

Machine Translation Summit XVI

<http://www.mtsummit2017.org>

Proceedings of MT Summit XVI

Vol.2 Commercial MT Users and Translators Track

Masaru Yamada, Mark Seligman

MT Summit XVI

September 18 – 22, 2017 -- Nagoya, Aichi, Japan

Proceedings of MT Summit XVI,

Vol.2 Commercial MT Users and Translators Track

Masaru Yamada (Kansai University)

&

Mark Seligman (Spoken Translation, Inc.), Eds.

Co-hosted by



International
Association for
Machine Translation

<http://www.eamt.org/iamt.php>



Asia-Pacific
Association for
Machine Translation

<http://www.aamt.info>



NAGOYA
UNIVERSITY

Graduate School of
Informatics, Nagoya University

http://www.is.nagoya-u.ac.jp/index_en.html

©2017 The Authors. These articles are licensed under a Creative Commons
3.0 license, no derivative works, attribution, CC-BY-ND.

Introduction

The Commercial MT Users and Translators Track at MT Summit XVI, to be held in Japan for the first time in 24 years, features twenty presentations in diverse fields of research from worldwide organizations including academic institutes, enterprises, and individuals in the translation and language technology industry. This Summit is the first since the practical deployment of neural machine translation (NMT), so many of the presentations involve related AI-driven MT technologies. Other studies go beyond traditional post-editing and efficiency scenarios to address the adoption of state-of-the-art MT across the industrial spectrum: topics include MT use cases in crisis scenarios or educational environments; terminology management and QA in systems combining customized MT engines; and many more. Some additional examples:

- Quality evaluation of NMT and comparison with SMT
- Detailed investigation of post-editing errors and efficacy
- Dictation of translation

Many presentations will document the acceptance and integration of neural machine translation technology and its application in real-life scenarios.

Commercial MT Users and Translators Track Co-Chairs
Masaru Yamada
Mark Seligman

Commercial MT Users and Translators Track Co-chairs

Masaru Yamada Kansai University

Mark Seligman Spoken Translation, Inc.

Commercial MT Users and Translators Track Committee

Jan Alexandersson

Srinivas Bangalore

Laurent Besacier

Michael Carl

Mike Dillinger

Kurt Eberle

Satoshi Enoue

Raymond Flournoy

Anthony Hartley

Ron Kaplan

Jeffrey Killman

Isabel Lacruz

Yves Lepage

Rei Miyata

Joss Moorkens

Ricardo Muñoz

Masaaki Nagata

Sharon O'Brien

Akiko Sakamoto

Xiaodong Shi

Michel Simard

Midori Tatsumi

Carlos Teixeira

Kirti Vashee

Yuji Yamamoto

Contents

Page

- 1 Zero-shot translation for low-resource Indian languages
Giulia Mattoni, Pat Nagle, Carlos Collantes and Dimitar Shterionov
- 11 Feature-rich NMT and SMT post-edited corpora for productivity and evaluation tasks with a subset of MQM-annotated data
Kim Harris, Lucia Specia and Aljoscha Burchardt
- 13 Usability of web-based MT post-editing environments for screen reader users
Silvia Rodríguez Vázquez, Sharon O'Brien and Dónal Fitzpatrick
- 26 Live presentations to a multilingual audience: personal universal translator
Chris Wendt
- 27 Towards a full-scale neural machine translation in production: the Booking.com use case
Pavel Levin, Nishikant Dhanuka, Talaat Khalil, Fedor Kovalev and Maxim Khalilov
- 38 The Interact Project and Crisis MT
Sharon O'Brien, Chao-Hong Liu, Andy Way, João Graça, André Martins, Helena Moniz, Ellie Kemp and Rebecca Petras
- 49 A Case Study of Machine Translation in Financial Sentiment Analysis
Chong Zhang
- 59 A New Methodology to Maximize the Strength of SMT and NMT
Yu Gong and Demin Yan
- 67 Rule-based MT and UTX Glossary Management – Honda's Case Dealing with Thousands of Technical Terms
Saemi Hirayama and Yuji Yamamoto
- 79 A detailed investigation of Bias Errors in Post-editing of MT output
Silvio Picinini and Nicola Ueffing
- 91 Terminology-based post-editing of neural MT using the structured glossary data format, UTX
Yuji Yamamoto
- 109 Harvesting Polysemous Terms from e-commerce data to enhance QA
Silvio Picinini

- 116 Translation Dictation vs. Post-editing with Cloud-based Voice Recognition: A Pilot Experiment
Julián Zapata, Sheila Castilho and Joss Moorkens
- 130 Will Neural MT be a Breakthrough in Terms of Post-Editing Productivity in English-to-Japanese Technical Translation?
Tsunao Mikasa and Nobuko Kasahara
- 142 The Impact of MT Quality Estimation on Post-Editing Effort
Carlos S. C. Teixeira and Sharon O'Brien
- 154 Utilizing Neural MT Engines in Industrial Translation
Toru Shishido
- 166 Comparative Evaluation of NMT with Established SMT Programs
Lena Marg, Naoko Miyazaki, Elaine O'Curran and Tanja Schmidt
- 179 Journey around Neural Machine Translation quality
Marco Ganci
- 206 A Reception Study of Machine Translated Subtitles for MOOCs
Ke Hu, Sharon O'Brien and Dorothy Kenny
- 214 TraMOOC - Translation for Massive Open Online Courses: Recent Developments
Joss Moorkens, Sheila Castilho, Federico Gaspari, Andy Way, Rico Sennrich, Antonio Valerio Miceli Barone, Valia Kordoni, Markus Egg, Maja Popović, Yota Georgakopoulou, Maria Gialama, Vilelmini Sosoni, Iris Hendrickx and Menno van Zaanen

Zero-Shot Translation for Indian Languages with Sparse Data

Giulia Mattoni

Pat Nagle

Carlos Collantes

Dimitar Shterionov

giuliam@kantanmt.com

patn@kantanmt.com

carlosc@kantanmt.com

dimitars@kantanmt.com

KantanMT.com, INVENT Building, Dublin City University Campus, Dublin 9, Dublin, IRELAND

Abstract

Neural Machine Translation (NMT) is a recently-emerged paradigm for Machine Translation (MT) that has shown promising results as well as a great potential to solve challenging MT tasks. One such a task is how to provide good MT for languages with sparse training data. In this paper we investigate a Zero Shot Translation (ZST) approach for such language combinations. ZST is a multilingual translation mechanism which uses a single NMT engine to translate between multiple languages, even such languages for which no direct parallel data was provided during training.

After assessing ZST feasibility, by training a proof-of-concept engine ZST on French↔English and Italian↔English data, we focus on languages with sparse training data. In particular, we address the Tamil↔Hindi language pair. Our analysis shows the potential and effectiveness of ZST in such scenarios.

To train and translate with ZST engines, we extend the training and translation pipelines of a commercial MT provider – KantanMT – with ZST capabilities, making this technology available to all users of the platform.

1 Introduction

Nowadays Machine Translation (MT) is an essential tool for the translation industry. The most used MT paradigms are Phrase-based Statistical Machine Translation (PBSMT) (Koehn et al., 2007) and Neural MT (Bahdanau et al., 2014; Sutskever et al., 2014; Cho et al., 2014). While PBSMT has been the state-of-the-art both in academia and industry for the last decade, recently NMT has showed great potential and in many cases has surpassed PBSMT (Bentivogli et al., 2016; Junczys-Dowmunt et al., 2016; Chung et al., 2016; Shterionov et al., 2017).

NMT, similar to PBSMT, is a data-driven MT paradigm, making it strongly dependent on the parallel data used for training. That is, the translation quality of an NMT system correlates with the quality and quantity of the training corpora. Freely accessible parallel corpora are available from various providers, such as: Opus¹, DGT-EC (European Commission)² and Linden/Clarin repository.³ Within the industry, MT

¹<http://opus.lingfil.uu.se/>

²<https://ec.europa.eu/jrc/en/language-technologies/dgt-translation-memory>

³<https://lindat.mff.cuni.cz/repository/xmlui/handle/11858/00-097C-0000-0023-625F-0>

systems are typically built with proprietary data – i.e., data with restricted access, provided by a translation vendor and tailored towards specific translation task(s) mainly because of data confidentiality requirements and/or because the data includes terminology and style that are specific for the translation task. Often, to build a custom MT system (i.e., an MT system that is customised according to a translation vendor’s requirements) that can produce high-quality translations, proprietary and non-proprietary data are concatenated.

For some language pairs, however, there is not enough available parallel data (proprietary or non-proprietary) to build MT systems of high quality to meet users’ requirements. This specifically applies to minority or low-resource languages – languages that have a low population density, are under-taught or have limited written resources or are endangered – as well as to language pairs **with sparse training data** – language pairs for which there have not been documented (human) translations that can be used as training data. Sparsity of training data for language pairs such as, e.g. Tamil or Hindi, which by itself are not low-resource languages, is a phenomenon that hinders the MT industry.

Aiming to overcome the data sparsity within the spectrum of the Indian languages, this paper investigates a zero-shot translation (ZST) (Johnson et al., 2016) strategy for the Hindi and Tamil languages. We built ZST engines on available parallel data to and from English (we use English as an intermediate or a pivot language).

To determine the viability and potential of ZST we first build a proof-of-concept (POC) ZST engine for high-resource languages (English, Italian, Spanish and French). Second, we build a ZST engine for Hindi and Tamil as well as Hindi and Tamil, using parallel corpora with English, Hindi and Tamil data as well as for English, Hindi and Tamil data to prove the applicability of ZST for sparse-data language pairs.

To determine the quality of these engines we compare their outputs with the results of (i) one-to-one NMT engines for the same language combinations and (ii) via a pivoting language. In particular, case (ii) boils down to using two different NMT engines – one that translates from the investigated source language into English and another that translates from English to the investigated target language.

We use the KantanMT⁴ platform to train and translate ZST engines. KantanMT is a custom Machine Translation (MT) platform that allows its users to build custom MT systems covering more than 75 languages. The analysis of the resulting quality is performed through comparison of quality evaluation metrics and the A/B testing interface of the KantanMT platform (KantanLQR™). As a provider of commercial MT solutions, the KantanMT platform is designed and tailored to train and deploy one-to-one translation engines (Phrase-based Statistical Machine Translation and NMT). The research and development of ZST engines imposes certain architectural requirements to the KantanMT platform. In this work, we also discuss the changes that such a platform requires in order to accommodate ZST technology.

The main contribution of this paper is two-fold: on the one hand it is the insights that we draw from our analysis of ZST as a means to tackle the problem of sparse data; on the other hand, we extend the pipeline of a commercial MT – KantanMT.com – making ZST available to KantanMT users.

This paper is organised as follows. In Section 2 we present relevant background and motivate our work; in Section 3 we discuss our MT training and translation pipeline, the changes we have done in order to accommodate ZST technology and we outline the data used to build our ZST engines; Section 4 is devoted to the analysis of the

⁴www.kantanmt.com

translation capabilities of these engines; we conclude our work and present our future plans in Section 5.

2 Background and Motivation

2.1 Zero-shot Translation

Zero-shot translation (ZST) (Johnson et al., 2016) is an approach to train a single NMT engine to translate between multiple languages. Such a multilingual engine can translate from a source to a target language without having seen explicit parallel corpora for that specific language pair during training. ZST exploits transfer learning to overcome the need of building one-to-one translation engines. According to (Johnson et al., 2016), an NMT engine can be trained as a multilingual ZST engine⁵ by simply augmenting the training data with a token before each segment stating the target language. In particular, a sentence S^{L1} in language $L1$ aligned to a sentence S^{L2} in language $L2$ will be augmented with a token $\langle 2L2 \rangle$. Following their findings we exploit a similar approach to augment each segment of the parallel training corpora with a token to indicate the target language. Moreover, we extend this data processing step to handle different tokenisation rules for each language correctly. We add one more token to indicate the language of origin⁶ of the specific sentence that will be used during tokenisation.

In the work of (Johnson et al., 2016; Ha et al., 2016) a single shared attention mechanism and a single ‘universal’ encoder-decoder across all languages is used. Firat et al. (2016) also present a multilingual approach that uses a shared attention mechanism. However, they use multiple encoders/decoders for each source and target language. Aiming at smallest possible alterations of our training and translation pipelines we focus on the single encoder/decoder model with shared attention. Such an architecture does not impose any changes to our platform (i.e. KantanMT), except in the preprocessing (both before training and before translation) step.

In (Johnson et al., 2016), the authors prove that mixing language pairs with little and large available data into a single multilingual NMT model produces a considerable translation quality improvement of the low resource language. This translation capabilities are due to the fact that all the parameters of the multilingual model are implicitly shared by all the language pairs. The analysis on multilingual NMT and zero-shot (or zero-resource) translation, given by (Firat et al., 2016), investigates multiple strategies for multi-way, multilingual translation engines. They show that an NMT engine trained on parallel data without data between two languages translates very poorly for these two languages. In contrast, adding pseudo-parallel data for these two languages to fine-tune the engine improves significantly the quality. They, also, investigate a more basic multilingual NMT engine – trained on two parallel corpora (with or without a finetuning corpus) and is focused to translate between two of these language pairs, in contrast to (Johnson et al., 2016) where the focus is on translating a plethora of languages with one engine.

Motivated by the promising results documented in the aforementioned publications, our main objective is to demonstrate that ZST is particularly beneficial when it comes to MT for language combinations with sparse parallel corpora. We aim to translate one particular language pair (Tamil→Hindi) with a single encoder-decoder with shared attention mechanism NMT engine while using English↔Tamil and English↔Hindi as

⁵In the remaining of this paper we refer to multilingual NMT engines that are trained according to the Zero Shot Approach as *ZST engines*.

⁶We refer to the language of a specific sentence as its *language of origin* to differentiate between source and target languages.

well as a small set of Tamil↔Hindi data.

2.2 Indian languages

Research, conducted on MT for Indian languages, mainly focuses on to- and from-English translation (Sindhu and Sagar, 2016; Antony, 2013). In the survey of Antony (2013) of MT systems for Indian languages there is only one Tamil-Hindi system.⁷ Even exploiting data-driven MT paradigms (such as PBSMT or NMT) that ease the creation and exploitation of MT systems even by non-linguistically informed users, the lack of parallel data is what restricts high-quality MT systems to be built. Ramasamy et al. (2012) present an English-Tamil PBSMT engine as well as a corpus of circa 200000 parallel sentences. Post et al. (2012) present parallel corpora for six Indian languages and English. Bojar et al. (2014) discuss the HindEnCorp dataset which constitutes of approximately 300000 parallel sentences. Another source for data are platforms like Opus and EMILLE. These resources, however are not sufficient (both quantity-wise as well as quality-wise) to build an efficient, domain oriented one-to-one MT engine between two Indian languages.

The aforementioned issues impose a translation gap between Indian languages. We exploiting ZST methodology in order to reduce this gap. We use various available parallel corpora, which we cleansed and organised, to training our ZST engines.

3 Zero Shot Translation Engines

3.1 Pipeline

The KantanMT platform has two main pipelines: one to train an MT engine and a second one to translate text with a selected MT engine. Figure 1 illustrates these pipelines.

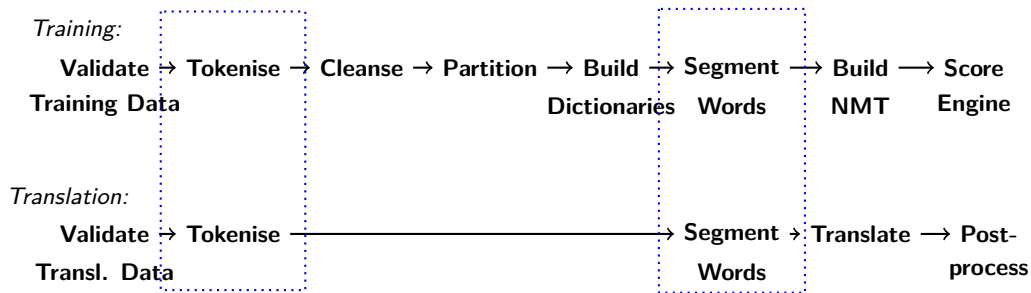


Figure 1: Abstract representation of the training and translation pipelines. Blue boxes indicate processing steps that are common for both pipelines. The input of the training pipeline is source and target data; the input of the translation pipeline is text to translate.

While their core processing mechanisms are different, as shown in Figure 1 they both use the same *tokenisation* step, as well as *word segmentation*. In practice, in the latter step a dictionary is used that is created in the *Build dictionaries* step in the training pipeline; this dictionary is then stored and reused during translation again in the *word segmentation* step.⁸

⁷We refer the interested reader to the (Antony, 2013) for more information on the system.

⁸We present more details about word segmentation and dictionaries in Section 4.

