because "without feedback, adult learners will experience anxiety, frustration, and often failure, and so will their teachers" [7: 15]. Feedback can become an extremely time-consuming activity, and more so for those who had to quickly adapt to online environments. Providing feedback in an efficient manner maximizes the intended effect concerns anyone engaged in online education. This article provides a brief overview of existing research on the topic of online translation teaching and the role of feedback with the goal of providing applied recommendations. In addition to the literature review, it presents examples of how the certificate and Masters program at Rutgers University has implemented these feedback practices into their courses.

Translation and Interpreting Technology Online, pages 208–215,
July 05-07, 2021.
https://doi.org/10.26615/978-954-452-071-7_024

2 First things first: feedback provision based on translation competence models

Feedback can only be built upon a solid framework that includes effective online teaching methodologies and previously established models of what is precisely being taught. Two basic notions here are "translation competence" and "socioconstructivist" teaching methodologies. "Translation competence" refers to the skillset required to translate at a professional level not possessed by all bilinguals [14, 15]. Existing models of "translation competence", such as those by the research group PACTE of the European Masters Association (EMT), represent a research-based framework to establish specific learning goals for each course or program. These competence frameworks are componential models, meaning that they consist of a number of subcompetences. In the case of the PACTE model [15], these subcompetence are: (1) linguistic (language, specialized language, drafting genres such as contracts or brochures), (2) extralinguistic (knowledge about specialized domains or areas of knowledge), (3) knowledge about translation (processes, ethics, strategies), (4) instrumental (TM tools, MT, online documentary sources, data mining strategies) as well as the main one, (5) the strategic subcompetence (the ability to mobilize all the other components to solve quickly any specific translation problem) [14]. These main competences in the models contain non-finite lists of sub skills for each component that research has shown professional translators possess, and consequently, assume to be those that translation students need to acquire. These models not only provide an overall framework to structure what type of translations are presented, but also to plan specific task-based activities, pre -or post- translation, that can help build translation competence [1, 5]. This is a topic that cannot be fully expanded here due to space constraints, but in Jiménez-Crespo [3] readers can get more information on how to build translation programs and courses up using these models of translation competence. They are a key foundation to provide a framework with specific learning goals, to scaffold learning activities, plan projects, testing and evaluations, etc. They are also key to direct any type of feedback towards the achievement and evaluation of those goals.

The second basic area of interest is online teaching methodology. Translation is generally not well-suited for what is known as "transmissionist" approaches [5]. These are the "classic" approaches in which the instructor lectures from a stand or a videoconference, students translate a model and then, in turn, receive the authoritative corrections from the instructor. Students are then assumed to "somehow" integrate this learning into their active competencies. Research has shown that translation is a "performance-based" skill and therefore, teaching cannot primarily be done by lecturing about "how" to do it, meaning teaching from the podium (or videoconference/ pre-recorded instructional videos) general principles or agreeing / disagreeing with students that their proposed translations are right or wrong. This is not fully conductive for students to actively integrate the learning contents in their future translation performance. Lecturing is definitely necessary for some parts of the learning process, but translation is about 20% "declarative knowledge", that is, knowledge "about" the process, and 80% "operational knowledge", that is, "how to do something" [14].

Currently, research in the didactics of translation shows that the most popular methodologies, especially for online contexts, are the “socioconstructivist” approaches [6], as well as “situated-learning” ones [2]. These perspectives indicate that translation learning is fundamentally an interactive, collaborative, “socio-personal process”. Learners are at the center of their learning process and they socially construct their knowledge. This means that participants discover knowledge by themselves through collaboration in real-life professional translation assignments or specific tasks in the overall cycle of translation production. In this regards, online learning is an ideal context to replicate real-life professional translation assignments. Instructors are seen as facilitators or guides rather than authoritative figures that have the final say in translation solutions, and they are in charge of creating real-life simulations such as the projects by Olvera-Lobo et al [12, 13]. In some of these online courses, translation assignments resemble freelance team jobs in which students can rotate in their roles as managers, terminologists, translators, and revisers, while trainers play the role of the client, as well as the guide that provides informative, effective feedback to point students towards possible strategies, resources and mechanisms to identify solutions to translation problems.

Last but not least, the ultimate goal of translation education is to produce “experts” in a specific skill, and according to cognitive science, feedback plays a fundamental role. To achieve expertise in any field, structured “deliberate practice” that results from “regular engagement in specific activities directed at performance enhancement in a particular domain”, with “appropriate difficulty and informative feedback” [16: 29], is in fact the most efficient way. Any practice, in this context, requires a constant dedication and processing translations at the “growing edge”. This means that activities have to be of increasing difficulty, but it requires participants to make an effort to complete them. The here then is how to organize challenging and engaging translation courses in which students receive this “informative feedback” that will make them grow.

3 Types of translation feedback online

Feedback in online courses will first and foremost depend on the nature or type of online instruction. Generally, online training can be synchronous or asynchronous, or a combination of both. Synchronous training involves training through videoconferencing systems in which all students meet at the same time, while asynchronous courses are organized as self-paced courses in which students do not meet at a specific time with other classmates or the trainer, but they do have specific deadlines to complete projects, assignments or quizzes, participate in forums or videoforums, watch prerecorded lectures, etc. In addition, translation training can be “process-oriented” and “product-oriented”. In “process-oriented” training, instructors focus on translation strategies rather than the target text: basic concepts and models prior to translation. In “product-oriented” approaches, classes are focused on analysis of errors or inadequacies related to style and content in progression of representative texts proposed. Feedback will thus depend on the combination of teaching approaches from the options above. Most importantly, it needs to be taken into account that feedback in online courses can be extremely time-consuming for the instructor. Therefore, providing effective, meaningful feedback that has the maximum impact on students learning is paramount for any effective instruction, both for the students to achieve their goals, as well as for the instructor to efficiently distribute his/her limited time resources.