In order to support both training and translating with a ZST engine, it is necessary to adapt these common steps (i.e., the tokenisation and the word segmentation) such that they meet the following requirements:

1. Training and test data are augmented with ZST tokens as defined in Section 2.
2. Different languages require different tokenisation rules which needs to be accommodated in the tokenisation step. That is, the training and test data sets would contain sentences in different languages (see Section 2). The tokeniser would required to know their language of origin and tokenise them according to language-specific rules.
3. Any ZST token is not affected by any consecutive preprocessing step.
4. During both training and translation the output of the neural network does not contain any ZST token.

To meet the first requirement the user needs to introduce ZST tokens for the source and the target data. The target data needs to be augmented with one ZST token, which indicates the *language of origin* of the data. E.g., if the target data is in English, each sentence needs the prefix **zst_en** (if a locale is specified, e.g., British English, the prefix is **zst_en_gb**). The source data, however, needs two ZST tokens: one to indicate the *language of origin* and another to indicate the target language. These have the same form as mentioned above with the first ZST token referring to the *language of origin* and the second one indicates the target language. Example 3.1 illustrates the source and target data, augmented with ZST tokens.

Example 3.1

Source (English, original): It helps for detachment of umbilical cord.

*Source (English, with ZST tokens): *zst_en * *zst_hi* It helps for detachment of umbilical cord.*

Target (Hindi, original): आपको ईमेल एलर्ट के लिए सबस्क्राइब किया गया है।

*Target (Hindi, with a ZST token): *zst_hi* आपको ईमेल एलर्ट के लिए सबस्क्राइब किया गया है।*

In order to meet requirements 2, 3 and 4, we modified the *Tokenisation* step as well as the *Word segmentation* step in our pipelines. The *Tokenisation* step is adapted to read the first from the two ZST tokens from each sentence of the *source* data and the only one ZST token from each sentence of the *target* data and extract the language and locale codes. Then it removes these ZST tokens. Next, each sentence will be tokenised according to tokenisation rules specific for the language and locale codes extracted from the ZST token.

The *Word segmentation* step, which is prior to the *Build NMT* step (in the training pipeline) or to the *Translation* step (in the translation pipeline) will split each word into subword units (Sennrich et al., 2016). During this step the ZST tokens may become segmented which would negatively impact the training of the network. We augment the *Word segmentation* with an extra step to recover any segmented ZST token.

Example 3.2 shows the form of the source and the target data prior to training. The @@ symbols are used as a delimiter for the word segmentation.

Example 3.2

Source (English, original): You have been subscribed to email alerts .

*Source (English, tokenized, word-segmented): *zst_hi* You have been sub@@ scribed to email al@@ er@@ ts .*

Target (Hindi, original): आपको ईमेल एलर्ट के लिए सबस्क्राइब किया गया है।

Target (Hindi, tokenized, word-segmented): आपको ईमेल एल@@ र्त के लिए सब@@ स्क्र@@ @ाइ@@ ब किया गया है ।

Figure 2 shows the changes that were introduced to our pipelines.

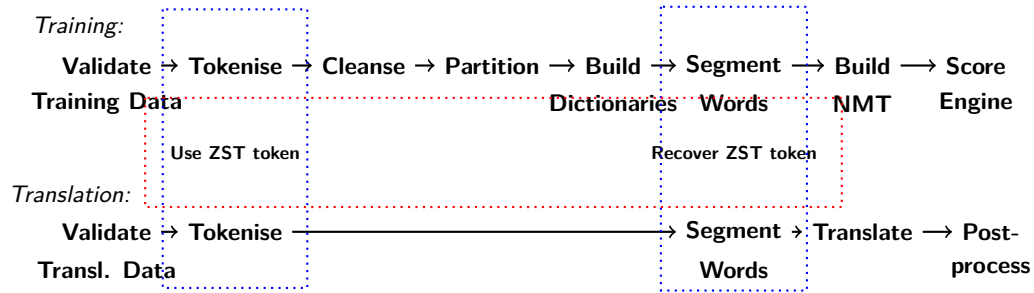


Figure 2: Abstract representation of the training and translation pipelines augmented with additional functionalities required to accommodate ZST. Blue boxes indicate processing steps that are common for both pipelines. The red boxes indicate the additional steps that are required for ZST. The input of the training pipeline is source and target data with ZST tokens; the input of the translation pipeline is text to translate with ZST tokens.

3.2 Engines

With the adapted pipelines we can now easily build ZST engines and use them to translate between language pairs for which parallel data was not provided. In particular, given parallel data set between languages $L1$ and $L2$ as well as between $L2$ and $L3$ we can build a ZST engine that translates a text in $L1$ into $L3$.

Example 3.3 Consider we have available parallel data between English (EN) and Tamil (TA) and between English and Hindi (HI). We use TA and HI data both as source and as target (aligned correctly with their EN counterpart), and the same for the EN data (aligned correctly with the TA and HI) and train a ZST engine:

Source	Target
English	Tamil
Tamil	English
English	Hindi
Hindi	English

This engine would allow us to translate from TA to HI, but also the other way round – from HI to TA. Moreover, it would translate from EN to HI or TA (and vice-versa) as well as from EN to EN.

Example 3.3 shows how we use the available parallel data both as source and as target, aligned correctly, in order to train a basic ZST engine. In general, given data for N languages all aligned with 1 other language (in Example 3.3 that is English) we can build a ZST engine to translate between all of the $N * (2 + N)$ source and target options, including (as in Example 3.3) translating between the same language.

The reason that a ZST engine requires the data to be used both as source and as target is that the neural network will learn to map unseen language pairs through their

Engine Name:	Languages:	Number of Sentences:	Source Words:	Target Words:	Used to translate:	Domain
ZST ₁	EN↔FR, EN↔IT	798 996	15 075 691	15 075 789	FR→IT	Legal
Pivot ₁	EN→FR	198 999	3 844 982	3 475 693	EN→FR	Legal
Pivot ₂	IT→EN	198 999	3 399 530	3 502 284	IT→EN	Legal
ZST ₂	EN↔TA, EN↔HI	1 009 892	15 284 069	15 284 069	TA→HI	General
ZST ₃	EN↔TA, EN↔HI, TA→HI	1 051 631	15 691 380	15 691 380	TA→HI	General, Technical
Pivot ₃	TA→EN	168 871	2 759 734	3 960 123	TA→EN	General
Pivot ₄	EN→HI	268 317	3 338 686	3 620 445	EN→TA	General
one-to-one ₁	TA→HI	41 739	365 571	546 584	TA→HI	Technical

Table 1: Summary of the data used to build ZST and one-to-one engines.

common language.⁹

In the scope of this work we build ZST engines with English, French, Italian data, as well as with English, Tamil and Hindi data. First, we build a proof-of-concept ZST engine on English-French, English-Italian data; we use this engine to translate between French and Italian. To test the performance of this engine we also build two One-to-one engines: one from French to English and another from English to Italian. We refer the latter engines as *Pivot* engines and use them in a sequence to derive an Italian translation, starting from a French text.

Next, we focus on the Indian languages and build two ZST engines - one on English-Tamil and English-Hindi data and a second one on the same English-Tamil and English-Hindi data as well as Tamil-Hindi data. Then we build three one-to-one engines: one Tamil-Hindi, one Tamil-English and a third one English-Hindi all using the same data as for the ZST engines.

Table 3.2 enumerates the available data and the engines we trained.

In Section 4 we present and discuss our findings from comparing the translation quality of these engines.

4 Experiments

We perform our analysis on the MT engines – ZST or one-to-one – enumerated in Table 3.2.

NMT setup. Our training and translation pipelines are based on the OpenNMT toolkit¹⁰ (Klein et al., 2017) version 0.7. As learning optimizer we use ADAM (Kingma and Ba, 2014) with learning rate 0.0005. We train our networks for at least 5¹¹ epochs on NVIDIA G520 GPUs with 4GB RAM (each model is trained on a single GPU). The maximum batch size is 50. The maximum input length used for training is 150.

Dictionaries. Each NMT engine is trained on two dictionaries – one for the source and one for the target data. For ZST engines, we use the concatenated source or target training data to build a source or target dictionary. The dictionaries are composed of word segments in order to increase the vocabulary capabilities of the network and avoid out-of-vocabulary (OOV) problems. We use byte pair encoding (BPE) Sennrich et al. (2016) of 40 000 operations to build the word segments.¹² We prepare the dictionaries from normal-cased (i.e., lower- and upper-cased) tokenised data.

⁹For more details we refer the interested reader to (Johnson et al., 2016).

¹⁰<http://opennmt.net/>

¹¹We present and analyse results of engines with the same number of epochs as to make the comparison fair.

¹²For data in Chinese, Japanese, Korean or Thai, our pipelines use dictionaries based on character-by-character segmentation (Chung et al., 2016). That is, each word segment in the dictionary is a single character. BPE is used for all other languages, including Tamil and Hindi.

Engine:	BLEU*	F-Measure*	Perplexity**
ZST ₂	0.21	3.26	17.12
ZST ₃	9.78	26.40	21.91
one-to-one ₁	8.20	22.16	78.96
Pivot ₃ + Pivot ₄	0.16	16.94	24.85

Table 2: Evaluation of our Indian engines. * - the higher the better; ** - the lower the better.

Result analysis. We began our experiments using a ZST engine consisting of Legal domain data acquired from the European Commission – DGT, which is freely available for use. We decided on a POC engine consisting of English↔French and English↔Italian parallel data sets. We also constructed two one-to-one engines for the same language pairs as the ZST (i.e., Pivot₁ and Pivot₂, see Table 3.2). We started by running 50 sentences of legal domain content that the engines had not seen during training. The translation test set content was in French and needed to be translated into Italian. First, the ZST engine translated the content from French to Italian. Next the same French legal content was translated through the French↔English engine (Pivot₁); then we used the output from this engine as input for the English↔Italian engine (Pivot₂).

We then evaluated both Italian outputs produced by the ZST₁ and Pivot₂ engines running an A/B testing with KantanLQR, KantanMT’s quality evaluation platform. A native Italian speaker with French fluency ranked the translations. The result of the A/B test was conclusively in favour of ZST, with our reviewer choosing 58 percent of the test segments from this engine as better quality than that of the pivot engines. With this result from our POC engine with high resource languages we began experimenting with low resource languages, in particular English↔Tamil and English↔Hindi.

The initial translation tests for our ZST Indian language engine were not as promising as we had hoped from the results of our POC engines. The output was not a complete translation to Hindi but a combination of all 3 input languages of English, Tamil and Hindi. From this result we concluded that we would need more parallel data in both language pairs and possibly aligned data for Tamil↔Hindi to help bridge the sparse data gap. We augmented our test data (the statistics of our data to built the ZST₃ engine, shown in Table 3.2).

We use BLEU (Papineni et al., 2002) and F-Measure (Melamed et al., 2003) to assess the quality of the Indian engines. We also report the perplexity of the engine scored after training is finished. To test whether indeed ZST can improve on one-to-one or pivot engines, we use the same test data set. It contains 500 sentences that are from the same domain of the one-to-one engine (one-to-one₁ in Table 3.2). Our results are summarised in Table 4.

While the enlisted scores for the given test set are in general very low, we observe that the best scores are achieved by the ZST₃ engine – the ZST engine which combines parallel data in different languages and a small set of Tamil↔Hindi data – the BLEU and F-Measure scores for the ZST₃ engine are the highest.

Furthermore, these results confirm that a ZST engine with parallel data for the languages of interest can significantly boost the translation capabilities (compare the scores of ZST₂ and ZST₃).

We ought to note that while these engines may not produce high-quality Tamil→Hindi translations (according to the evaluation metrics reported in Table 4) they show that ZST has a potential and deserves further investigation. Our direct ef-

forts are in bringing a Tamil→Hindi engine together with other Indian languages to industry standards.

5 Conclusions and Future Work

In this paper, we present our first Zero Shot Translation engines for languages with sparse training data. We observed that while ZST produces good quality output for high resource languages (with good training data), it is not performing as good for the Tamil↔Hindi language pair that we used as our main use case. However, our ZST engine that combines multiple-source data and Tamil↔Hindi performs better than the rest of the Indian engines.

Our results showed that further experiments on zero shot translation are needed. First, we will focus on data analysis in order to understand which data combinations are useful for ZST and which are not. Next, we intend to test ZST for other language combinations in order to evaluate which language families or specific languages could benefit the most from such a translation approach.

In addition, with this work we adapted the training and translation pipelines of a commercial MT provider to support ZST engines. In the future we aim to further improve these pipelines and provide more and better ZST services to the users.

References

- Antony, P. J. (2013). Machine translation approaches and survey for indian languages. *IJ-CLCLP*, 18(1).
- Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *CoRR, Accepted for oral presentation at the International Conference on Learning Representations (ICLR) 2015*, abs/1409.0473.
- Bentivogli, L., Bisazza, A., Cettolo, M., and Federico, M. (2016). Neural versus phrase-based machine translation quality: a case study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*.
- Bojar, O., Diatka, V., Rychlý, P., Stranak, P., Suchomel, V., Tamchyna, A., and Zeman, D. (2014). Hindencorp – hindi-english and hindi-only corpus for machine translation. In Chair, N. C. C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).
- Cho, K., van Merriënboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of EMNLP 2014*, Doha, Qatar. Association for Computational Linguistics.
- Chung, J., Cho, K., and Bengio, Y. (2016). A character-level decoder without explicit segmentation for neural machine translation. In *Proceedings of the ACL*, Berlin, Germany.
- Firat, O., Sankaran, B., Al-Onaizan, Y., Yarman-Vural, F. T., and Cho, K. (2016). Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 268–277.

- Ha, T., Niehues, J., and Waibel, A. H. (2016). Toward multilingual neural machine translation with universal encoder and decoder. In *Proceedings of the Thirteenth International Workshop on Spoken Language Translation (IWSLT '16)*, Seattle, WA, USA.
- Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Yonghui Chen, Z., and Thorat, N. (2016). Google’s multilingual neural machine translation system: Enabling zero-shot translation.
- Junczys-Dowmunt, M., Dwojak, T., and Hoang, H. (2016). Is neural machine translation ready for deployment? A case study on 30 translation directions. *CoRR*, abs/1610.01108.
- Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.
- Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. (2017). OpenNMT: Open-Source Toolkit for Neural Machine Translation. *ArXiv e-prints*.
- Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of ACL 2007, demonstration session*, Prague, Czech Republic.
- Melamed, I. D., Green, R., and Turian, J. P. (2003). Precision and Recall of Machine Translation. In *NAACL-HLT*.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, Stroudsburg, PA, USA.
- Post, M., Callison-Burch, C., and Osborne, M. (2012). Constructing parallel corpora for six indian languages via crowdsourcing. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 401–409, Montréal, Canada. Association for Computational Linguistics.
- Ramasamy, L., Bojar, O., and Žabokrtský, Z. (2012). Morphological processing for english-tamil statistical machine translation. In *Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages (MTPIL-2012)*, pages 113–122.
- Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.
- Shterionov, D., Nagle, P., Casanellas, L., Superbo, R., and O’Dowd, T. (2017). Empirical Evaluation of NMT and PBSMT Quality for Large-scale Translation Production. In *EAMT*.
- Sindhu, D. and Sagar, B. (2016). Study on machine translation approaches for indian languages and their challenges. In *Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT), 016 International Conference on*, pages 262–267. IEEE.
- Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*.

Feature-rich NMT and SMT post-edited corpora for productivity and evaluation tasks with a subset of MQM-annotated data

Kim Harris

Lucia Specia and Aljoscha Burchardt

Abstract

This presentation will discuss the creation and practical use of a large data set created through an unprecedented large-scale collaboration between MT R&D and translation experts. It contains post-edited and annotated industry data for four morphologically rich language pairs (EN-DDe, EN-CS, DE-EN, EN-LV). A subset of “almost perfect” sentences also contains MQM error annotations for further detailed analysis and profiling for recurring error patterns. The post edits were performed by professional translators and the data is freely available for further use. The data used for post-editing comprised 20,000 to 45,000 sentences of industry data (IT, life sciences) depending on the language pair. The post-editing of all four language pairs was performed using PET (Aziz, W. et al). Several crucial and novel data points were taken during the post-editing: time logging, keystroke logging quality evaluation of the post-editing effort by the translator upon completion of the post-editing. The recording of this information during the post-editing phase allows for specific features and novel combinations of features to be used for a variety of research- and user-oriented purposes, including establishing the actual post-editing effort by translators based on time and keystrokes and comparing these results to the perceived level of quality of the post-edited sentence, establishing correlations between certain characteristics such as sentence length and post-edit time, or post-edit time and human quality evaluation. The datasets also measure post-editing productivity and are used to detect error patterns in the MT output. This would allow users of MT to adequately assess a) the use of MT in general, b) the actual productivity gains achieved in two different systems or across languages, domains and other data subsets such as long sentences or sentences containing certain grammatical constructs or terminology. For two language pairs identical sets of source sentences comprising 30,000 sentences respectively were post-edited for NMT and SMT output, allowing for a variety of innovative comparisons to be done on the results of the two given the unique data points that were collected during post-editing. In addition, the creation of MQM-annotated subsets of these post-edits for typical industry domains provide

information about error patterns and support feature-oriented quality estimation and evaluation currently unknown to MT quality evaluation and estimation and can be used to improve the MT output

Usability of web-based MT post-editing environments for screen reader users



UNIVERSITÉ
DE GENÈVE



SWISS NATIONAL SCIENCE FOUNDATION



CTTS

Centre for Translation
and Textual Studies



Silvia Rodríguez Vázquez, Sharon O'Brien, Dónal Fitzpatrick

silvia.rodriguez@unige.ch · sharon.obrien@dcu.ie · donal.fitzpatrick@dcu.ie

18-22 September 2017

MT SUMMIT XVI – Nagoya, Japan

Motivation



Advocacy for TEnT accessible design



But why 



❖ Potential **social impact**

- The inaccessible design of popular TEnTs prevents qualified translators with visual and motor impairments from accessing the job market

“Translation tools: help or hindrance?” (Owton & Mileto 2011)

- Translator-Computer Interaction based on:
 - Keyboard-only input
 - Text-to-speech and/or text-to-Braille output
- Other interaction modes: not practical, too time consuming
 - Use of mouse simulation commands
 - Scripting
 - Collaboration with sighted assistant/colleague



❖ Recent research interest on **user-centred factors in translation technology design and evaluation**

- Usability-UX
 - Involvement of end users at design stage (Bota et al. 2013)
 - Usability of FOSS CAT (Veiga Díaz & García González 2015)
 - CAT usability modelling (Krüger 2016)
 - User Interface needs of **post-editors** (Moorkens & O'Brien 2017)
- Multimodal TEnT
 - Mobile **post-editing app** (Torres Hostench et al. 2017)
 - Interactive Translation Dictation (Zapata 2016)
- Ergonomics (Teixeira 2015)



Request for Proposal (RFP)
 “Computer-Assisted Translation (CAT) Tool for facilitating the provision of reference and translation services”

February 2017

Food and Agriculture Organization of the United Nations (FAO)

Accessibility as part of evaluation criteria



n) Ergonomics	
0120	– <u>Available keyboard shortcuts</u> : Some keyboard shortcuts are available.
0121	– <u>Customisable shortcuts</u> : The keyboard shortcuts are customisable.
0122	– <u>Interface Customisation</u> : Whether users are able to customise the interface. Please specify how and to what extent (e.g. size, location, arrangement, background colours of windows, fonts and letter size of menus and of the text displayed in the editor, contents and location of toolbars, etc.) this can be achieved. The software should work on dual screens; in particular, it should be possible to undock panes, if any, and move them to a second screen.
0123	– <u>Learning Curve</u> : As we deal with a number of external translators/revisers experienced with existing Cat Tools, we expect a low learning curve for rapid adoption of a new CAT tool.
0124	– <u>Accessibility</u> : accessibility features are available for people with disabilities.
0125	– <u>OCR features and speech recognition</u> : OCR features exist and some speech recognition software is compatible with the software.



❖ STILL: Scarcity of translation technology research focusing on end-users with special needs

- Exploratory Single Case Studies (Rodríguez Vázquez & Mileto, 2016)
 - Blind user interaction with different versions of SDL Studio
- Questionnaire for blind and visually impaired translators (Rodríguez Vázquez & Mileto, 2016)
 - Low levels of satisfaction with current state-of-the-art desktop CAT
 - Poor interaction CAT-AT (assistive technology)
 - Lack of comprehensive technical support
 - User guides: incomplete + inaccessible
 - *Fluency Now*: Most popular MT-integrated TEnT among users, not necessarily among LSP

No research work found on accessibility of translation tools and MT/post-editing



Goal: Explore the potential of **web-based MT-integrated TEnT** as a more suitable solution for blind translators



Selection Criteria

- Integration of MT
- Free access
- All main components, including post-editing environment, are web-based
- The basic accessibility requirements to enable exploration of the following pages are met: sign up, log in, project creation, post-editing environment

Tools chosen for study:



Method



❖ Classic usability study approach

- Task + questions about user experience
- Summative evaluation
- Remote, asynchronous usability evaluation (Petrie et al. 2006, Murphy et al. 2016)

❖ Snowball sampling

- The Round Table mailing list (approx. 150 subscribers)
<http://lists.screenreview.org/listinfo.cgi/theroundtable-screenreview.org>

INSTRUCTIONS

1. Conduct a simple **post-editing exercise** with each tool
2. Report every problem encountered via a **frustration experience form** (Lazar et al. 2007, Ceaparu et al. 2004)
3. Fill in a **post-task questionnaire** after each exercise
 - Based on Computer System Usability Questionnaire (CSUQ) (Lewis 1995)

Participants - Profile



16 blind translators agreed to participate (consent form)



11 tested at least 1 tool



9 tested both tools

10 blind translators ←



→ 10 blind translators

Female



72%

N=8

Male



27%

N=3

- **Age:** 18-24 (N=2), 25-34 (N=6), 35-44 (N=3)
- **Nationality:** Austria (N=3), Germany (N=2), Italy (N=2), Canada (N=1), Egypt (N=1), Poland (N=1), UK (N=1)
- **Education:** Translation background; university degree (BA/MA) (completed N=9; ongoing N=2)
- **Current occupation:** translator (N=6), public administration (N=1), web analyst (N=1), transcription service manager (N=1)
- **Computer skills** (*self-assessment, 5-point scale*): Adequate (N=1), Good (N=5), Excellent (N=5)

Participants – Use of user agents



Operating System	Windows	Windows
Browser*	Google Chrome (N=3) Mozilla Firefox (N=8) IE (N=1)	Google Chrome
<i>*(2 participants used 2 different browsers)</i>		
Assistive technology[†]	Screen reader only (N=2), screen reader & Braille refreshable display (N=8), per tool	
<i>†(3 participants used 2 different screen readers)</i>	Screen reader: 8 participants used JAWS, 4 participants used NVDA	

CSUQ – Measurement of usability

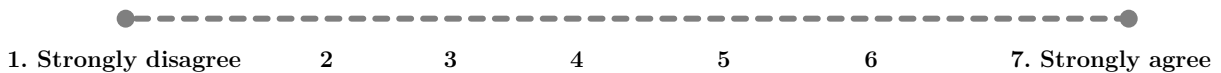


		ITEM		
Overall Usability	System Usefulness	1	Overall, I am satisfied with how easy it is to use this system	Overall satisfaction
		2	It was simple to use this system	
		3	I can effectively complete my work using this system	
		4	I am able to complete my work quickly using this system	
		5	I am able to efficiently complete my work using this system	
		6	I feel comfortable using this system	
		7	It was easy to learn to use this system	
		8	I believe I can become productive quickly using this system	
		9	I felt confident using the system	
	Fit for Information Quality purpose	10	The system gives error messages that clearly tell me how to fix problems	
		11	Whenever I make a mistake using the system, I recover easily and quickly	
		12	The information (such as online help, messages, and other documentation) provided with this system is clear	
		13	It is easy to find the information I needed	
		14	The information provided with the system is easy to understand	
		15	The information is effective in helping me complete the tasks and scenarios	
		16	The organization of information on the system screens is clear	
		17	I found the various functions in this system were well integrated	
		18	This system has all the functions and capabilities I expect it to have	
		19	Overall, I am satisfied with this system	

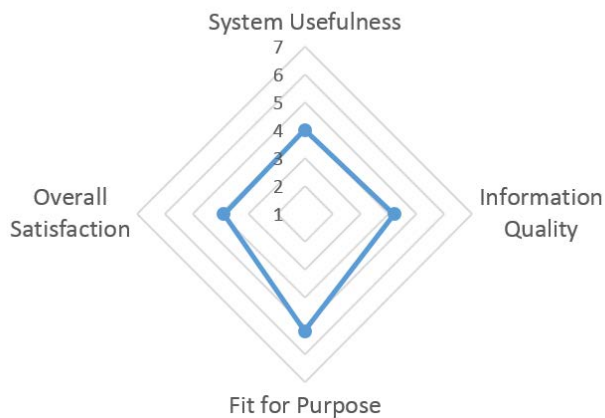
CSUQ Scores (I)



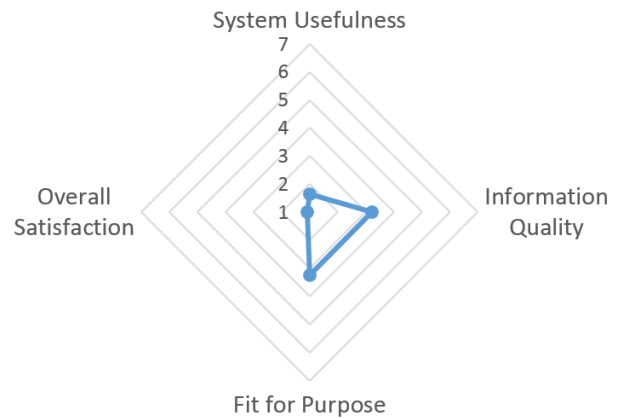
❖ Overall scores



MateCat (CSUQ scores)



Memsource (CSUQ scores)



CSUQ Scores (II)



❖ Overall scores



	Subscale						Overall	
	System usefulness		Information quality		Fit for purpose			
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
	1.64	0.635	3.23	1.020	3.25	0.707	2.37	1.134
	4.00	0.316	4.21	0.476	5.19	0.441	4.20	0.514
p-value (t-test)	<0.001		0.051		0.081		<0.001	

CSUQ Scores (III)



❖ If we look closer, per item (highlights)



	System usefulness			
	7. It was easy to learn to use this system		8. I believe I can become productive quickly using this system	
	Mean	SD	Mean	SD
	3.11	2.315	1.89	1.536
	4.40	2.118	3.60	2.458
p-value (t-test)	0.225		0.086	

Confidence in having successfully completed the task

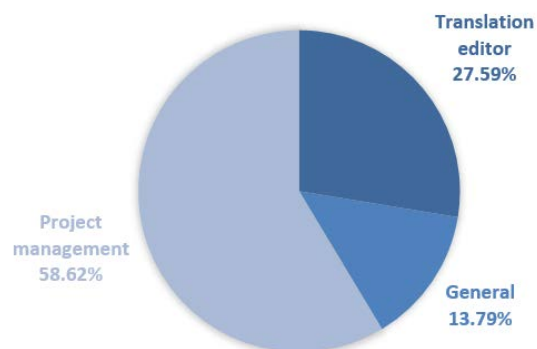
7-point scale, 1 (Not confident at all) and 7 (Very confident)

	1 (80%, N=8) 3 (10%, N=1) 5 (10%, N=1)
	1 (20%, N=2) 4 (10%, N=1) 5 (10%, N=1) 6 (20%, N=2) 7 (40%, N=4)



- Most **problematic steps** during the translation exercise
(“What were you trying to do?”)

	#	%	⌚ (\bar{x} , in min)
Create a new project	9	31.03%	20'
Edit target segment (general)	5	17.24%	37'
Set up the project	5	17.24%	8'
Edit MT suggestions/post-edit	2	6.90%	15'
Upload source file	2	6.90%	30'
Navigate through main menu	2	6.90%	3'
Sign up	2	6.90%	6'
Read translated segments	1	3.45%	2'
Export the target file	1	3.45%	5'



Technical **problem encountered**
(“What happened?”)

Solution or coping strategy
(“How did you solve the problem?”)

	#	%
Non labelled buttons/fields	10	29.41%
Button not working	6	17.65%
Not possible to read own translated text	5	14.71%
Not possible to post-edit	5	14.71%
Lack of content structure	3	8.82%
Lack of information & feedback	3	8.82%
Cursor got stuck in edit field	1	2.94%
Not possible to export	1	2.94%

	#	%
I was unable to solve it	13	44.83%
I figured out a way to fix it myself without help	8	27.59%
I ignored the problem or found an alternative solution	6	20.69%
I knew how to solve it because it has happened before	1	3.45%
I asked someone for help.	1	3.45%



	#	%	Time lost (\bar{x})
Edit target segment (general)	5	17.24%	37'
Edit MT suggestions/post-edit	2	6.90%	15'
Read translated segments	1	3.45%	2'

- ✓ Considered as important (N=2) or **very important (N=6) steps** to complete the translation task
- ✓ Related-problems encountered considered as frustrating (N=2) or **very frustrating (N=6)**

P01: “I could not edit the MT suggestions effectively. I could view the suggestions, but the only way to edit them that I could find was to copy them into the edit field; however, when I did that, the edit field still appeared to be empty and I couldn't edit the text I had just copied and pasted. When I decided to simply write the translation myself, I couldn't read what I had just typed in either; my braille display and screen reader showed an empty edit field.”

P11: “I entered Web Editor. Then, not without difficulties, I found my way to the target segment column. And then I started to write in it. The problem is, however, that NVDA would report what I have just written, but I went back with my edit field cursor, it only read “blank”[...] As long as I am not in full control of target-text editing, I cannot complete even a single segment of my translation.”

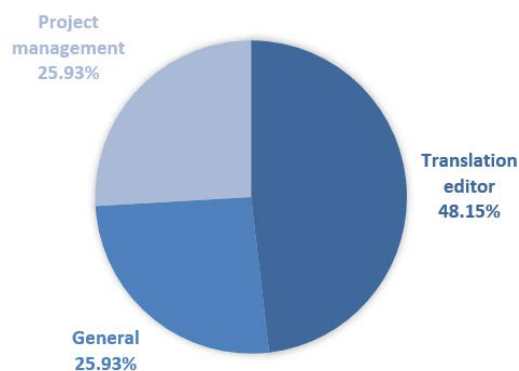
P05: “It wasn't marked up as being an edit field, the target segment was just a line of text. Therefore I couldn't find how to edit this.”



- Most **problematic steps** during the translation exercise

(“What were you trying to do?”)

	#	%	\bar{x} , in min)
Edit MT suggestions/post-edit	6	22.22%	9'
Sign up and login	6	22.22%	10'
Upload source file	3	11.11%	8'
Revise translation	3	11.11%	3'
Edit target segment (general)	2	7.41%	13'
Export the target file	2	7.41%	23'
Set up the project	2	7.41%	3'
Navigate through main menu	1	3.70%	10'
Copy source to target	1	3.70%	15'
Check MT/TM metadata	1	3.70%	5'





Technical **problem encountered** ("What happened?")

	#	%
Screen reader failure	6	21.43%
Button not working	5	17.86%
Not possible to post-edit	3	10.71%
Not possible to sign up	3	10.71%
Lack of information & feedback	3	10.71%
Lack of structure	2	7.14%
Not possible to locate access to editor	2	7.14%
Not possible to export	1	3.57%
Not possible to read long segments	1	3.57%
Manual search/find of segments	1	3.57%
Difficulty editing text	1	3.57%

Solution or **coping strategy** ("How did you solve the problem?")

	#	%
I figured out a way to fix it myself without help	10	37.04%
I was unable to solve it	7	25.93%
I ignored the problem or found an alternative solution	6	22.22%
I asked someone for help.	2	7.41%
I tried again	1	3.70%
I restarted the program	1	3.70%



	#	%	<i>Time lost (\bar{x})</i>
Edit MT suggestions/post-edit	6	22.22%	9'
Revise translation	3	11.11%	3'
Edit target segment (general)	2	7.41%	13'
Copy source to target	1	3.70%	15'
Check MT/TM metadata	1	3.70%	5'

- ✓ Considered as **important** (N=7) or **very important** (N=6) steps to complete the translation task
- ✓ **Variability** observed in **levels of frustration** related to problems encountered

P01: "Starting at the 4th segment, Jaws started behaving oddly while I was trying to read and edit the translation - speech output did not only read everything out loud twice, it also randomly read parts of the following lines."

"I discovered that this only happened when the tags in the target segment hadn't been put in place yet; once I had selected 'Guess Tags' this was no longer an issue. [...] Checking the translation via Braille display worked well, though."

P07: "While I was revising certain (longer) segments, I was no longer able to read the end of the segment, neither using speech output nor with my Braille display."

P15: "MateCat had automatically inserted the MT suggestion. But below the translation it indicated a symbol mismatch. When reading the translation, I noticed that there were strange symbols in the middle of the sentence. When I tried to move the cursor to these symbols to delete them, MateCat crashed, and I had to restart it. This happened several times."

Overall research indicators



- ❖ **None of the tools tested could be professionally used by blind translators in their current form**
 - **BUT:** MateCat could be fully accessible only with minor changes
- ❖ **Blind translators are more resourceful than we thought!**
 - Advanced IT competence (use of multiple AT and browsers), so they can easily adapt
 - But want to be treated as their sighted peers
- ❖ **We need to look for designed-for-all solutions**
 - Tools for blind translators only; e.g. EasyTrans (Al-Bassam et al. 2016): **not** the preferred approach by real end users!