Studies on the provision of online feedback emerged from the early days of the WWW [11]. For those interested in this topic, research on the role of feedback in regular face to face settings can also help provide a sounder foundation [9]. As a broad summary, it has been found that the overall translation process and the quality of the outcome partially correlate to (1) the type of feedback employed, (2) how it is administered, and (3) how it is presented. In addition, it should be added that online feedback does not only come from the instructor, but (4) it can also be provided by peers (e.g. group work online or collaboration in documents in cloud-based servers, exchanging translations or exchanging access as reviewers in cloud-based CAT tools), professionals (internships) or the crowd (e.g. asking questions in Proz.com or participating in crowdsourcing initiatives), as well as the provision of automated feedback (in terms of automated quizzes, multiple choice selection, etc.). Neunzig and Tanqueiro [11] published the first and most comprehensive classification of the types of feedback in online translation courses. They conducted a study to identify how different types of feedback correlate to acquisition of learning objectives and improvements in the quality of students' translations. Online learning technologies have evolved since the publication of this study (for example, videoconferencing now allows to conduct lectures, break up students into virtual discussion rooms, collaboration in cloud-based translation documents in real life during lectures in Google Docs or Word 365). Nevertheless, the main categories of feedback are still relevant for anyone interested in the provision of efficient feedback. These categories are first and foremost categorized depending on (1) how the feedback is administered, and (2) when it is administered:

1. How the feedback is administered: Individual or non-individual feedback – provided to the entire group.

2. When it is administered:

2.1. Delayed Individual feedback. Delayed individual feedback is the most common form of feedback in online courses, and entails providing a corrected translation with comments. This is done for exams or individual graded assignments. This type of feedback tends to be highly beneficial for grammar or language mistakes, but a bit less so for other types of translation errors. In order to improve the efficiency of this type of feedback, strategies such as having an online final translation portfolio in which students submit a final "polished" translation of "publishable" quality for each graded assignment, guarantees that the time invested in providing feedback by instructors is meaningfully and actively integrated by the student.

2.2. Individual feedback – Contiguous. This type of feedback refers to interventions while the students are conducting the translation through prompts or pop-up messages. It can be informative, just indicating whether the answer provided is true or false, or indicating the type of error that happened. Contiguous feedback can also be corrective, and it is divided into simple corrective feedback and elaborate feedback. Elaborate feedback takes the form of guidelines or prompts that help students find the right answer, or indications of the most appropriate strategy to find an acceptable solution. It could also entail presenting a possible solution with an explanation. Building this type of feedback into the learning process can be complex and time consuming. One possible way to integrate this type of feedback into a synchronous class involves having a class connected to a videoconferencing system and at the same time have all students be part of a cloud-based shared doc (in Google Docs or Microsoft 365 for example). A student can type a proposed solution and the instructor, and students, can comment real time on the proposal by student, correcting any possible proposed rendering.

2.3. Non-individual feedback–Anticipatory. This involves providing instructions and guidelines, key problem-solving strategies and-or attention to key translation problems and issues prior to engagement in the translation. Normally translations need to include the "translation brief", a notion brought to us by functionalist theories of translation (Nord 1997) where instructions for the translation are provided. This includes the intended audience (e.g. Latin America, US or Spain for Spanish), intended purpose, function of the translation (e.g. a company wants to get a product sold in this market, a local government wants to advertise its historical and cultural features to boost tourism in an online website), etc. Anticipatory feedback can be provided in terms of an extended video that presents the translation and identifies the main challenges or "difficulties", what the PACTE group refers to as "rich points" [15] that connect the actual translation assignment to the learning goals for the translations. These video presentations do not include actual solutions to any problem, but rather, point at the problem, frame it, and direct students to possible mechanisms to solve it. This is similar to the most effective feedback found by Neunzig and Tanqueiro [11], "elaborate feedback" that provides guidelines and possible ways to solve problems, rather than the solution itself. For example, if a text includes any measurements, students can be reminded that km or hectares need to be adjusted for a US audience that is not acquainted with them. Rather than indicating the solution, anticipatory feedback points at the resources to solve the problem, such as Google conversion tools. Similarly, a specialized text about international trade can contain specialized terminology that can be found in terminology databases such as the IATE European Union Database[1]. Nevertheless, it is key for students in this terminological search to learn that they need to include the specialized domain of the source text in question in the search so that they obtain accurate results. Another example can be dialectal variation. In the case of Spanish, students can be directed to find the most frequent term for international or neutral Spanish, or the term preferred for any country the translation is

[1] https://iate.europa.eu/

intended to. For this purpose, dialectal variation tools such as Diatopix[2] can be used (only for English, French, Spanish and Portuguese).

2.4. Non-individual – Delayed. This type of feedback to the entire group can come in form of a video explaining the main problems in a graded translation assignment summarizing the main issues, how to solve them, main strategies to avoid common pitfalls in the exam or assignment. It can also come in form of a translation sample from the course with the most common issues critically discussed, pointing especially at solutions to common translation problems, lack of problem identification by students (a common issue for novices), etc.

3. Additional types of feedback

3.1. Anticipatory or delayed - Consult with translation model. This was found to be the least effective of them all, and though it is extremely useful in some contexts (i.e. for translation analysis, criticism, etc.) it is not recommended as general method. It is possible to produce as a collaborative effort in the course, synchronously or asynchronously (in a shared doc or using a discussion forum), a translation model that is the result of a training session and thus, the result of intensive feedback and students-students and students-instructor collaboration.

3.2. Simple delayed individual feedback. Students provide a translation and the professor grades and comments on the translation. This is one of the basic approaches and it is helpful to some extent, but in itself it is time-consuming and it is not fully efficient to increase the quality of students' performance (if compared to other more interactive methods based on socioconstructivist approaches that, nevertheless, do include commentary and editing by instructors).