Future Work



- ❖ **In-depth analysis of qualitative data gathered**
 - Levels of frustration; correlation with time lost
 - Technical difficulties logged could provide insights for TEnT developers about what aspects to test (“accessibility check list”)
 - Send report to TEnT providers
- ❖ **Observation study with selected participants**
 - Interaction with more advanced TEnT features
- ❖ **Parallel usability study with sighted translators**
 - Comparison of CSUQ scores
 - Comparison of user preferences regarding information quality and user interface

Thank you



Silvia Rodríguez Vázquez, Sharon O'Brien, Dónal Fitzpatrick



UNIVERSITÉ
DE GENÈVE

silvia.rodriguez@unige.ch · sharon.obrien@dcu.ie
donal.fitzpatrick@dcu.ie



References (I)

- Al-Bassam, Dina, Hessah Alotaibi, Samira Alotaibi, and Hend S. Al-Khalifa. 2016. "EasyTrans: Accessible Translation System for Blind Translators." In *Computers Helping People with Special Needs: 15th International Conference, ICCHP 2016, Linz, Austria, July 13-15, 2016, Proceedings, Part II*, edited by Klaus Miesenberger, Christian Bühler, and Petr Penaz, 583–586. Cham: Springer. doi:10.1007/978-3-319-41267-2_83.
- Bota, Laura, Christoph Schneider, and Andy Way. 2013. "COACH. Designing a New CAT Tool with Translator Interaction." In *Proceedings of Machine Translation Summit XIV*. Nice, France.
- Ceaparu, Irina, Jonathan Lazar, Katie Bessiere, John Robinson, and Ben Shneiderman. 2004. "Determining Causes and Severity of End-User Frustration." *International Journal of Human-Computer Interaction* 17 (3): 333–56. doi:10.1207/s15327590ijhc1703_3.
- Krüger, Ralph. 2016. "Contextualising Computer-Assisted Translation Tools and Modelling Their Usability." *Trans-Kom - Journal of Translation and Technical Communication Research* 9 (1): 114–48.
- Lazar, Jonathan, Aaron Allen, Jason Kleinman, and Chris Malarkey. 2007. "What Frustrates Screen Reader Users on the Web: A Study of 100 Blind Users." *International Journal of Human-Computer Interaction* 22 (3): 247–269. doi:10.1080/10447310709336964.
- Lewis, James R. 1995. "IBM Computer Usability Satisfaction Questionnaires: Psychometric Evaluation and Instructions for Use." *International Journal of Human-Computer Interaction* 7 (1): 57–78.
- Moorkens, Joss, and Sharon O'Brien. 2017. "Assessing User Interface Needs of Post-Editors of Machine Translation." In *Human Issues in Translation Technology: The IATIS Yearbook*, edited by Dorothy Kenny. Oxford, UK: Routledge.
- Murphy, Emma, Enda Bates, and Dónal Fitzpatrick. 2010. "Designing Auditory Cues to Enhance Spoken Mathematics for Visually Impaired Users." In *Proceedings of the 12th International ACM SIGACCESS Conference on Computers and Accessibility*, 75–82. ASSETS '10. New York, NY, USA: ACM. doi:10.1145/1878803.1878819.

References (II)

- Owton, Tara, and Fiorenza Mileto. 2011. "Translation Tools and Software - Help or Hindrance?" *EBU Newsletter - The Voice of Blind and Partially Sighted People in Europe*. <http://www.euroblind.org/newsletter/online/2011/november-december/newsletter/online/en/newsletter/feature/nr/899/>.
- Petrie, Helen, Fraser Hamilton, Neil King, and Pete Pavan. 2006. "Remote Usability Evaluations With Disabled People." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1133–1141. CHI '06. New York, NY, USA: ACM. doi:10.1145/1124772.1124942.
- Rodríguez Vázquez, Silvia, and Fiorenza Mileto. 2016. "On the Lookout for Accessible Translation Aids: Current Scenario and New Horizons for Blind Translation Students and Professionals." *Journal of Translator Education and Translation Studies* 1 (2): 115–35.
- Teixeira, Carlos S. C. 2015. "Researching the Interface of Translation Tools: Ergonomic Considerations." presented at *Points of View on Translator's Competence and Translation Quality*, Cracow, Poland.
- Torres-Hostench, Olga, Joss Moorkens, Sharon O'Brien, and Joris Vreke. 2017. "Testing Interaction with a Mobile MT Postediting App." *The International Journal for Translation and Interpreting Research* 9 (2): 138–50. doi:10.12807/ti.109202.2017.a09.
- Veiga Díaz, María Teresa, and Marta García González. 2015. "Usability of Free and Open-Source Tools for Translator Training: OmegaT and Bitext2tmx." In *Translation and Openness*, edited by Peter Sandrini and Marta García González, 115–30. Innsbruck: Innsbruck University Press. <http://doi.org/10.15203/2936-88-2-6>.
- Zapata, Julian. 2016. "Translating On the Go? Investigating the Potential of Multimodal Mobile Devices for Interactive Translation Dictation." *Revista Tradumàtica*, no. 14: 66–74. doi:10.5565/rev/tradumatica.180.

Live presentations to a multilingual audience: personal universal translator

Chris Wendt

Abstract

The fact that just about everybody carries an internet-connected smartphone enables us to break the language barrier for in-person meetings, presentations, talks and lectures. Using the smartphone in a smart way to connect the audience with the speaker allows all audience members to follow along and participate, regardless of language. Presentations often include specialized vocabulary, like people's names, names of products, company specific acronyms and abbreviations. This is not a challenge for text translation, but in speech translation unknown words are mapped to the phonetically closest known word, which can have a catastrophic effect in translation. Customization of the speech recognition system helps here. We are showing a new and very convenient method to customize the speech recognition system, thus providing a useful automatic interpretation and translation of the speech. It uses the slide deck's content, slides and notes, to customize the SR system at the start of the session, allowing the speaker to use the terms used in the deck, adding the specific terms to the SR system's standard vocabulary. The system displays the transcript in the speaker's language, or a language of choice on the presentation screen, and each audience member may follow in their own language, on their own device. The system extends to the audience. The audience member can ask a question in any of the supported language, speaking or typing into the mobile device. All audience members can read the question in their own language, as well as the answer of the presenter. This makes microphone-runners practically unnecessary. A benefit for audience members with hearing loss is the availability of full transcripts of what everybody is saying. We'll show a few examples situations where this has proven useful. This session provides a view into conducting truly multilingual presentations with full audience participation regardless of language and hearing abilities.

Toward a full-scale neural machine translation in production: the Booking.com use case

Pavel Levin
Nishikant Dhanuka
Talaat Khalil
Fedor Kovalev
Maxim Khalilov

pavel.levin@booking.com
nishikant.dhanuka@booking.com
talaat.khalil@booking.com
fedor.kovalev@booking.com
maxim.khalilov@booking.com

Abstract

While some remarkable progress has been made in neural machine translation (NMT) research, there have not been many reports on its development and evaluation in practice. This paper tries to fill this gap by presenting some of our findings from building an in-house travel domain NMT system in a large scale E-commerce setting. The three major topics that we cover are optimization and training (including different optimization strategies and corpus sizes), handling real-world content and evaluating results.

1 Introduction

Booking.com is one of the largest online companies in the world operating in 43 different languages, connecting millions of daily visitors to 1.4 million bookable accommodations while offering both parties multilingual support and information every step of the way. Given the company's fast growth and a rising need for more high quality translated content, machine translation (MT) is becoming an increasingly attractive option to automate this difficult task.

Our experiments [9] consistently show the superiority of neural machine translation (NMT) systems over the more traditional statistical ones, even when we benchmark them against the well-established and tested general purpose systems. Therefore our recent focus has been on tailoring and improving our own in-house NMT systems to make them practical and effective for us. This work highlights some of the main learnings on our journey and should be of interest to anyone looking to deploy a custom NMT system.

In particular we focus on the following three major topics:

- **Optimization and training**

At Booking.com we have collected tens of millions of travel domain specific human-translated parallel sentences, which in theory allows us to train very flexible models with hundreds of millions of parameters. However learning such system can be computationally expensive which often translates to unacceptably long product development iteration cycles. To address this we first analyze how convergence is affected by different optimization techniques (Section 2.2), including in a multi-GPU environment. Second, we look at how the quality of a trained system improves as a function of the training corpus size (Section 2.3).

- **Handling real-world content**

Real world text comes with many challenges which have to be addressed. Section 3 presents some practical considerations for dealing with named entities and rare words.

- **Quality evaluations**

When building an MT system with customer-facing output, setting up a good quality evaluation loop can be one of the most important aspects. In this part we show how in addition to the BLEU metric [12], the de facto standard for automatic MT scoring, we employ human evaluation of translation adequacy and fluency. We take a close look at how the two approaches correlate. Further, we share our experience developing our business sensitivity framework, which helps us proofread the final translation identifying particularly pernicious errors.

2 Optimization and training

2.1 Model architecture

The core of our translation pipeline is based on OpenNMT [7], which is a Lua written framework for training encoder-decoder neural architectures. Usually, both the encoder and the decoder recurrent neural networks (RNNs), in our case typically long short-term memory (LSTM) units [5], each with 4 layers. We always use (global) attention layer with input feeding to help the model learn faster by keeping a “memory” of past alignment decisions [10]. For European languages we use “case features” (see Section 3.1) as additional input variables from the “cases” embedding space [14]. The main word embeddings are concatenated with the case embeddings to form the inputs to the encoder. At each layer of the encoder the RNNs are bi-directional [13]. Both the encoder and the decoder use residual connections between layers [4] as well as the dropout rate of 0.3 [16].

2.2 Optimization and model fitting

2.2.1 Single-GPU environment

To optimize the training of our NMT system in single-GPU environments, we evaluated different algorithms primarily based on their speed of convergence and translation output quality. The dataset used was English-German property descriptions with one million parallel sentences. We conducted experiments with four well-known optimizers: stochastic gradient descent (SGD) with learning rate decay, Adam [6], Adagrad [3] and Adadelata [18]. Our SGD decay strategy is based on a combination of the perplexity score and epoch number, meaning we decay current learning rate by a multiplicative factor of 0.7 if current epoch’s validation perplexity does not decrease, and after each epoch after the 9th epoch. Our initial learning parameters for SGD, Adam, Adagrad and Adadelata are 1.0, 0.0002, 0.1, and 1.0 respectively. We ran the model for 20 epochs and used both perplexity per epoch and BLEU score after every five epochs on the validation set of 10,000 sentences to measure the performance. Our results are summarized in Table 1 and Figure 1.

As can be seen in Table 1 and Figure 1, we observed that initially Adam converged faster as expected because it applies momentum on a per parameter basis, but SGD took over as soon as decay started and outperformed Adam thereafter. The perplexity reached by SGD in the 9th epoch was already achieved by Adam in the 6th. But from the 10th epoch onward, as soon as SGD learning rate starts decaying indefinitely, Adam’s perplexity is consistently worse than that of SGD. However, there was no decrease in perplexity from 15th till 20th epoch, so SGD already converged by epoch 15. We also observed that Adagrad performed very poorly on our model. Adadelata was much better than Adagrad but still slightly behind Adam and SGD.

Optimizer	Perplexity				BLEU				Time per epoch
	5	10	15	20	5	10	15	20	
SGD with decay	2.37	2.15	2.06	2.06	43.74	45.10	45.84	46.58	6h 11m
Adam	2.26	2.16	2.18	2.24	44.89	45.33	45.21	44.78	+40m
Adagrad	38.75	19.82	15.21	12.55	1.4	2.25	2.56	3.14	+14m
Adadelta	2.62	2.42	2.36	2.32	42.43	43.42	44.35	44.07	+54m

Table 1: Performance of different optimizers on training English-German translation model reported every 5 epochs. Each experiment was conducted in a single NVIDIA Tesla K80 GPU.

We further validated our results using BLEU scores every 5 epochs. The results were mostly consistent with what we observe by looking at perplexity. In terms of time taken per epoch, SGD was the fastest. Adam was about 10% slower in comparison.

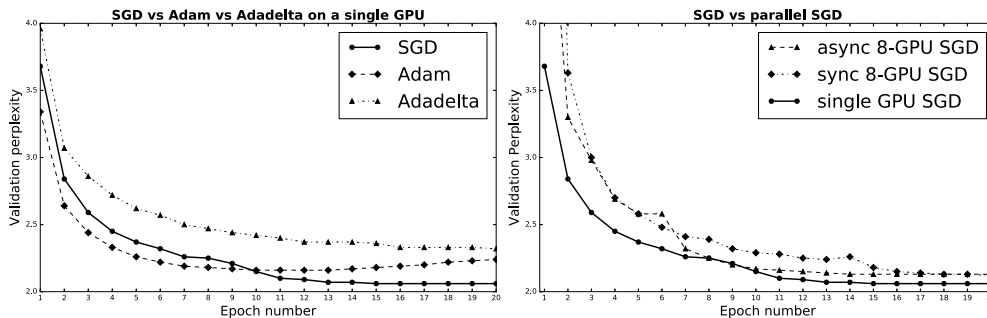


Figure 1: Model convergence. The subplot on the left shows model convergence for three different optimizers: SGD, Adam and Adadelta. Adagrad in our setting did so poorly that it would not fit in the plot (its validation at epoch 20 was above 12). The right subplot compares the convergence of SGD on a single GPU to those of SGD run on an 8-GPU cluster using synchronous and asynchronous parameter updates.

2.2.2 Multi-GPU environment

Next we experimented with the use of multiple GPUs by using data parallelism technique which trains batches in parallel on different GPUs. On a single GPU our model takes 6h11m per epoch on average, and we usually see it converging around 15th epoch, which means training a model on only 1 million sentences takes about 4 days. 15 epochs on a corpus of size 10M could easily translate to around 40 days¹. In an attempt to speed up our development cycle, we ran some experiments with synchronous and asynchronous SGD (with decay) on a cluster of 2, 4, 6 and 8 GPUs. The main difference between these two approaches is that in synchronous mode all gradients are accumulated and parameter updates are synchronized, while in asynchronous each GPU calculates its own gradient and communicates with the “master copy” of parameters independently and asynchronously. This master copy of parameters is stored on a single dedicated GPU which is not used for training. To achieve a faster convergence through better parameter initialization, only one GPU works for the first 6,000 iterations in async SGD.

As can be seen in Figure 2, average time per epoch came down as we added more hardware:

¹Reported estimates do not account for any time related to model checkpointing.

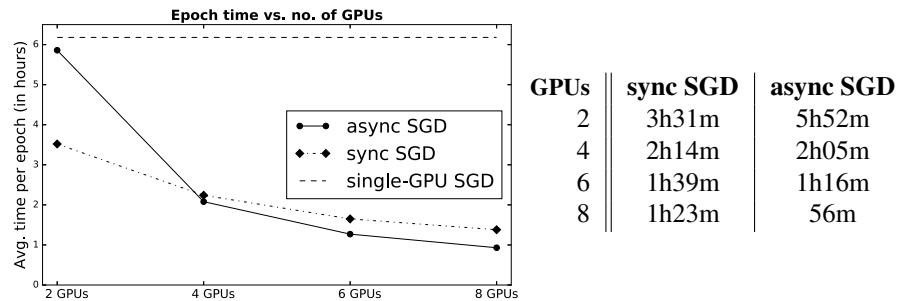


Figure 2: Time per 1M iterations taken by synchronous and asynchronous SGD. On a single GPU the same model takes 6.18 hours.

from 6h11m to 1h23m for sync and to only 56 minutes for async. Note that with 2 GPUs, async takes almost the same time as non-parallel SGD (around 6 hours) while sync is much faster at 3h31m. The reason for that is that 2-GPU async is almost equivalent to a single GPU model as async blocks one GPU completely to store the master copy of parameters and is not used for training. Because async mode skips the overhead of parameter synchronization, it was expected that it would be faster than sync, so we also looked at the quality as measured by perplexity. During the first epoch sync perplexity is much worse than that of async due to only 1 GPU working in async for first 6,000 iterations resulting in better parameter initialization (this cannot be seen in Figure 1 which has been cropped for better visibility; sync has first epoch perplexity of 9.61, compared to 5.61 for async and 3.68 for single-GPU SGD). However, for all remaining epochs their scores are very similar. Single-GPU SGD, on the other hand, performed noticeably better in the first half of the training, but gets quite similar to multi-GPU models eventually (although still marginally better). Overall we are very happy with async’s performance as it is able to reduce the training time by about 85%.