4 How feedback is incorporated at Rutgers University online translation courses

The feedback loops implemented at asynchronous online translation courses at Rutgers University entail primarily a dynamic gradual process. In each unit, students receive initially anticipatory feedback in the form of video presentations on the unit translations (as well as any other theoretical presentations, readings, etc.) focusing on the "rich points" in each translation and possible problem-solving strategies (students need to implement them and find their own solution alone or in pairs-groups depending on the assignment). These "rich points" are connected to the learning goals for each unit and they are representative of the most common issues on the prototypical textual genres assigned (business letter, recipes, children's stories, purchase contracts, medical inserts, UN resolution, research paper, patents, etc.). Students, in groups or individually, submit a draft of the translation. An online forum then opens where the instructor directs students to provide commented solutions for these "rich points".

[2] http://olst.ling.umontreal.ca/diatopix/?lg=en

Over several days, students and instructor engage in a discussion (on forums or videoforums) on the main issues, and the instructor proactively comments and presents and guides students towards the best solutions and problem-solving strategies. Students then have to incorporate the comments from the forums in their translation drafts and provide a "final translation". This translation is then graded and returned with comments and this can be considered as delayed individual feedback. Over the course of the online unit, students have already received group or non-individual anticipatory feedback, elaborate feedback in the forums, and peer feedback. In addition to the individual delayed feedback from the instructor, and the group delayed feedback in the form of a video that summarizes the main issues most students had in their final version of the translation, this methodology provides a richer approach in terms of online feedback and a better approach to help students incorporate in their own learning style and progression the teachings from each unit. Alternatively, in synchronous teaching, the use of videoconferencing combined with shared cloud documents can provide an ideal platform for instructor-entire class, instructor- class groups to collaborate on translation assignments, leading to the production of a group translation version that is the result of a collaborative learning process.

5 Conclusions

Translation feedback in didactic contexts can become an extremely time-consuming activity, and more so for those who had to quickly adapt to online environments due to the Covid pandemic. Providing feedback in an efficient manner that maximizes the intended effect on the learner should be the main goal of anyone engaged in online education. This paper has reviewed the significance of providing effective feedback in online environments to both improve the learning process of students, while taking into considerations the time limitations for instructors. It has presented an updated categorization of online feedback based on the publication by Neunzig and Tanqueiro [11] and how this has been applied in the online translation program at the undergraduate and graduate program at Rutgers University. To finish with, it should be mentioned that anonymous students course evaluations rate similarly online and face to face courses over the years, witness to the fact that the online feedback model developed helps students perceive that both environments are equally suited for their translation competence acquisition process.

References

1. González Davis, M.: Multiple voices in the translation classroom. Amsterdam-Philadelphia: John Benjamins (2004).
2. González-Davis, M., Enríquez-Raido, V: Situated learning in translation and interpreting training. New York-London, Roultledge (2018).
3. Jiménez-Crespo, M. A.: Building from the ground up: on the necessity of using translation competence models in planning and evaluating translation and interpreting programs.

Cuadernos de ALDEEU, Special Issue, Translation and Interpreting Training, 11-42 (2013).

4. Jiménez-Crespo, M. A.: Translation training and the Internet: two decades later. TIS: Translation and Interpreting Studies 9, 33-56 (2015).
5. Kelly, D.: Handbook for Translation Trainers. Manchester, St. Jerome (2018).
6. Kiraly, D.: A Social Constructivist Approach to Translator Education - Empowerment from Theory to Practice. Manchester: St Jerome (2000).
7. Knowles, M.: Self-directed learning: A guide for learners and teachers. Chicago, Follett Publishing Company (1975).
8. Luo, X.: Translation in the time of COVID-19. Asia Pacific Translation and Intercultural Studies (2021). DOI: 10.1080/23306343.2021.1903183
9. Massey, G., and Brändli, B.: Collaborative feedback flows and how we can learn from them: investigating a synergetic learning experience in translator education". In: Kiraly, D. (ed.) Towards Authentic Experiential Learning in Translator Education, pp. 177-199. Göttingen: Mainz University Press (2021).
10. Massey, G.: Process-Oriented Translator Training and the Challenge for E-Learning. Meta: Translators' Journal 50 (2), 626–633 (2005).
11. Neunzig, W., Tanqueiro, H.: Teacher Feedback in Online Education for Trainee Translators. Meta: Translators' Journal 50 (4) (2005). https://www.erudit.org/en/journals/meta/1900-v1-n1-meta1024/019873ar.pdf
12. Olvera Lobo, María D. et al.: Translator Training and Modern Market Demands. Perspectives: Studies in Translatology 13(2), 132–142 (2005).
13. Olvera-Lobo, María D. et al.: Teleworking and Collaborative Work Environments in Translation Training." Babel 55 (2), 165–180 (2009). http://www.ugr.es/~robinson/2009_babel_55.pdf
14. PACTE. "Investigating Translation Competence: Conceptual and Methodological Issues. Meta: Translators' Journal 50, 609-619 (2005). https://ddd.uab.cat/pub/artpub/2005/137444/meta_a2005v50n2p609.pdf
15. PACTE. Researching Translation and Interpreting Competence by PACTE Group. [Benjamins Translation Library, 127]. Amsterdam-Philadelphia: John Benjamins (2017).
16. Shreve, G. M.: Translation and expertise: the deliberate practice. Journal of Translation Studies 9: 27–42 (2006).

The use of corpora in an interdisciplinary approach to localization

Parthena Charalampidou[1] [0000-0002-5047-228X]

[1] Aristotle University of Thessaloniki, University Campus, 54124, Greece

pchar@frl.auth.gr

Abstract. Translation Studies and more specifically, its subfield Descriptive Translation Studies (Holmes 1988/2000) is, according to many scholars (Gambier 2009; Nenopoulou 2007; Munday 2001/2008; Hermans 1999; Snell-Hornby et al. 1994 e.t.c), a highly interdisciplinary field of study. The aim of the present paper is to describe the role of polysemiotic corpora in the study of university website localization from a multidisciplinary perspective. More specifically, the paper gives an overview of an on-going postdoctoral research on the identity formation of Greek universities on the web, focusing on the methodology of corpora compilation and analysis with methodological tools and concepts from various fields, such as translation studies, social semiotics, cultural studies, critical discourse analysis and marketing. The objects of comparative analysis are Greek and French original and translated (into English) university websites as well as original British and American university website versions. Up to now, research findings have shown that polysemiotic corpora can be a valuable tool not only of quantitative but also of qualitative analysis of website localization both for scholars and translation professionals working with multimodal genres.