2.3 The importance of corpus size

In order to see how much benefit we get from an increased corpus size, we compared models trained on 1M, 2.5M, 5M, 7.5M and 10M sentences. For fair comparison we report the learning curves as a function of number of iterations (training time) and not the epoch number. Figure 3 shows our findings.

Essentially there were no major surprises. It appears that given enough iteration the model with more distinct sentences will have a higher BLEU score. Notice how in the beginning smaller datasets are actually winning, but given enough training time the model is starting to take full advantage of more data. The largest corpus size of 10M does not have the best performance at the end of 90M iterations, however as we shall see in Section 4.3 this is in fact not true and according to human evaluations 10M gives the best results which are simply not captured by the BLEU metric.

3 Handling real-world content

3.1 Tokenization and case features

In our final models we use byte-pair encoding (BPE) tokenization procedure [15]. BPE is a compression technique which was recently adapted to find optimal tokens for sequence composition in sequence-to-sequence learning tasks. In theory the technique should find a perfect compromise between using word-level translation (and dealing with out-of-vocabulary entities)

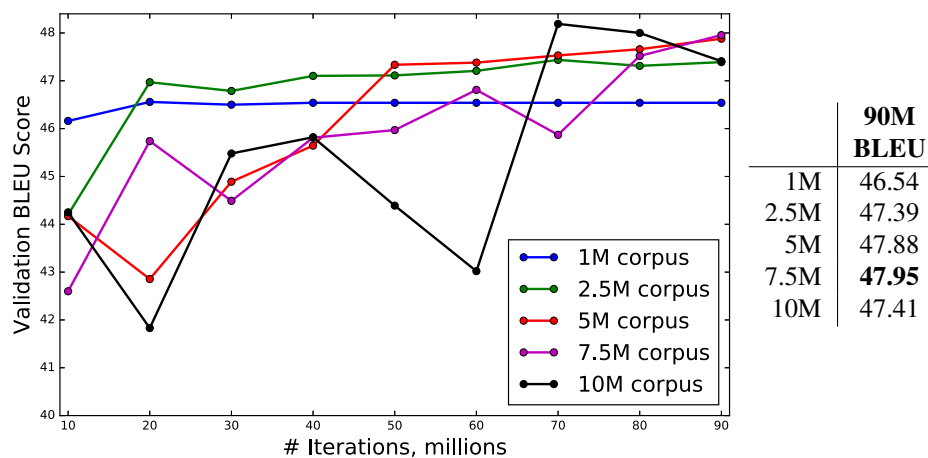


Figure 3: Performance (measured by BLEU score) of a model trained on different corpus sizes, reported every 10M iterations.

and character-level translation (and dealing with much longer sequences of tokens). The procedure is very straightforward. We start with a set of tokens which is the list of acceptable characters and iteratively grow it, at each step adding a concatenation of two items already in the list which is the most frequent in our corpus. The number of iterations can be viewed as the algorithm’s only hyperparameter. We can either apply BPE to the source and the target sentences separately, or we can apply them to the combined corpus. Based on our experiment (see Table 2) we decided ended up with the joint version.

	50k-Vocab baseline^[9]	Joint BPE				Separate BPE			
		30k	50k	70k	90k	30k	50k	70k	90k
Epoch 5	39.54	43.75	43.46	43.40	41.23	42.81	42.35	39.73	N/A
Epoch 10	40.95	44.55	44.52	43.81	43.81	43.39	43.48	43.51	
Epoch 15	42.01	45.08	45.91	46.14	45.75	43.58	43.23	45.17	
Epoch 20	42.15	46.31	46.43	46.61	45.62	45.22	46.00	45.90	

Table 2: Comparison of the BLEU scores of identically trained models with different BPE configurations, as well as the baseline with a vocabulary of 50,000 most common words (see [9] for more details on the baseline model). All experiments were run on 1M corpus. We found 70,000 tokens (70k) jointly trained BPE to have the highest validation BLEU score. Because we saw a strong pattern which made it clear that separately trained BPE 90k model was not going to win, we decided to not run that experiment as it is also the most expensive one.

Apart from applying BPE tokenization we also use case features preprocessing. This allows us to map the same words and word pieces spelled with different cases to the same embeddings while also passing the casing information separately. For example raw terms *book*, *Book* and *BOOK* would all be mapped to the same token *book*, but would have different accompanying case feature values. Case features get their own embeddings which get combined with token embeddings during the translation [14]. In theory this greatly increases the encoding and decoding efficiency of the system, which we also observed in practice through much better

performance over not using case features.

Raw source	Offering a restaurant with WiFi, Hodor Ecolodge is located in Winterfell.
Tokenized source	offering ^C a ^L restaurant ^L with ^L wi [■] fi ^C ^{■,N} ho [■] dor ^L ecolodge ^L is ^L located ^L in ^L winter [■] fell ^L ^{■.L}
Tokenized Output	die ^C ho [■] dor ^L ecolodge ^C in ^L winter [■] fell ^L bietet ^L ein ^L restaurant ^L mit ^L wlan ^U ^{■.N}
De-tokenized output	Die Hodor Ecolodge in Winterfell bietet ein Restaurant mit WLAN.

Table 3: A typical sentence describing an accommodation translated from English to German. Before being fed into the encoder, the sentence is first tokenized using byte-pair encodings. Notice how the words “Hodor” and “Winterfell” which never occurred in our training corpus are split into pieces which are understood by the encoder. The symbol ■ indicates no space between two neighboring word pieces. The superscripts are case features (C: true case, L: lower case, U: all capitals, N: non-alphabetic)

3.2 Handling named entities

Text in the travel domain contains a large amount of entities. There is almost always some destination involved, a property name, distances, times, etc. Although many NMT researchers report results on end-to-end neural networks [1, 2, 17], we often found RNN encoder-decoder architecture insufficient to produce acceptable results, mainly due to mishandled named entities. This section outlines our approach to processing such entities which drastically improves the translation output quality.

As an example, mistranslated distances constitute one of the most common error types when NMT is applied naively on raw text, even with very large corpus sizes (over 10M parallel sentences). Interestingly NMT often correctly converts between kilometers and miles for commonly occurring distances (e.g. 5km, 10 miles); however, the number of distance-related mistakes in our validation set is too large to be left untreated. Another common type of error is related to times and dates (12 vs 24 hour clock times, different date formats).

Source sentence	Winterfell Railway Station can be reached in a 55-minute car ride .
Pure NMT translation	Den Bahnhof Winterfell erreichen Sie nach einer 5-mintigen Autofahrt .
NMT with distance placeholders	Den Bahnhof Winterfell erreichen Sie nach einer 55-mintigen Autofahrt .

Table 4: Translation of a sentence involving a distance using a BPE-based NMT model and an identically trained model with placeholder preprocessing. These types of errors are critical, however they are not adequately reflected in the BLEU score or decoder perplexity change.

In most such cases we used a set of manually created templates to search for entities and replace them with special placeholders. As our team does not understand most of the languages

that we build MT systems for, we get some help from our in-house language specialists (translators). The template refinement cycle goes as follows. We come up with a set of reasonable regular expressions to identify named entities of a certain type in both languages and run them on our parallel corpus. Then we take the set of sentences where the numbers of recognized entities differs between the source and the target. We then look at the breakdown of most common entities in either language which did not have corresponding parallel counterparts, and refine our regular expressions accordingly. At translation (prediction) stage, we preprocess the input to replace all named entities with corresponding placeholders, run the translation, then substitute back the named entities parsed according to the target language format. This simple approach dramatically improves the translation output quality for sentences which involve problematic named entities.

4 Quality evaluation

Unlike simple classification or regression tasks, sequence learning problems are much more difficult to evaluate. The problem comes from the fact that there can be many possible solutions and it is hard (and often impossible) to compare the model output to all valid “true values”. To assess the quality of translations automatically, a useful heuristic is the so-called BLEU score [12] which roughly measures the degree of word overlap between the model translation and a human translation. BLEU score is attractive because it is completely automatic given translated sentences and corresponding model predicted sentences. However, multiple problems have been noted in using BLEU score alone. As a purely counting-based metric, BLEU will favor translation which have more common words and n-grams with the reference translation, regardless of the sentence grammar. It would also penalize models which rephrase the sentence in a way which uses different words from the reference sentence, while preserving its meaning.

In this section we first describe how we leverage our in-house linguistic expertise to score our models in a relevant way (Section 4.1). Then we analyze how BLEU score correlates with human metrics (Section 4.3).

4.1 Human evaluation loop

Our main human evaluation is based on adequacy/fluency methodology² which, as the name suggests, is based on two criteria: *adequacy* and *fluency*. Adequacy shows to what degree the meaning of the source sentence is preserved, while fluency scores how grammatically well-formed (from the native speaker’s perspective) the translated segment sounds. Each sentence is scored by two independent professional translators from English to German (native German speakers). For the experiments in Section 4.3 we chose 200 randomly selected sentences and translators with at least one year of experience professionally translating Booking.com content.

Additionally we use human evaluators to score the quality of entity handling (as described in Section 3.2). For that task each sentence known to contain a specific entity type is given a binary score of whether or not the entity is translated correctly. We found having a separate evaluation specific to entities in addition to adequacy and fluency is important as it helps us to decide on tokenization procedure, entity handling procedures, etc.

4.2 Business sensitivity analysis

One important shortcoming of the BLEU score is that it says nothing about the so-called “business sensitive” errors. For example, the cost of mistranslating “Parking is available” to mean “There is free parking” is much greater than a minor grammatical error in the output. Typically

²<https://www.taus.net/academy/best-practices/evaluate-best-practices/adequacy-fluency-guidelines>

	Precision		Recall		F1 Score	
	EN	DE	EN	DE	EN	DE
Free parking	0.97	0.96	0.93	0.94	0.95	0.95
Non-free parking ^a	0.79	0.83	0.89	0.85	0.84	0.84
Not about parking	1.00	0.99	1.00	1.00	1.00	0.99
Average	0.97	0.96	0.97	0.96	0.97	0.96

(a) Performance of English and German components of our BSF framework measured with a hold-out set of 500 examples.

		German prediction		
		Free parking	Non-free parking ^a	Not about parking
English prediction	Free parking	99.4%	0.5%	0.1%
	Non-free parking ^a	5.1%	94.6%	0.3%
	Not about parking	< 0.1%	< 0.1%	99.9%

(b) The result of applying BSF to our English/German corpus, expressed in matches normalized by the total English volumes. For example out of all English sentences which BSF annotated as “Free parking”, 99.4% also get predicted as “Free parking” in German, while 0.5% of those get identified as “Non-free parking”^a and 0.1% as not about parking at all.

^a Non-free parking can either be a sentence about clearly paid parking, or it can be something ambiguous as “There is parking available nearby”

Table 5: Business-sensitive translation errors analysis for English-German pair for the “parking availability” aspect.

it is very difficult to detect such errors because doing so requires some understanding of the sentence *meaning*. Even so, given the potentially huge cost of such mistakes, we have developed a basic “business sensitivity framework” (BSF) layer to detect certain specific types of errors.

The way it works is rather straightforward. It is a two-stage system, where we first identify the sentences with a particular sensitive aspect (e.g. parking availability, pet policy, etc.) then we apply two classifiers (one to the source sentence, the other to the translation) to identify the predicted values of this aspect (e.g. “free parking”, “pets not allowed” etc.) Finally, BSF flags the sentence as problematic if the predicted aspect values differ between the source and the translation. For the first layer of finding relevant sentences, we learn word and phrase embeddings by training word2vec [11] on our full (monolingual) corpora. Then we pick a few “seed” words or phrases (e.g. “pet”, “dog”, “cat” for the pet policy aspects) and expand the list by looking at those words’ word2vec cosine distance neighborhoods. After our language specialists proofread the list, it is used to identify the relevant sentences via simple keyword matching. For the classification task we use a bag-of-ngrams linear model approach [8].

As an example, Table 5 shows the BSF performance for “parking availability” aspect in English → German translation.

4.3 BLEU score vs human-based metrics

While BLEU score is very convenient to use because it can be computed automatically, the main metrics we really trust are human-based (see Section 4.1). Here we look at how the BLEU scores from our English-German corpus size experiment of Section 2.3 are correlated with adequacy/fluency metrics.

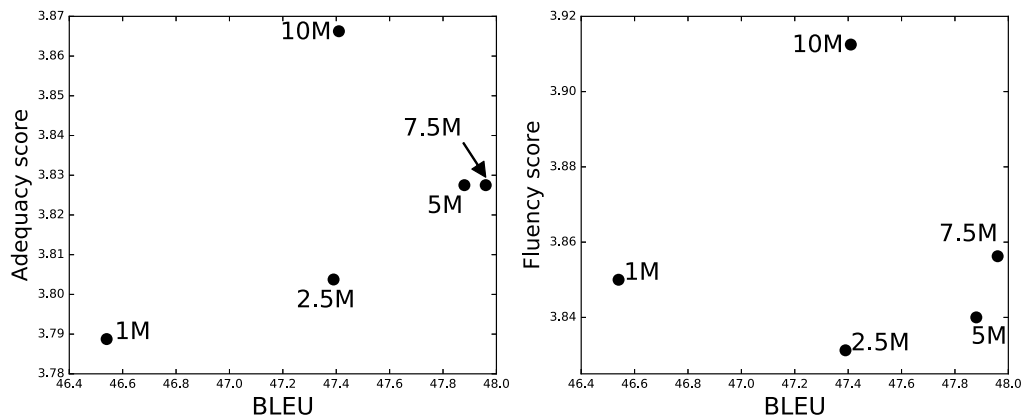


Figure 4: BLEU against adequacy/fluency scores for English-German corpus size experiment from Section 4.1

The results are shown in Figure 4. The training with the corpus size of 10M clearly gives the best performance according to human evaluation, however this is not reflected in the BLEU score. As we can see the correlation between human metrics and BLEU score is rather tenuous. In particular, had we only looked at BLEU, we could have easily made the wrong conclusion about our experiment from Section 2.3.

5 Conclusion

We have presented our approach to developing a large scale NMT system, specifically focusing on practical considerations. We presented the performance of different optimization strategies for model training in single- and multi-GPU environments. We found that a combination of Adam and SGD with learning rate decay works the best on a single GPU, and asynchronous SGD parallelization is a great strategy to dramatically speed up the training. We presented the advantages of BPE tokenization for machine translation and argued in favor of preprocessing named entities for better quality translation. Finally, we presented our approach of dealing with critical translation mistakes through our business sensitivity framework and argue that despite being the main metric in research, BLEU score alone can be a poor way of tracing MT system improvement.

In the future we are going to continue running optimization related experiments, particularly around better strategies for taking advantage of multiple GPUs. In order to leverage our massive monolingual corpora that are not translated, we are also focusing more on the research topics of model pre-training and similar techniques. Other important research topics to us are domain adaptation and user-generated content.

Acknowledgements

We would like to thank our Language Specialists for providing invaluable human feedback and to Darina Kozlova for her important advice on human evaluation and for patiently coordinating all that work.

References

- [1] Cho, K., Merriënboer, B., Bahdanau, D., and Yoshua, B. (2014). On the properties of neural machine translation: Encoder-decoder approaches. In *Proceedings of SSST-8*, Doha, Qatar.
- [2] Crego, J., Kim, J., Klein, G., Rebollo, A., Yang, K., Senellart, J., Akhanov, E., Brunelle, P., Coquard, A., Deng, Y., et al. (2016). Systran’s pure neural machine translation systems. *arXiv preprint arXiv:1610.05540*.
- [3] Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159.
- [4] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.
- [5] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.*, 9(8):1735–1780.
- [6] Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- [7] Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. (2017). OpenNMT: Open-Source Toolkit for Neural Machine Translation. *arXiv preprint arXiv:1701.02810*.
- [8] Langford, J., Li, L., and Strehl, A. (2007). Vowpal wabbit online learning project. *Technical report*, <http://hunch.net>.
- [9] Levin, P., Dhanuka, N., and Khalilov, M. (2017). Machine translation at booking.com: Journey and lessons learned. In *Proceedings of the 20th International Conference of the European Association for Machine Translation (EAMT)*, pages 80–85.
- [10] Luong, M.-T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *Conference on Empirical Methods in Natural Language Processing*.
- [11] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.
- [12] Papineni, K., Roukos, S., Ward, T., and jing Zhu, W. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318.
- [13] Schuster, M. and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.
- [14] Sennrich, R. and Haddow, B. (2016). Linguistic input features improve neural machine translation. *arXiv preprint arXiv:1606.02892*.
- [15] Sennrich, R., Haddow, B., and Birch, A. (2015). Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.
- [16] Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research*, 15(1):1929–1958.