Keywords: polysemiotic corpora, university website localization, multimodal analysis and corpora.

1 Introduction

Several studies, that focus on websites and their communication and interaction with users, adopt approaches from the fields of cultural and marketing studies. These studies mostly aim at revealing the cultural differences that exist in the online marketing of companies and organizations (Simin, Tavangar and Pinna 2011; Salerno–O' Shea 2006; Dormann and Chisalita 2002; Leonardi 2002; Robbins and Stylianou 2002; Schmid-Isler 2000; Marcus and Gould 2000; Sheppard and Scholtz 1999; Russo and Boor 1993; del Galdo 1990 e.t.c.). A lot of research has also been conducted with tools and methodologies from the fields of linguistics and more specifically, text linguistics and critical discourse analysis for the study of textual genres such as websites and textual functions in multimodal texts (Santini 2010, 2007, 2006; Bateman 2008; van Leeuwen 2008; Askehave and Nielsen 2004; Lemke 2002; Yli-Jokipii 2001; Fritz 1999; Storrer 1999; Wee 1999; Landow 1997; Martin 1997; Bohle 1990; Reiss 1971/2002 e.t.c.). As

Translation and Interpreting Technology Online, pages 216–226,
July 05-07, 2021.
https://doi.org/10.26615/978-954-452-071-7_025

multimodal texts combine more than one semiotic system to create meaning and achieve their communicative goal, there is an urgent need for the adoption of methodological tools from the field of semiotics. This is obvious in various studies, which focus on the multimodality of genres, using semiotics-oriented methodological tools (Tomášková 2015, 2011; Jewitt 2009/2011; Ventola and Guijarro 2009; Baldry and Thibault 2006; O'Halloran 2004 e.t.c.). The application of theories and methodological tools from the above-mentioned fields could help us a) recognize the university website communicative function, b) study whether this function is retained or not in different language versions and c) highlight the parameters that define its retention or modification.

For the systematic and comparative analysis of our research material, we have created a mini corpus, comprising of both pictorial and verbal elements with the UAM Image Tool (O' Donnell 2008). These corpora consist of the homepages of Greek and French university websites, in original and universalized (Floros and Charalampidou 2020) versions, as well as the original versions of American and British websites. In the second case, English is used as an original language, so it is a good point of reference and comparison with versions that are localized in English. The main criterion for the analysis of the specific language versions was the familiarity of the researcher with the respective languages. Additionally, in most Greek and French university websites, the alternative version provided, in most cases, is English, a phenomenon with an increasing tendency in other countries as well (Callahan 2012).

2 Research Questions

University website localization is a challenge both for the translation scholar and the translation professional. It is also of great research interest due to the multimodal nature of the texts involved (hypertexts) as well as the cultural and social dimensions it can take. However, it has been rather unexplored, up to now, in the international literature. There is only limited literature on the interlingual study of university websites, which focuses mainly on content transfer and language policies (Apperson 2015, Tomášková 2015, Callahan 2012, 2006, Simin, Tavangar and Pinna 2011, Bernardini, Ferraresi,Gaspari 2010). Also, very few studies refer to Greek university websites (Callahan 2012, 2006). However, these are insufficient, as the first one is limited to the study of verbal choices in each website (Callahan 2012), and the second one examines solely the macrolevel. Also, given the fact that the second research was conducted fifteen years ago (Callahan 2006), it displays major differences from the current online image of Greek universities. Additionally, according to literature up to now, there is no systematic study focusing on the way Greek higher education institutions approach a foreign audience, and no attempts have been made to create multimodal corpora with Greek university website homepages. Taking into consideration the research gap that exists in the field our research has attempted to:

a) Map the translation practice in the genre of university websites in Greece and interrelate the localization choices, both on macro- and micro-level, with the cultural background of the receivers as well as with the new social and economic conditions on national and international level.

b) Define the way the identity of Greek universities is projected on the web and compare it with the identity of English- and French-speaking universities.

With these goals in mind the following steps have been taken:

- Compilation of a polysemiotic corpus with Greek, French and English university websites with the UAM Image Tool (O' Donnell 2008).
- Recording of content and hyperstructure localization techniques in Greek universalized versions.
- Recording of microtextual localization techniques in Greek universalized versions on verbal and optical level.
- Recording of microtextual localization techniques in French universalized versions on verbal and optical level.
- Comparison of Greek university websites with French and British/American ones on macro- and micro-level.
- Association of localization techniques on macro- and micro-level with cultural, social and economic factors.

3 Methodology

The preliminary study includes six Greek and six French university websites both in their original (French/Greek) and their universalized version as well as six British/American university websites in their original version. The corpus was drawn from the THE World University Rankings for 2019 comprising of universities in a similar ranking. The systematic study and observation of the sum of websites in all four language versions (Greek, French, localized English and original English) required the compilation of a mini corpus (Zanettin 1994). The corpus belongs to the category of specialized corpora which include specialized texts of a specific type and are used for the study of a specific type of language (Hunston 2002: 14). According to Flowerdew (1993:232) corpora which are small in size are adequate when the study focuses on a specific domain. Additionally, specialized comparable corpora (multilingual and monolingual) are used, among others, to track functional translational equivalents ("units" that can be compared on the denotational, the connotational and the pragmatic level) (Tognino – Bonelli 2002).

In the first part of the research we compiled a polysemiotic corpus using the UAM Image Tool (O' Donnell 2008), which is free software that allows the annotation of images, that is the introduction of verbal interpretative information (Habert 2005; Leech 2005). The term 'polysemiotic' is used here to denote corpora that do not include solely one semiotic system, namely language, image or sound but rather combine the annotation of images with the verbal elements that anchor the pictorial meaning. Next, images were annotated manually on different levels and sublevels. Methodological tools from the field of social semiotics (*isotopies* (Greimas 1966), *anchorage* (Barthes 2007), *metafunctions* (Kress and van Leeuwen 1996) constituted the basis for the annotation of the pictorial elements of university websites.

The notion of *isotopy* is a key term in social semiotics and has been suggested by Greimas (1966) who, since the late 1960s, has been a central figure in the Paris School of Semiotics. His theory of structural semantic isotopy can be applied both on lexical and non-lexical units allowing for the description of the coherence and homogeneity of meaning in a multimodal text, such as the website, by connecting figures different from one another. Through the detection of repetitive *semes* (parts of the meaning of a word) the isotopies in a text can be identified and, thus, content analysis is enhanced. Since the aim of our research on the macrolevel was to detect similarities and differences regarding content in university websites, the notion of *isotopy* was adopted.

The multisemiotic nature of the website genre also calls for the study of meaning creation through the synergy of image and text. Another concept that was drawn from the field of social semiotics was that of *image-text relation*. Acccording to Barthes (2007:50-59), the iconic message can be divided into a) literal and b) symbolic. This distinction actually refers to the separation of the denotational description of an image from the connotations that it bears. Taking for granted that every image is polysemous Barthes (2007: 46) suggests that through verbal messages the receiver of the message is directed to the selection of specific signifieds related to the image's signifiers and to the avoidance of others. The verbal message's function in relation to the visual one is called *anchorage* and elsewhere than in advertising its principal function may be ideological since the reader can be directed to a preselected concept (Barthes 2007, 48). In the light of these notions we were able look for the connoted verbo-pictorial messages within university websites and correlate them to their communicative function in each linguistic version.

The third semiotic-oriented tool which allowed for the comparative study of persuasive multimodal meaning making was the grammar of visual design that Kress and van Leeuwen (2006/1996) have suggested. More specifically, these authors (2006/1996) developed a method of social semiotic analysis of visual communication, based on Halliday's social semiotics, and suggested a descriptive framework of multimodality, based on three metafunctions, namely:

a) the *representational metafunction*, which describes what is represented in an image and includes either (i) conceptual processes, which can be attributive highlighting one represented participant of the two depicted or suggestive involving just one represented participant, or (ii) presentational processes that function as a narrative;

b) *the interpersonal metafunction*, which is the representation of relations between image and viewer and describes (i) mood, that is the participant's gaze that can constitute either an offer or a demand, (ii) perspective, which defines the power relations between image and viewer as well as the degree of the viewer's involvement, and (iii) social distance, which refers to the degree of familiarity between image and viewer.

c) the compositional metafunction, which refers to the codes that operate in the layout of an image to produce meaning and create textual coherence such as (i) salience, (ii) reading path, (iii) vectors, (iv) compositional axes, and (v) centers and margins.

For example, for the annotation of the *interpersonal metafunction* in images (Kress and van Leeuwen 1996) various sublevels were created such as:

- the shooting angle on the horizontal and the vertical axis which defines power relations and social distance between the viewer and the represented participants on the image,
- the offer or demand of information, depending on a straight or oblique gaze of the represented participants
- the distance of the shot (close, medium, long) which defines the degree of contact between the viewer and the represented participants.

In the following figures the sublevels of the interpersonal metafunction are depicted:

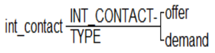


Fig. 1 Interpersonal function: Contact

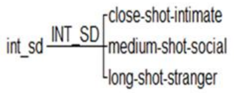


Fig. 2 Interpersonal function: Social distance

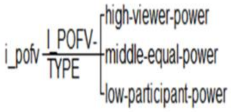


Fig. 3 Interpersonal function: Point of view-vertical axis

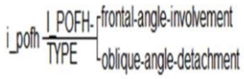


Fig. 4 Interpersonal function: Point of view – horizontal axis

Besides qualitative analysis, the tool provides information regarding the frequency and distribution of the images' characteristics. For the polysemiotic analysis of website content, we added the verbal elements that relate to the image. That is, during annotation, we added the field "text" which *anchors* (Barthes 2007) to the annotated image.

In the following stages of the research further analysis of verbal content will be attempted through the compilation of a second type of corpora using the method of corpora compilation through the internet. Using the WebBootCAT tool, available in SketchEngine (Kilgarriff 2013; Kilgarriff and Grefenstettey 2003), which provides automatic annotation, we will attempt to create a corpus with university websites' verbal elements. This corpus will allow an in-depth study of the verbal realization of the discourse under study through the creation of concordances, statistical charts e.t.c. We are also planning to align the original and universalized versions of Greek university websites using SDL Trados Studio. For the analysis of verbal content we are planning to study the localization of the websites' communicative function (Reiss 1971/2002) and the verbal devices that realize it. The results from the polysemiotic corpus and the parallel verbal corpus are going to be associated with the sociocultural context on the basis of Hofstede's cultural dimensions (Hofstede 1991) and Hall's high- and low-context cultures (Hall 1976).

The methodology adopted, including both types of corpora is depicted in the following figure:

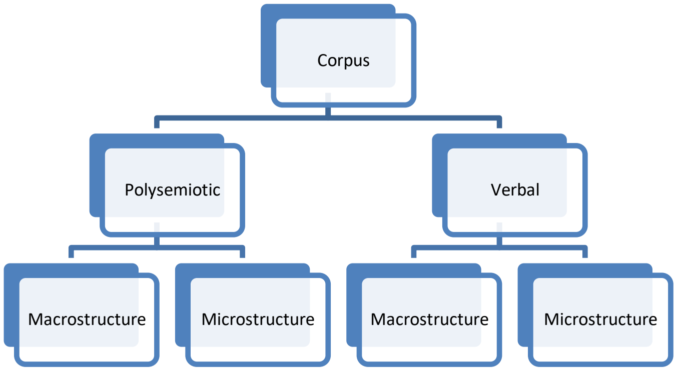


Fig. 5 Methodology for the interlingual study of university websites

The whole process includes two parallel corpora that include Greek and French websites (original and universalized versions) and a comparable monolingual corpus (Mc Enery and Wilson 1996; Peters et al. 1996).