- [17] Xing, W., Lu, Z., Tu, Z., Li, H., Xiong, D., and Zhang, M. (2017). Neural machine translation advised by statistical machine translation. In *Proceedings of AAAI 2017*, San Francisco, CA, USA.
- [18] Zeiler, M. D. (2012). Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

The INTERACT Project & Crisis MT

Sharon O'Brien, Chao-Hong Liu, Andy Way (DCU)
João Graça, André Martins, Helena Moniz (Unbabel)
Ellie Kemp, Rebecca Petras (Translators without Borders)



[INTERACT](#)
[International Network on](#)
[Crisis Translation](#)



[@CrisisTrans](#)

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under the Marie Skłodowska-Curie grant agreement No 734211.



Overview of Interact Project

- International Network on Crisis Translation
- EU-funded Marie Curie Networking Project (with research outputs)
- Based on the main premise that:
 - In today's age of globalisation, communication during a crisis must be multilingual and multilingual crisis communication is enabled through *translation*

Structure

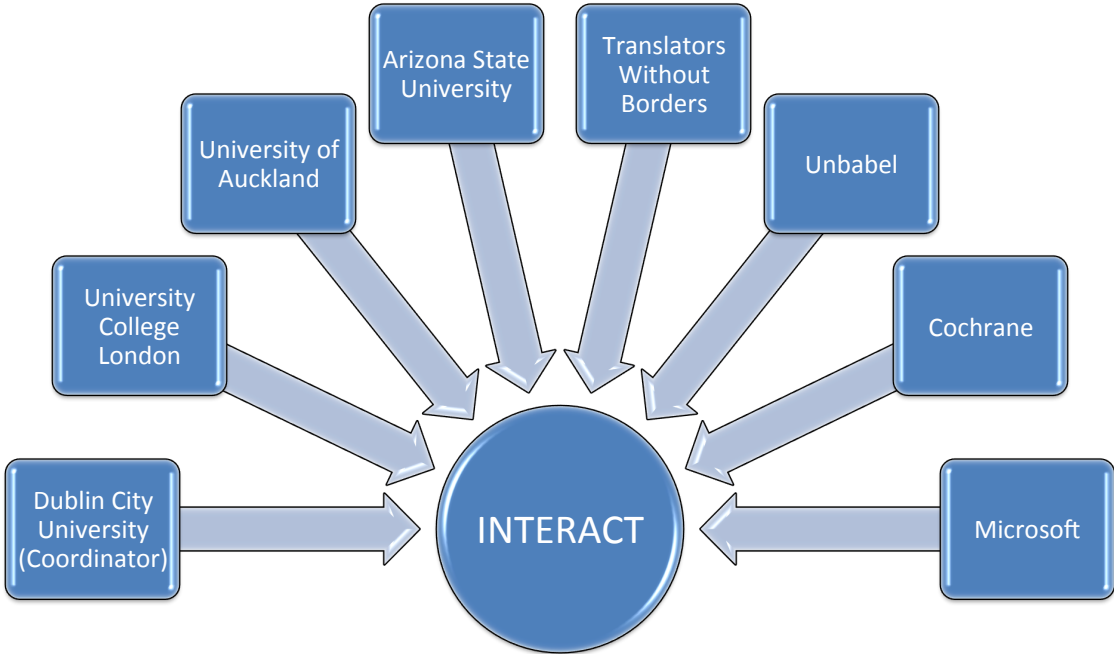
- Brief overview of project
- The role of and challenges for MT in crises
- Previous work
- What we plan to do for MT

3

What do we mean by ‘Crisis’?

- “An event that is expected to lead to a dangerous situation, whether it is an emergency or a disaster”,
Lighthouse Readiness Group
- Project focus:
 - Written Translation
 - Health Content

Partners in the INTERACT Project

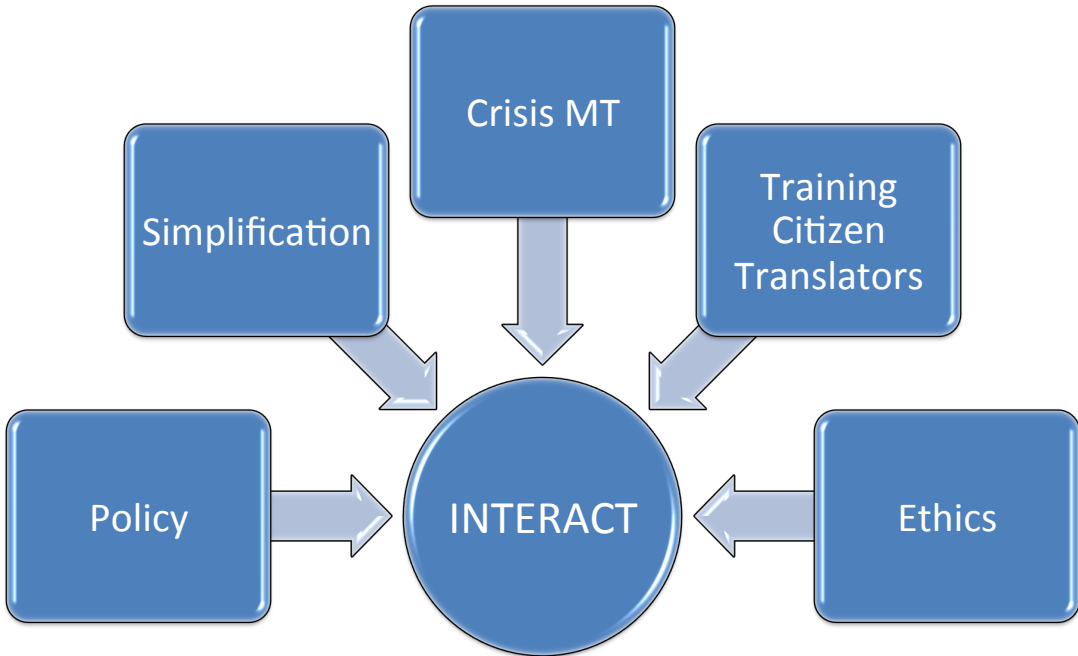


INTERACT/Crisis MT

sharon.obrien@dcu.ie

5
@CrisisMT

Research Work Packages

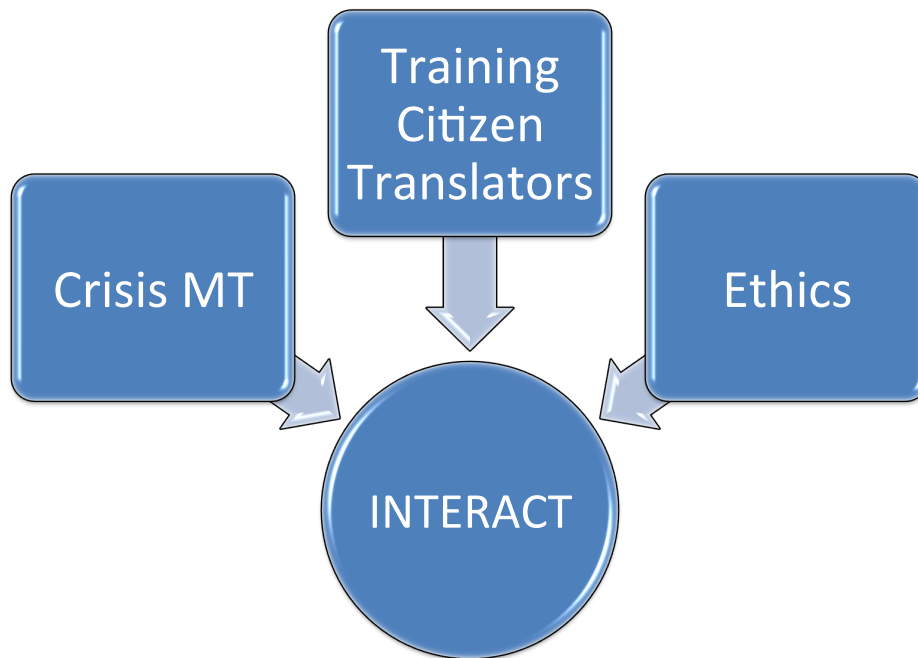


INTERACT/Crisis MT

sharon.obrien@dcu.ie

6
@CrisisMT

Focus For Today



INTERACT/Crisis MT

sharon.obrien@dcu.ie

7
@CrisisMT

Why is (Machine) Translation Important in a Crisis?

- Clear, accurate, timely information is essential in a crisis
 - e.g. Seeger 2006; Fischer 2008; World Health Organisation 2012; Infoasaid 2012; Santos Hernández and Morrow 2013
- Greater cooperation between humanitarian agencies and linguistic volunteers is required (Harvard Humanitarian Initiative 2011)
- but...
- Little to no recognition of the fact that those in need of information may not speak the dominant 'response' language

INTERACT/Crisis MT

sharon.obrien@dcu.ie

@CrisisMT 8

Nature of Crises

- May have sudden onset
- Unpredictable language combinations and information needs
- Unpredictable duration
- Highly stressful
- Lives are at risk

- But also consider: the 4Rs of Disaster Management:
 - Risk
 - Response
 - Recovery
 - Resilience

i.e. Translation is not just required during the *response stage*

Examples of Previous & Ongoing Work

- MT in response to the Haiti Earthquake (Lewis 2010)
 - “Cookbook” for SMT in crisis situations (Lewis et al. 2011)
- DARPA’s Lorelei (Low Resource Languages for Emergent Incidents) project
- TWB’s Rule-Based MT systems
 - English-Kurmanji, English-Sorani via Apertium

Challenges for Crisis MT

- Unknown time of occurrence (no training in advance?)
- Unknown language pairs
- (Often) low resource languages
- (Often) no/little parallel data for affected languages
- (Sometimes) highly specialised (communicating risk of disease, nuclear threat etc.)
- (Sometimes) no/low power and Internet connections
- Text to Speech may be required too

Pivoting as Potential Solution?

- E.g. Arabic-speaking refugees arrive on a Greek Island. Responders speak Greek and limited English. Refugees speak Arabic, some have limited English, and no Greek
- For emergency response purposes, e.g. ascertaining the state of health of refugees,
 - we need Arabic < > Greek translation urgently
- We do not have sufficient translators/ interpreters
- We have Arabic-to-English and English-to-Greek Engines, but no Arabic-to-Greek engines

Pivoting as Potential Solution

- Our questions: Could we use:
 - Arabic < > English < > Greek in this situation?
 - What is the quality like?
 - What impacts the quality?

INTERACT/Crisis MT

sharon.obrien@dcu.ie

@CrisisMT 13

Approaches to Pivoting

- Naïve Approach (Utiyama & Isahara 2007)
 - Translate from A to B
 - Use B as source to translate to C

INTERACT/Crisis MT

sharon.obrien@dcu.ie

@CrisisMT 14

Approaches to Pivoting

- Interpolated Direct Engine (Wu and Wang 2007)
 - Applies only to PBMT
 - A-to-C phrase table is derived from available trained A-to-B and B-to-C models
 - Combined with language model available for C to build direct engine from A-to-C

Approaches to Pivoting

- Neural Interlingua Approach (Johnson et al. 2016)
 - ‘One’ Neural Network is trained with A-to-B and B-to-C sentence pairs
 - NN is used to translate A-to-C even though no A-to-C sentence pairs are used in the training
 - Performance improved if small amount of A-to-C sentences are used in training

Our Pivot Triplets

1. Greek < > English < > Arabic
2. German < > English < > Arabic
3. French < > English < > Swahili

The task will involve identifying relevant training resources, building sample crisis MT engines for health content, and evaluating the results using standard MT evaluation techniques (automatic and human evaluation metrics).

The Human Factor

- (Citizen) Translators
- End Users, e.g. First Responders
- How do we train these people to work with MT in crises?

The Human Factor

- Development of training materials for Citizen Translators who volunteer in Crisis Scenarios
- One specific focus will be post-editing
 - E.g. Through the Unbabel crowdsourcing platform and TWB's and Cochrane's Networks
- Train, evaluate, re-design, train, evaluate...
- Peer review via the crowd (e.g. using Unbabel's online MQM annotation tool)
- Train the trainer
- Training in support resources, e.g. Slándáil, ReliefWeb, TWB harvest of terminology from Sphere Handbook, etc.

The Human Factor

- Focus on end users too:
 - Ethical and informed use of MT for crisis scenarios
 - Use of Quality Estimation for *triaging* MT output so unreliable translation is not even seen by post-editors/end users

To Conclude

- Follow our progress on Twitter:
[@CrisisTranslation](https://twitter.com/CrisisTranslation)



A Case Study of Machine Translation in Financial Sentiment Analysis

Chong Zhang v-chong.zhang@lionbridge.com
Department of Linguistics, Stony Brook University

Matteo Capelletti Matteo.Capelletti@lionbridge.com
Lionbridge Technologies, Inc

Alexandros Poulis Alexandros.Poulis@lionbridge.com
Lionbridge Technologies, Inc

Thorben Stemann v-Thorben.Stemann@lionbridge.com
Lionbridge Technologies, Inc

Jane Nemcova Jane.Nemcova@lionbridge.com
Lionbridge Technologies, Inc

Abstract

The European research project Social Sentiment Indices powered by X-Scores (SSIX) intends to allow Small and Medium-sized Enterprises (SMEs) to take advantage of social media sentiment data for the finance domain. The project aims to overcome language barriers and realize a financial sentiment platform capable of scoring textual data in different languages.

Our approach to achieve this goal takes maximum advantage of human translation while keeping costs low by incorporating machine translation. In the long run, we intend to provide a tool that helps SMEs to expand into new markets by analyzing multilingual social contents.

In this paper, we investigate how sentiment is preserved after machine translation. We built a sentiment gold standard corpus in English annotated by native financial experts, and then we translated the gold standard corpus into a target corpus (German) using one human translator and three machine translation engines (Microsoft, Google, and Google Neural Network) which are integrated in Geofluent to allow pre-/post-processing. We then conducted two experiments. One meant to evaluate the overall translation quality using the BLEU algorithm. The other intended to investigate which machine translation engines produce translations that preserve sentiment best.

Results suggest that sentiment transfer can be successful through machine translation if using Google and Google Neural Network in Geofluent. This is a crucial step towards achieving a multilingual sentiment platform in the domain of finance. Next, we plan to integrate language-specific processing rules to further enhance the performance of machine translation.

1. Background

Over the past two years, Lionbridge has been involved as a leading industrial partner in the European funded [SSIX project](#) (Social Sentiment Index, 2015 - 2018). During the project (which will be completed in February 2018), we have developed a platform for detecting opinions about stocks, companies and their products as expressed in social media and other media sources. For example, we can extract content from Twitter, StockTwits, news, company blogs, etc and analyze sentiment associated to each content.

In Lionbridge, we conceive the SSIX platform as a supporting tool for our sales representatives. Our goal is to make it easier to detect the following aspects:

- What are the needs of our customers
- What prospects may be entering within our areas of expertise
- What are the weak and strong points of our competitors

We consider such knowledge as strategic to trigger appropriate action in real time. For example, we can track customers' needs on social media and adjust our services accordingly in real time; we can detect events that are relevant to our interests and deal with them strategically.

In the past, a sales representative would need to search different sources in an *accessible locale* to find relevant discussion of new products or market updates. This was done in the past manually to a large extent. Such manual approach may not be ideal for many reasons: it is prone to missing information, slow in response time, and expensive in terms of human labor.

Now the SSIX platform offers the possibility to partially automate the search. It allows search terms and media channels to be defined, and it notifies users of changes amongst public opinion. It allows us to see what people say about products and companies in real time. Furthermore, this is not restricted to a specific language and locale. Thanks to the integrated technology of Lionbridge GeoFluent (GeoFluent, Lionbridge Inc.), we can overcome the language barrier and provide financial sentiment analysis across languages.

2. Introduction

One of the primary targets of the SSIX project is sentiment analysis in the financial domain across multiple languages. The work has started with English, where a three-way validated sentiment gold standard has been developed and has been used to train the sentiment classifier. The work on English can rely on several available resources, such as text normalization tools, polarity lexica and distributed word representations that allow the development of a sentiment classifier for English to be based on pre-existing resources.

The work started with building a three-way validated sentiment gold standard corpus for English (Hürlimann et Al., 2016). Three experts in the domain of finance annotated the English corpus manually, and their sentiment scores were reconciled for consistency. This gold standard corpus was used to train and test the SSIX sentiment classifier.

Addressing languages different from English, however, is a more complex issue that raises a series of questions. Resources for other languages may neither be as readily available, nor as good in quality. This raises the question whether it is possible/sufficient to rely on the resources we have for English to address sentiment classification for other languages. Suppose, as it is in fact the case, that we want to develop a sentiment classifier for German when we already have a working version for English. Is there a way to capitalize on the resources developed for English to create a classifier for German?

To answer this question, we suggest at least three approaches:

1. Create a gold standard corpus for German from the ground up, manually annotate and cross review it, and then train the new classifier on it. We call this the *Native* approach.
2. Take the English sentiment gold standard corpus, translate it (either manually or automatically) to German, and train the German classifier on it. We call this the *Derived* approach.
3. Use machine translation to convert the German input to English, and feed the English translations to the English classifier. We call this the *Direct Translation* approach.