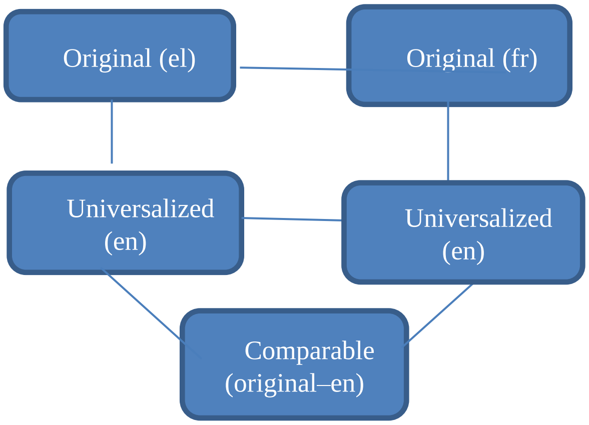


Fig. 6 The five university website versions under study and the comparison relations among them.

In this way, the statistical results regarding pictorial elements from the tool can be combined with the verbal information introduced and thus, enhance analysis of two semiotic systems in parallel.

4 Preliminary results

The comparative analysis of Greek and English/American monolingual corpora, which focused on the study of the denotational and connotational polysiotic meaning of the two versions, has revealed differences in the isotopies used in each case. These differences have been related to different educational ideologies in the respective sociocultural contexts (Charalampidou 2018). The divergence on the isotopic level with reference to online university content depicts the way universities define themselves as well as the way they define their target audience.

The isotopies that prevail in British and American websites express the universities' attempt to project an image in line with the needs of the market. They emphasize on *quality of research, teaching, facilities* and *working environment* which is indicative of their effort to provide *proof of excellence* which will lead to a high score in external assessments. The isotopies selected by British and American universities are, in almost all of the cases, interconnected with the isotopy of *value of giving* reflecting the university's submission to market rules, similarly to Fairclough 's (1993) findings in university brochures. In literature, British and American universities are described as extremely competitive and commodified (Saunders 2010; Hill and Kumar 2009; Olssen and Peters 2005; Hill 2003; Torres and Schugurensky 2002) and this is projected through the isotopies found in their websites. Either in the beginning or at the end of the homepage there is a link through which the user can donate to the university. On the contrary, in Greek university websites a diverging communication strategy seems to be followed in line with a different educational tradition. The values projected are those of *knowledge, continuity and longevity of education and knowledge* and *hellenicity*. The university is self-projected as a place where knowledge and education are generously offered without expecting any rewards and thus, fits a more Humboltian model of education. The use of a different rhetoric in Greek universities is not a surprise since their main resources come from the government and they do not rely on users' donations or students' fees.

Additionally, the compilation of monolingual polysemiotic corpora of the original and universalized Greek and French university websites allowed the study of the translation strategies adopted in each locale and their association with cultural characteristics and marketing principles (Charalampidou and Grammenidis forthcoming). Adopting a translation-oriented approach to localization we defined the notion of translation strategy as an umbrella term that can include the concept of localization strategy in multimodal genres, such as university websites, and attempted to apply functional translation theories to university website localization. What we found was that, although Greek university websites aspire to reach a wider audience, they do not seem to take into consideration the undefined cultural background of the receivers or the expectations that such an audience might have. They address the mean international student retaining the verbopictorial discourse that they use to address Greek-speaking students. On the other hand, French university websites make an attempt to respond to the needs and expectations of an international audience by modifying operative landing content and in many cases recreating content that projects the values promoted by the Bologna Declaration (for more details see Charalampidou and Grammenidis forthcoming).

The conclusions that have been drawn from the first stages of the research reveal that polysemiotic corpora allow the translation studies scholar to adopt an in-depth translation- and semiotics-oriented approach to localization taking into consideration cultural, ideological and marketing parameters. A corpus restricted to text only would limit the research to the observation and analysis of verbal persuasive means, leaving aside their interaction with pictorial elements. However, images very often constitute the basic or even the exclusive means of meaning coneveyance. The compilation of polysemiotic corpora can also be of great use in the context of translator training for the development of multimodal literacy to translation students as well as for their training in the translation of multimodal genres. Since the postdoctoral research is on-going the next step involves creating verbal parallel corpora in SketchEngine and extending the study with statistical results regarding operative verbal devices. In this way, more objective conclusions can be drawn with reference to divergence or convergence in operative discourse depending on cultural dimensions. The extension of the corpora to include a greater number of university websites is also required in order to reach safer conclusions.

References

1. Apperson, G.: How University Websites Portray Study Abroad. Elon Journal of Undergraduate Research in Communications 6 (2), (2015), http://www.elon.edu/docs/e-web/academics/communications/research/vol6no2/01_GinaApperson.pdf, last accessed 2016/12/15.
2. Askehave, I., Nielsen, E.A.: Web Mediated Genres – A Challenge to Traditional Genre Theory. In Working paper nr.6, Centre for Virksomhedskommunikation, Aarhus School of Business (2004).
3. Baldry, A., Thibault, P.J.: Multimodal Transcription and Text Analysis. Equinox, London (2006).
4. Bateman, J.: Multimodality and Genre. A Foundation for the Systematic Analysis of Multimodal Documents. Palgrave Macmillan, Basingstone and New York (2008).
5. Barthes, R.: Εικόνα-Μουσική-Κείμενο [Image-Music-Text]. Translated by G. Spanos. Plethron Publications, Athens (2007).
6. Bernardini, S., Ferraresi, A., Gaspari F.: Institutional academic English in the European contexts: a web-as-corpus approach to comparing native and non-native languages. In: López, A. L., Jiménez R.C. (eds.) English in the European context: The EHEA Challenge. Peter Lang, Bern, pp. 27-54 (2010).
7. Bohle, R. Publication Design for Editors. Prentice Hall, New Jersey (1990).
8. del Galdo, E.: Internationalisation and translation. Some guidelines for the design of human-computer interfaces. In: Nielsen, J. (ed.) Designing User Interfaces for International Use, pp. 1-10. Elsevier, Amsterdam (1990).
9. Callahan, E.: Cultural Similarities and Differences in the Design of University Websites. Journal of Computer-Mediated Communication 11, 239-273 (2006).
10. Callahan, E. and Herring S. C.: Language Choice on University Websites: Longitudinal Trends. International Journal of Communication 6, 322-355 (2012).
11. Charalampidou, P.: In search of the myth in multicultural website design: the case of English university website versions in the British, the American and the Greek locale. In: Frangopoulos, M., Zantides, E. (eds). Design as Semiosis, Special Issue of PUNCTUM. International journal of Semiotics 4(1), 35-62 (2018).

12. Charalampidou, P., Grammenidis. S.: Addressing the international student: translating French and Greek university websites into English. mTm 12, (forthcoming in 2021).
13. Fairclough, N.: Critical Discourse Analysis and the Marketization of Public Discourse: The Universities. Discourse and Society 4 (2), 133-168 (1993).
14. Floros, G., Charalampidou, P.: Website localization: Asymmetries and terminological challenges. The Journal of Internationalization and Localization 6(2), 108-130 (2020).
15. Dormann, C., Chisalita, C.: Cultural values in web site design. In ECCE-11: Eleventh European Conference on Cognitive Ergonomics, September 8-11, Catania, Italy (2002).
16. Flowerdew, J.: Concordancing as a tool in Course Design. System 21 (2), 231-244. Exeter (1993).
17. Fritz, G.: Coherence in Hypertext. In: Bublitz, W., Lenk, U., Ventola, E. (eds.) Coherence in Spoken and Written Discourse. How to Create it and How to Describe it, pp. 221-232. John Benjamins, Amsterdam (1999).
18. Gambier, Y.: Vers de nouvelles perspectives traductionnelles et traductologiques. In: Bulut, A., Uras-Yilmaz (eds) Proceedings of the International Colloquium of Translation: Translation in all its Aspects with Focus on International Dialogue. Istanbul, pp. 32-47 (2009).
19. Greimas A. J.: Structural Semantics: An Attempt at a Method, trans. Daniele McDowell, Ronald Schleifer and Alan Velie. University of Nebraska Press, Lincoln, Nebraska (1966/1983).
20. Habert, B.: Portrait de linguiste(s) à l'instrument. Texto, X/4 (2005).
21. Hall, E. T.: Beyond Culture. Anchor Press/Doubleday, Garden City, NY (1976).
22. Hermans, T.: Translation in Systems. Descriptive and System-oriented Approaches Explained. St. Jerome, Manchester (1999).
23. Hill, D.: Global neo-liberalism, the deformation of education and resistance. The Journal of Critical Education Policy Studies 1(1), 1-28 (2003).
24. Hill, D., Kumar, R.: Global neoliberalism and education and its consequences. Routledge Studies in Education and Neoliberalism 3. Routledge, London (2009).
25. Hofstede, G.: Cultures and Organisations: Software of the Mind. Mc Graw-Hill, New York (1991).
26. Holmes, J.S.: Translated! Papers on Literary Translation and Translation Studies. Rodopi, Amsterdam (1988).
27. Hunston, S.: Corpora in Applied Linguistics. Cambridge University Press, Cambridge (2002).
28. Jewitt, C.: The Routledge Handbook of Multimodal Anlaysis. Routledge, London (2009/2011).
29. Kilgarriff, A.: Terminology finding, parallel corpora and bilingual word sketches. In: The Sketch Engine ASLIB 35th Translating and the Computer conference, London (2013),
30. www.kilgarriff.co.uk/publications2.htm, last accessed 2018/10/8.
31. Kilgarriff, A., Grefenstettey, G.: Introduction to the special issue on the web as corpus. Computational Linguistics - Special issue on web as corpus 29 (3), 333-347 (2003).
32. Kress, G., Van Leeuwen, T.: Reading images: The grammar of visual design. Routledge, London (1996/2006).
33. Landow, G.P.: Hypertext 2.0: The Convergence of Contemporary Critical Theory and Technology. John Hopkins University Press, Baltimore (1997).
34. Leech, G.: Adding Linguistic Annotation. In: Wynne, M. (ed.) Developing Linguistic Corpora: a Guide to Good Practice, pp. 17-29. Oxbrow Books, Oxford (2005).
35. Lemke, J.: Multimedia Genres for Science Education and Scientific Literacy. In: Sclheppegrell, M., Colombi M. C. (eds.) Developing advanced Literacy in First and and Second Languages. Erlbaum, New York (2002).