The three approaches obviously differ in quality, efficiency and costs. Each approach has its advantages and disadvantages, which are briefly outlined below.

2.1. The Native Approach

Building a new Gold Standard corpus from scratch, as in the Native approach, is expensive, but potentially very rewarding. The most prominent benefit is that no translation is taking place and the native expert judgments are on “first hand” data. Creating such a gold standard is both costly and time-consuming, as we need more than one annotator (at least 3) to agree on the sentiment of each piece of text in order to ensure good quality data. Considering that the sample should contain several thousands of tweets and that a domain like Finance needs judgments made by specialists, the cost may quickly skyrocket. On the other hand, the only variable in the Native Approach is the agreement of the annotators, provided their individual domain knowledge and familiarity with the exchange media (tweets) does not lead to vastly different sentiment scores for the same data. *Due to the conditions of its design and implementation, we could assume that once available, such a gold standard would be the standard against which any other approach should be benchmarked.*

2.2. The Derived Approach

In this approach, instead of building a new corpus and annotating it manually, we use the already existing English language gold standard and translate it to German. This approach presupposes that a statement with positive sentiment in English remains positive in German and vice-versa for negative judgments. Several translation methods are available: It can either be done manually, via machine translation, or in a hybrid way, using computer aided translation tools or post-translation review by human translators. We can also take advantage of the fact that only some words are sentiment-bearing thus targeting these words in context for optimal translation and ignoring the rest.

If we use human translation, the task of creating a translated GS will be cheaper than the creation of a native GS, in the sense that one domain expert will probably be enough, where previously three were needed. Certainly, the cost and time decrease drastically when using machine translation, but the resulting data, especially in a technical domain such as finance, may be of lower quality. Machine translation could, for instance, systematically map an English term to a German term which is synonymous in some other domain, but which is not relevant to the financial domain.

A human-reviewed machine translation is surely the safest approach if one wants to speed up the process and keep costs limited. This may actually reveal error patterns in the translation that can be fixed in post-processing.

2.3. The Direct Translation Approach

Instead of training a new classifier on German data, we translate the German input text to English and feed it to the English classifier. Clearly, translation here can mean only machine translation, as we will be dealing with large amounts of input data to be processed in real time. This approach can also add further costs as machine translation on large amounts of data comes at a cost.

The translation-based approaches in 2 and 3 face a number of issues related to the domain and the specificity of the text involved. Spelling errors, uncommon abbreviations and rhetorical text are all extra challenges that need to be tackled.

Input normalization and output optimization are strategies that can be pursued to improve the quality and accuracy of the translation. First, we may remove elements like repeated characters or delete unknown strings. During post-analysis of translated material, we can map common MT mistakes to the desired output, for instance, terms that need a specific translation in the domain of reference. There is a large range of operations that can be performed – some language-specific, some more general. In this respect, **GeoFluent** [2] is specifically designed not only to support automatic translation but also in preparing the input and correcting the output of the translation process (pre- and post-processing of the data).

3. Setup

The work discussed in this paper is a contribution to the Derived and Direct Translation approaches.

Within the scope of the SSIX project, we built a sentiment gold standard corpus for English, annotated by native experts from the domain of finance (Hürlimann et Al., 2016). The gold standard corpus was translated into a target corpus in German by a domain expert. At the same time, it was also translated into German by three machine translation engines. These are Microsoft, Google, and Google Neural Network, which are integrated in Lionbridge GeoFluent [2]. We used GeoFluent to introduce pre-/post-editing, such as DO-NOT-TRANSLATE rules to tackle special financial terms and text normalization rules.

In SSIX, we intend to take maximum advantage of human translation while keeping the cost low by incorporating the machine translation component. Our objective is to use manually translated data as a benchmark and examine machine translation outputs: their quality and preservation of sentiment in the financial domain.

A crucial prerequisite for our approach is that the sentiment of the gold standard corpus can be transferred to the target corpus after translation. If the sentiment is lost after translation,

either by human or by machine, we cannot use our previous research results, i.e. the English sentiment classifier, and implement either the Derived approach or the Direct Translation approach. The only viable option left would be the Native approach, which is bound to have high costs. As a result, to meet the prerequisite and make decisions for further actions, we must investigate the impact of machine translation on the sentiment quality of the gold standard corpus. We have conducted two experiments to study how machine translation influences sentiment, as discussed below.

4. Experiment 1

The first experiment was designed to find out the quality of each machine translation engine. In this experiment, we selected a sample of 700 English tweets from Twitter and StockTwits relative to the financial domain. This data set was selected for its clarity in expressing sentiment. For example, textual data that did not offer valuable information such as containing only URLs was filtered out to reduce noise.

During the experiment, this sample was translated into German simultaneously by one human translator and the three machine translation engines mentioned above, namely Microsoft, Google, and Google Neural Network, as integrated in Lionbridge GeoFluent. The human translator is a native speaker of German and a domain expert in finance.

To evaluate translation quality for the three machine translation engines, we calculated their BLEU scores (Koehn et al., 2007; for source code see References). Using human translation as the reference, the three machine translations were each compared to the human translation to see how close they are to the professional human translation¹.

The results are summarized in the table below. They suggest that Google and Google Neural Network performed better than Microsoft on 1-gram, and Microsoft performed better than Google and Google Neural Network on 2-grams, 3-grams, and 4-grams.

Engine	1-gram	2-grams	3-grams	4-grams
Microsoft	0.901470798	0.865873923	0.786125067	0.684824095
Google	0.963509145	0.846959705	0.728174371	0.605465403
Google Neural Network	0.963340387	0.846025029	0.727096883	0.604167208

Table 1. BLEU score for machine translations

The 1-gram is used to assess how much information is retained after translation. Clearly Microsoft has lost more information than both Google and Google Neural Network. Among 2-grams, 3-grams, and 4-grams calculations, 4-grams is believed to be the most correlated with judgements made by native speakers of the target languages (Papineni, K., et al., 2002).

¹ We understand that BLEU score is meant to evaluate translations on a corpus level. However, due to time and resource limitations, at this stage we can only investigate the current data sample size. We consider expanding our data size and reduplicating this experiment in order to confirm our results in future.

Our results suggest that Microsoft produced the most similar translations to human translator. Google and Google Neural Network performed more poorly in comparison.

However, we must notice that the BLEU algorithm was not sufficient for our purposes because it only evaluates translation quality in the respect of approximating human translation. Since the purpose of SSIX is to build a sentiment platform, *we consider the quality of translation is the best when there is minimal discrepancy in sentiment between the original texts and the translations*. Using our criterion, we need to explore the sentiment preservation. That is why we conducted Experiment 2.

5. Experiment 2

4.1 Experiment Design

For Experiment 2, we selected a subset of the previous sample ($N = 200$). We had to reduce the size of our sample because Experiment 2 required much more human resources than Experiment 1. To keep the time and expense cost under control, we chose a subset of the previous sample.

This experiment was designed to investigate whether translations (regardless of whether they came from human translators or machine engines) can maintain the sentiment from the original texts. As the first step, we recruited two German financial domain experts and they assigned sentiment to all four translations. The experts were kept away from the original English texts and their sentiment.

The sentiment scores assigned by the domain experts ranged from 1 to 10, 1 being the most negative, and 10 being the most positive. If the assigned pair of scores for a certain line of text diverged from each other for more than 2 points (including 2), we asked a third domain expert to evaluate the text again and chose the more appropriate sentiment score from the two alternatives.

For example, the human translator translated a certain tweet into German: *"Der miterlebte Fortschritt ist echt atemberaubend."* - *Stifel Analyst, nachdem er Teslas Fabrik zum vierten Mal gesehen hat \$TSLA* <https://t.co/nD7KECoM6V>

Its original English tweet is: *The progress witnessed is truly stunning.* - *Stifel analyst after seeing Tesla's factory for the fourth time \$TSLA* <https://t.co/nD7KECoM6V>

One of our domain experts assigned the German translation a sentiment score of 3, and the other assigned it a 10. Since there was a big gap between the two scores, the third domain expert evaluated the translation, and chose 10 from the pair of 3 and 10. As a result, the sentiment score for this tweet is 10.

4.2 Results and Discussions

After the data were evaluated and reconciled in the above way, we performed some statistical analysis on the results. We used a mixed linear regression model, which was implemented with the lmer4.0 package in R (Federico et al., 2014; Guzman et al., 2012). Compared with a linear regression model, a mixed effects model can explicitly model individual character-

istics. In our design, we used the item as a random intercept to capture the variance of each translated item to maximize the differences we could find between compared sets.

We are mainly concerned with the following two questions:

- Do human translations preserve sentiment?
- Does machine translation preserve sentiment?

To answer the first question, we need to compare the sentiment of the English gold standard corpus with the sentiment of human translation. If there was no significant difference between the sentiment scores of English gold standard and human translation, we would know the sentiment did not change too much; if a significant difference was found, then the sentiment is already lost in human translations.

After calculating our data set, results showed that there was no significant difference between the sentiment of English gold standard and human translation (Figure 1). In other words, the difference between gold standard sentiment (mean = 5.674) and human translation sentiment (mean = 5.536) was not large enough for us to draw the conclusion that they are different on a statistical level. This proves that human translation can preserve sentiment from the original texts. The results are what we desire to see because human translation is believed to be more reliable than machine translation. If human translation could not preserve sentiment, it is unlikely that machine translation can.

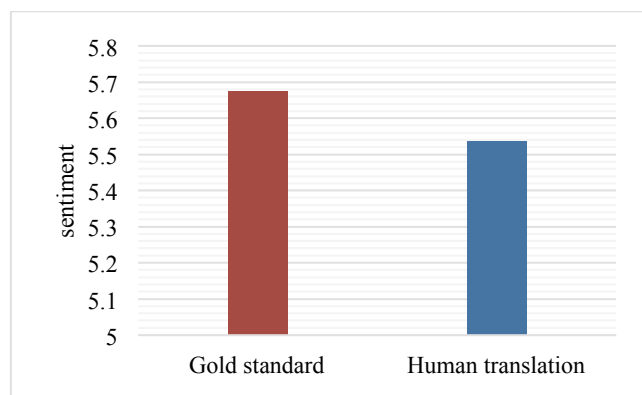


Figure 1. Sentiment Comparison: Gold standard vs. Human

Next, we try to answer the second question and assess the performance of machine translation engines on sentiment preservation. We compared the sentiment of the English gold standard with the sentiment of machine translations. Our results suggested that there were significant differences between the three pairs, i.e. English gold standard vs. Microsoft, English gold standard vs. Google, and English gold standard vs Google Neural Network (Table 2).

Engine	t-value	p-value
Microsoft	t = -3.574	p < .001
Google	t = 2.038	p < .05
Google Neural Network	t = 3.101	p < .01

Table 2. Results for Sentiment Comparison (Gold standard vs. Machine)

The visualization of the result can be found in Figure 2². Here Microsoft shows stronger diversion from the original sentiment in the gold standard, and Google produced the sentiment that was the closest to the original.

We also notice that compared to the gold standard sentiment mean, both human and machine translations have sentiment with lower means. At least two factors attribute to this fact. One is that translations have “neutralized” sentiment, drawing its mean closer to the grand mean (i.e. 5.5) because translations always lose information to an extent. The other is due to our domain experts. We used different groups of domain experts for annotating sentiment of English and German data, who are English and German native speakers respectively. Our German annotator could be more conservative or negative in assigning sentiment scores.

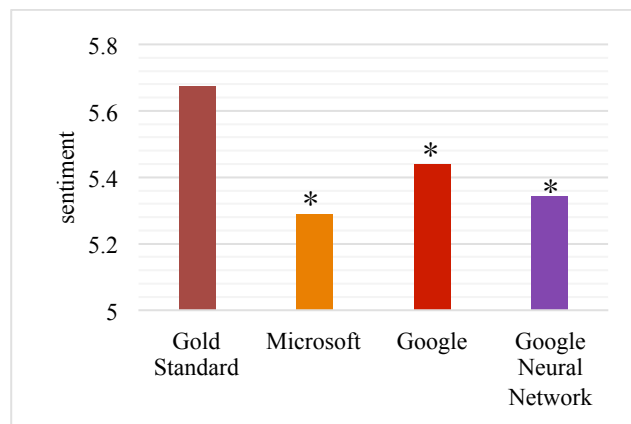


Figure 2. Sentiment Comparison: Gold standard vs. Machine

These results indicate that translations generated by machine engines are not of the desired high quality and look to be at risk of losing or distorting sentiment. However, they do not imply that machine translation is without merit. Since we have established that human translation is successful in preserving sentiment, we can use human translation as the benchmark to compare machine translations. If the sentiment assigned to a given machine translation engine does not deviate significantly from that of human translation, we can conclude that the engine has produced sentiment scores comparable to human translation.

The three comparisons discussed above showed that there are significant differences between the sentiment of human translation and Microsoft, which indicates that the Microsoft engine did not produce translations whose sentiment was alike to human translation (Table 3). The visualization is provided in Figure 3.

² The * on top of the bars indicated significance

Engine	t-value	p-value
Microsoft	t = -2.16	p < .05

Table 3. Results for Sentiment Comparison (Human vs. Machine)

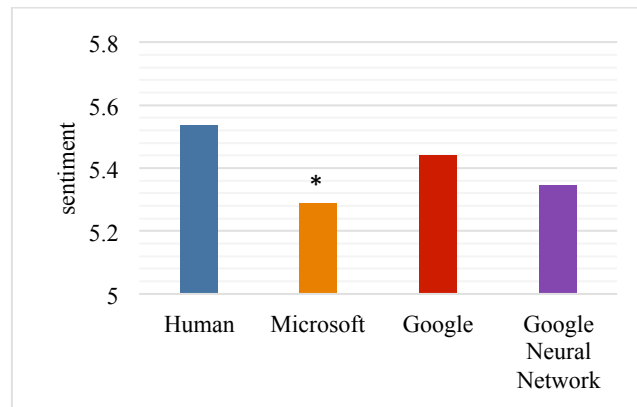


Figure 3. Sentiment Comparison: Human vs. Machine

Crucially, there was no significant difference between the sentiment scores of human translations and both Google and Google Neural Network. This means that the sentiment scores from Google and Google Neural Network does not differ significantly from human translation. This proves that these two engines' performance was in line with human performance, and consequently in these cases, sentiment can be considered as successfully preserved.

6. Conclusion

In this paper, we provide evidence that sentiment can be preserved after translation of an English gold standard corpus into German by machine engines, namely Google and Google Neural Network when they are integrated in GeoFluent. With this prerequisite fulfilled, we can either use the Derived approach to convert English data to another language and subsequently train a sentiment classifier on that data. Alternatively, we can use the Direct Translation approach to transfer multilingual data to English and use our already built English sentiment classifier. As these approaches do not need a human translator, time and costs can be greatly reduced, without an apparent, major loss in quality for the purposes of sentiment analysis. This is a crucial step for building an affordable multilingual sentiment platform in the domain of finance, to overcome the language barriers and help SME to analyze multilingual social content.

We have many directions for further research in the future that go from the integration of more language-specific processing rules in GeoFluent to enhancing the performance of machine translation, to benchmarking financial sentiment classifiers trained with Native and Derived approaches.

ACKNOWLEDGMENTS

This work is funded by the SSIX Horizon 2020 project (Grant agreement No 645425).³

References

Federico, M., Negri, M., Bentivogli, L., Turchi, M., & Kessler, F. F. B. (2014). Assessing the Impact of Translation Errors on Machine Translation Quality with Mixed-effects Models. In *EMNLP* (pp. 1643-1653).

GeoFluent (Lionbridge Inc.) <http://www.lionbridge.com/GeoFluent/>

Guzman, F., & Vogel, S. (2012). Understanding the Performance of Statistical MT Systems: A Linear Regression Framework. *Proceedings of COLING 2012*, 1029-1044.

Hürlimann M., Davis B., Cortis K., Freitas A., Handschuh S., Fernández S. (2016 September). *A Twitter Sentiment Gold Standard for the Brexit Referendum*. Paper presented at the Proceedings of the 12th International Conference on Semantic Systems, Leipzig, Germany.

Koehn, P., & Schroeder, J. (2007, June). Experiments in domain adaptation for statistical machine translation. In *Proceedings of the second workshop on statistical machine translation* (pp. 224-227). Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). 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.

Social Sentiment Index, 2015 – 2018, <https://ssix-project.eu/>

Source code for calculating the BLEU score:

http://www.nltk.org/_modules/nltk/translate/bleu_score.html

³ The sole responsibility of this publication lies with the author. The European Union is not responsible for any use that may be made of the information contained therein.

A New Methodology to Maximize the Strength of SMT and NMT

MT Summit XVI

Yu Gong
August 14th, 2017

vmware®

© 2014 VMware Inc. All rights reserved.