36. Leonardi, P.: Cultural variability in web interface design: Communicating US Hispanic cultural values on the Internet. In: Sudweeks, F., Ess, C. (eds.) Proceedings Cultural Attitudes Towards Communication and Technology 2002. Murdoch, Western Australia: School of Information Technology, Murdoch University, pp. 297-315 (2002).
37. Marcus, A., Gould, E.W.: Cultural Dimensions and Global Web User-Interface Design: What? So What? Now What? In: Proceedings of the 6th Conference on Human Factors and the Web (2000), http://www.amanda.com/resources/hfweb2000/hfweb00.marcus.html, last accessed 2019/3/1.
38. Martin, J.R.: Analysing genre: functional parameters. In Frances, C., Martin, J.R. (eds.) Genre and Institutions: Social Processes in the Workplace and School, pp. 3-39. Continuum, London and New York (1997).
39. McEnery, T., Wilson, A.: Corpus Linguistics. Edinburgh University Press, Edinburgh (1996).
40. Munday, J. (2001/2008) Introducing Translation Studies: Theories and Applications. London/New York: Routledge.
41. Nenopoulou, T.: Θέσεις και αντιθέσεις: από τη μετάφραση στη μεταφρασεολογία [Positions and juxtapositions: from translation to Translation Studies]. In: 20 Χρόνια Τ.Ξ.Γ.Μ.Δ [20 years Department of Foreign Languages, Translation and Interpreting], Anniversary volume. Diavlos, Athens (2007).
42. O' Donnell, M.: Demonstration of the UAM CorpusTool for text and image annotation. In: Proceedings of the ACL-08: HLT Demo Session (Companion Volume), pp. 13–16 (2008), https://www.uam.es/proyetosinv/woslac/DOCUMENTS/Presentations%20and%20articles/ODonnellACL08.pdf, last accessed 2018/1/7.
43. O'Halloran, K.: Multimodal Discourse Analysis: Systemic functional perspectives. Continuum, London and New York (2004).
44. Olssen, M., Peters, M.: Neoliberalism, higher education and the knowledge economy: From the free market to knowledge capitalism. Journal of Educational Policy 20 (3), 313-345 (2005).
45. Reiss, K.: La critique des traductions, ses possibilités et ses limites.Translation Artois, C. B. Presses Université (1971/2002).
46. Peters, C., Picchi, E., Biagini, L.: Parallel and Comparable Bilingual Corpora in Language Teaching and Learning. In: Botley, S., Glass, J., McEnery, T., Wilson, A. (eds). Proceedings of Teaching and Language Corpora 1996. UCREL Technical Papers 9 (Special Issue), Lancaster University 1996, pp. 68-82 (1996).
47. Robbins, S. Stylianou, A.C.: A Study of Cultural Differences in Global Corporate Websites. Journal of Computer Information Systems 42, 3-9 (2002).
48. Russo, P., Boor, S.: How fluent is your interface? Designing for international users. In: Proceedings INTERCHI '93 Conference on Human Factors in Computing Systems: INTERACT '93 and CHI '93, pp. 342-347. ACM Press, Amsterdam (1993).
49. Salerno-O'Shea, P.: A Comparative Analysis of Website Expressions of National Culture and Mediation. In: CLCWeb: Comparative Literature and Culture 8.2 (2006), http://docs.lib.purdue.edu/clcweb/vol8/iss2/2/, last accessed 2018/2/2.
50. Santini, M.: Interpreting Genre Evolution on the Web. In: EACL 2006 Workshop: NEW TEXT – Wikis and blogs and other dynamic text sources. Preface to the Proceedings. ERCIM news (2006).
51. Saunders, D. B.: Neoliberal Ideology and Public Higher Education in the United States. Journal for Critical Education Policy Studies 8 (1), 41-77 (2010).
52. Schmid-Isler, S.: The Language of Digital Genres: A Semiotic Investigation of Style and Iconology on the World Wide Web. In: Proceedings of the 33rd Hawaii International

Conference on System Sciences (2006), http://csdl2.computer.org/comp/proceedings/hicss/2000/0493/03/04933012.pdf, last accessed 2018/3/7.

53. Sheppard, C., Scholtz, J.: The Effects of Cultural Markers on Web Site Use. In Proceedings of the 5th Conference on Human Factors & the Web (1999), http://zing.ncsl.nist.gov/hfweb/proceedings/sheppard, last accessed 2018/3/7.
54. Simin, S., Tavangar, M., Pinna, A.: Marketing and Culture in University Websites. CLCWeb: Comparative Literature and Culture 13.4 (2011), http://dx.doi.org/10.7771/1481-4374.1703.
55. Snell-Hornby, M., Pöchhacker, F., Kaindl, K. (eds): Translation Studies. An Interdiscipline. John Benjamins, Amsterdam & Philadelphia (1994).
56. Storrer, A.: Kohärenz in Text und Hypertext. In: Henning L. (ed.) Text im digitalen Medium. Linguistische Aspekte von Textdesign, Texttechnologie und Hypertext Engineering, pp. 33-65. Westdeutscher Verlag, Opladen (1999).
57. Tognini-Bonelli, E.: Functionally complete units of meaning across English and Italian. In: Altenberg, B., Granger, S. (eds). Lexis in Contrast: Corpus-based Approaches, pp. 73–95. John Benjamins, Amsterdam 2002.
58. Tomášková, R.: A walk through the multimodal landscape of University Websites. In: Brno Studies in English 41(1) (2015),
59. http://www.phil.muni.cz/plonedata/wkaa/BSE/Articles%20in%20Press/BSE_2015-41(1)-XX_Tomaskova_-_Article_in_Print.pdf, last accessed 2018/2/5.
60. Tomášková, R.: Advertising Education: Interpersonal Aspects in the Genre of University Websites. In: Hopkinson, C., Tomášková, R., Blažková, B. (eds.) Power and Persuasion: Interpersonal discourse strategies in the public domain, pp.44–73. University of Ostrava, Ostrava (2011).
61. Torres, C. A., Schugurensky, D.: The political economy of higher education in the era of neoliberal globalization: Latin America in comparative perspective. Higher Education 43(4), 429-455 (2002).
62. van Leeuwen, T.: Discourse and Practice. New Tools for Critical Discourse Analysis. Oxford University Press, Oxford (2008).
63. Ventola, E., Arsenio J. M. G.: The World Told and the World Shown. Multisemiotic Issues. Palgrave Macmillan, Basingstoke and New York (2009).
64. Wee, C.K.A.: A systemic-functional approach to multi-semiotic texts. (Unpublished Honours thesis). National University of Singapore (1999).
65. Yli-jokipii H.: The local and the global: an exploration into the Finnish and English Websites of a Finnish company. IEEE Transactions on Professional Communication 44 (2), 104-113, 2001.
66. Zanettin, F.: Parallel Words: Designing a Bilingual Database for Translation Activities. In: Wilson, A., McEnery, T. (eds) Corpora in Language Education and Research: a Selection of Papers from Talc 94.UCREL technical papers, 4, pp. 99-111. UCREL, Lancaster (1994).

Author Index