SMT vs. NMT

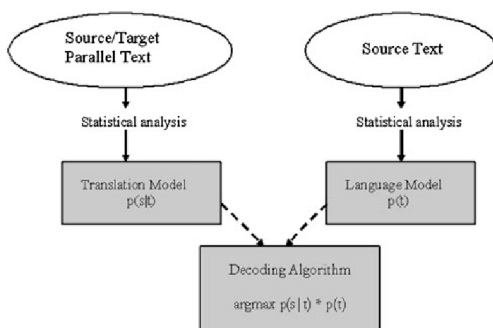


Figure-1

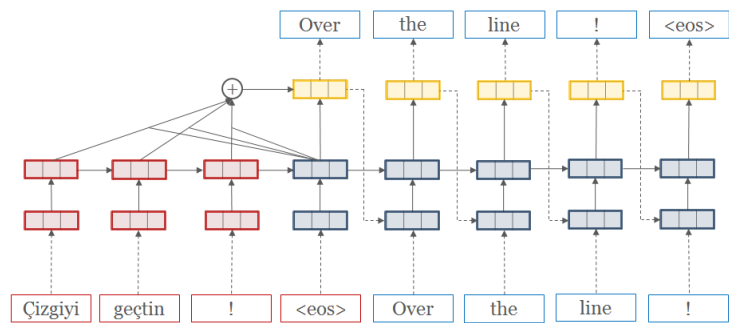


Figure-2

1. Karunesh Arora, Sunita Arora, Mukund Kumar Roy, [Speech to speech translation: a communication boon](#), 2013

2. <http://opennmt.net/>

vmware®

Which one is better?

- More and more attention to Neural MT
- Improved translation quality over SMT
- A milestone in machine translation



vmware®



Is it true?

The Data

- Selection of “real world” customer data collected over a three month period
- Catalogue of technical tools
- German → English
- ~ 5,000 Segments

Automatic Evaluation Results

	NMT	Moses
BLEU	23.68	47.98
METEOR	28.46	38.26

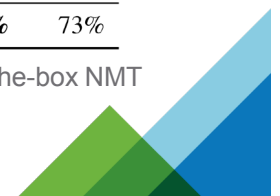
Table 1: BLEU and METEOR scores.

Manual Evaluation Results

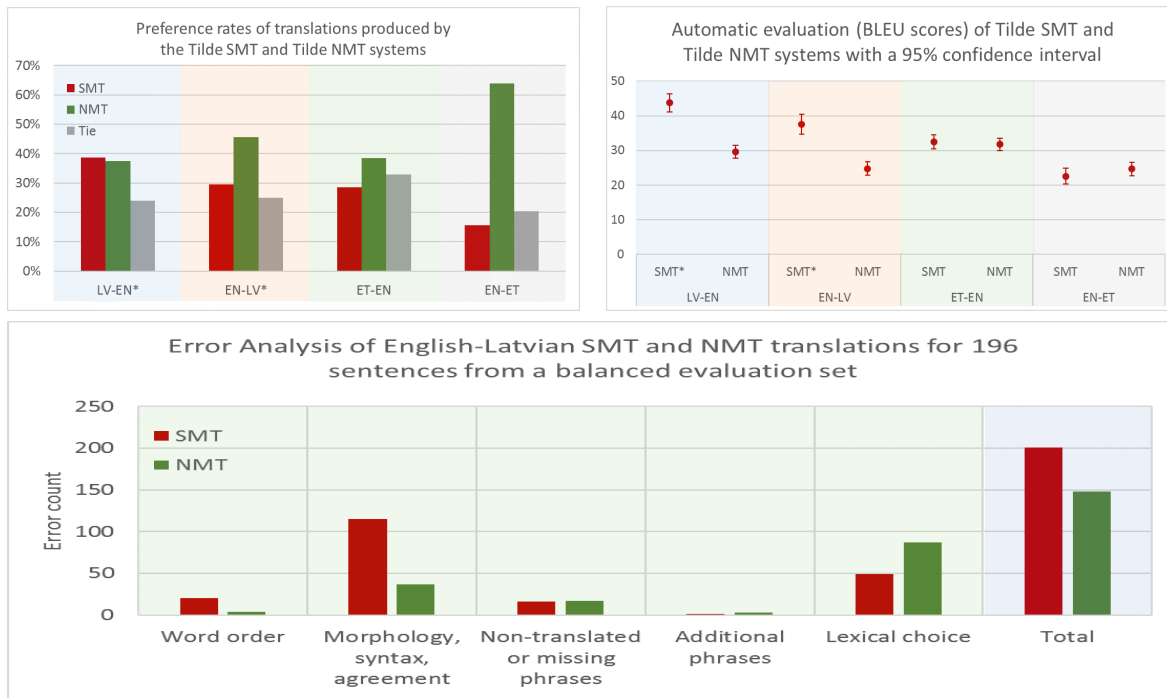
	#	NMT	Moses
formal address	138	90%	86%
genitive	114	92%	68%
modal construction	290	94%	75%
negation	101	93%	86%
passive voice	109	83%	40%
predicate adjective	122	81%	75%
prepositional phrase	104	81%	75%
terminology	330	35%	68%
tagging	145	83%	100%
sum	1453		
average		89%	73%

Anne Beyer, Vivien Macketanz, Aljoscha Burchardt and Philip Williams, Can out-of-the-box NMT Beat a Domain-trained Moses on Technical Data? EAMT 2017

vmware®



Another Example



vmware® <https://www.tilde.com/about/news/316>



Some Findings

- Professional translators prefer translations of NMT systems over translations of the SMT systems*
- NMT systems are better at handling word ordering and morphology, syntax and agreements (including long distance agreements) than the SMT systems*
- SMT systems are better at handling terminologies than the SMT systems
- Human comparative evaluation is crucial when comparing MT systems from fundamentally different approaches*

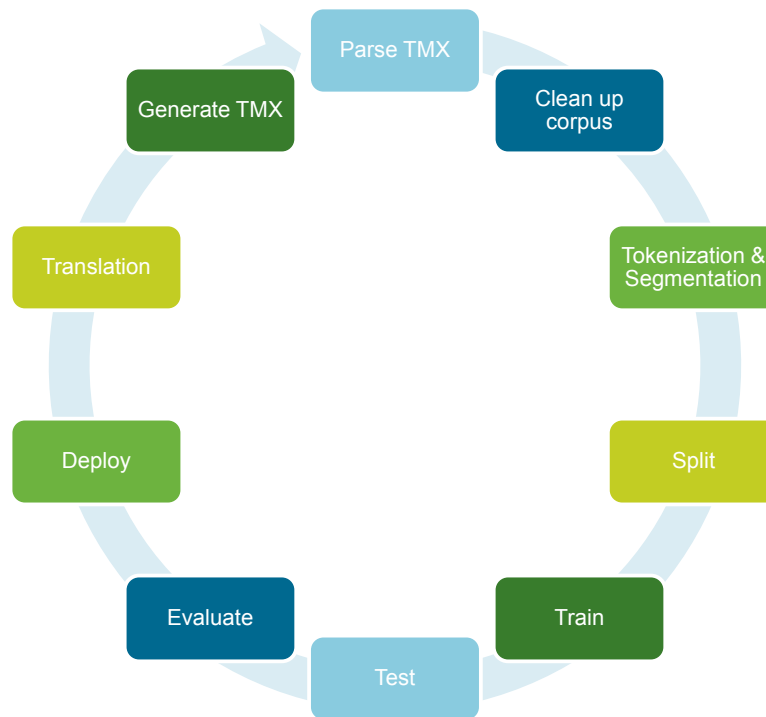
<https://www.tilde.com/about/news/316>
vmware®



What Can We Do?

vmware®

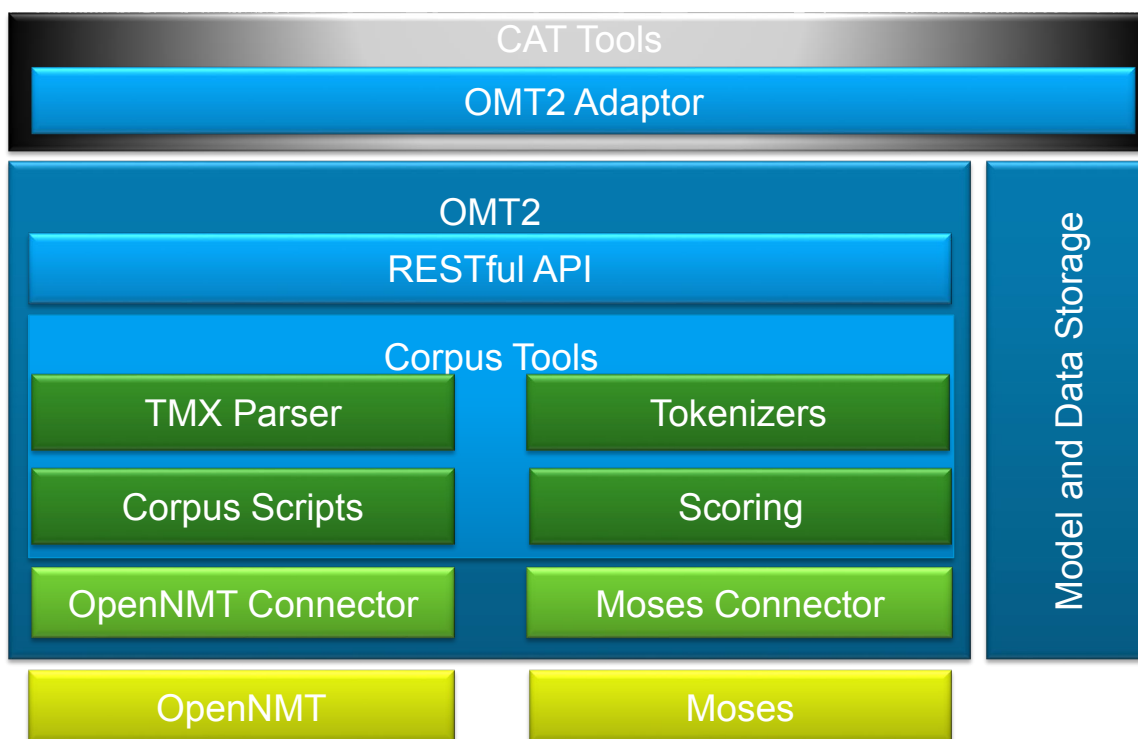
Open Machine Translation Toolset (OMT²)



vmware®

- Streamline the process of creating workable MT models
- Help users choose the best model by evaluating MT output
- Integrate machine translation into enterprise localization process
- Enable users to try the latest machine translation technology with least effort

Architecture



vmware®



OMT² Features

- Parse TMX: Extract corpus from Translation Memory eXchange (TMX).
- Clean up corpus: Remove garbage tags, those sentences length of which are not suitable for training a MT model.
- Tokenization & Segmentation: Call third party tools to tokenize or segment corpus.
- Split corpus: Split the original corpus by randomly selecting sentences for different purposes: training, validating and testing.
- Train: Train an MT model by calling OpenNMT or Moses scripts.
- Score: Use BLEU to give users a sense of how good the model is.
- Select the best model: automatically choose the model with highest BLEU score.
- Translation: Enable users to translate content by RESTful API.

vmware®



OMT² RESTful API

Request

https://translate.eng.vmware.com:5000/omt2/api/v1.0/getTranslation?src=en_US&tgt=zh_CN&str=hello

Response

```
{ "translation": { "SMT": "你好", "NMT": "你好" } }
```

vmware®



Sample Output in CAT Tools

1 Po the Panda is the laziest animals in all of the Valley of Peace, but unwittingly becomes the chosen one when enemies threaten their way of life.

SMT: 宝熊是和平谷中最懒惰的动物，但是当敌人威胁生活方式时，不知不觉地成为选择的动物。

NMT: 熊猫是所有和平谷中最懒的动物，但是当敌人威胁他们的生活方式时，它不知不觉地变成了一个被选中的人。

vmware®



Demo

vmware®



Q&A

vmware®



Thank You

gongy@vmware.com

vmware®



Rule-based MT and UTX Glossary Management – Honda's Case Dealing with Thousands of Technical Terms

MT Summit 2017 (Nagoya, Japan)

Saemi Hirayama

CAT tool leader, Honda R&D Americas, Inc.

Yuji Yamamoto

Founder/representative, CosmosHouse

Contents

- 1. Speakers**
- 2. Honda MT overview**
- 3. Issue 1: MT migration**
- 4. Issue 2: term inconsistency**
- 5. Terminology management continues**

Speakers

Saemi Hirayama

An in-house translator/CAT tool leader at the Ohio Center of Honda R&D Americas, Inc.

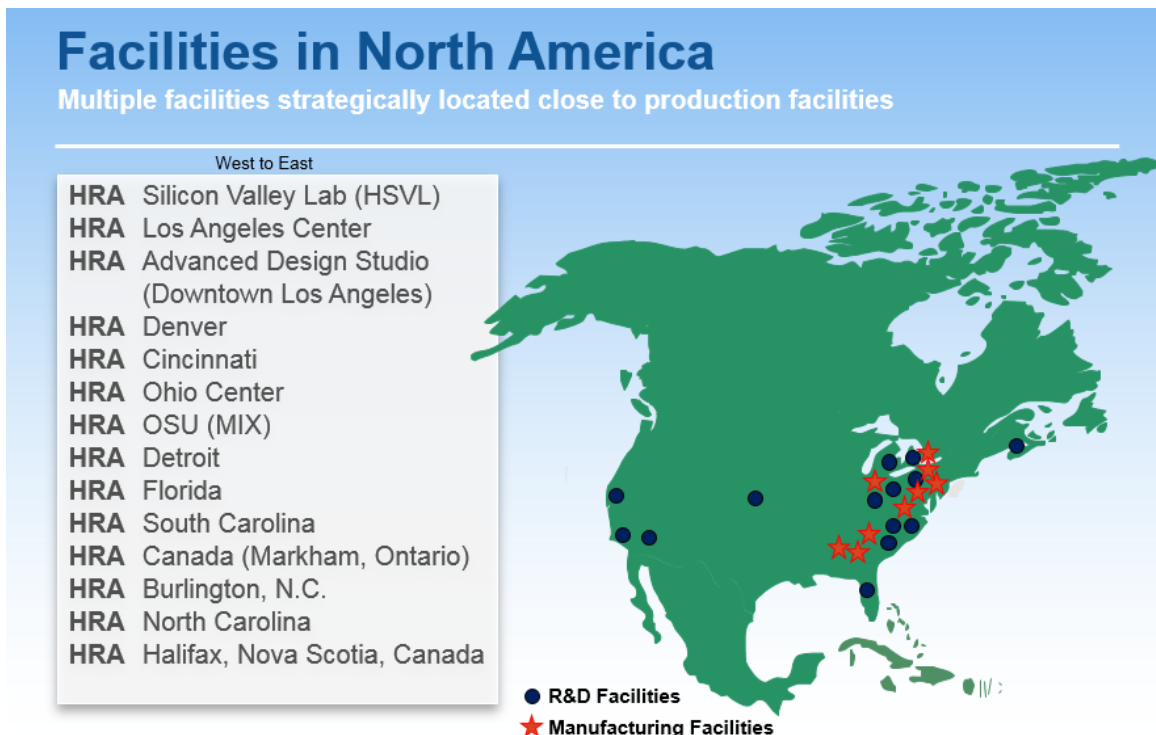
■ Yuji Yamamoto

Founder/representative, CosmosHouse

<<http://cosmoshouse.com>>

UTX team leader at AAMT (Asia-Pacific Association for Machine Translation)

Honda R&D Americas Inc. (hereafter referred to as Honda R&D)



**Honda R&D Americas
Inc.
Creating
New Value
in the U.S.**



- **Product Research & Development**
- **Product Styling Design**
- **Environmental Technology Development**
- **Safety Technology Research and Development**

Honda R&D's needs

- 1. JA to EN, EN to JA**
- 2. Technical documents written by engineers**
- 3. Used for translation needs by global associates in daily business operations**
- 4. Term-level accuracy and consistency are important**
- 5. Speed is crucial**

Honda R&D MT overview

Over a decade ago, Honda R&D adopted an RBMT (Rule-based Machine Translation) system

The current MT is a RBMT subsystem

Engineers use it to translate documents and emails

In-house translators also use it to process translation requests from engineers

Honda R&D MT overview

- Honda Jargon dictionaries categorized and added to the MT**
- A feedback function added to the web-based MT for mistranslations/ unregistered terms to keep the dictionaries up-to-date**

Honda R&D MT achievement

- **In-house translations reduced and outsourcing cost cut by half**
- **Significant translation speed increase**
- **Better communication with accurate technical terms**

Why was RBMT chosen at Honda R&D?

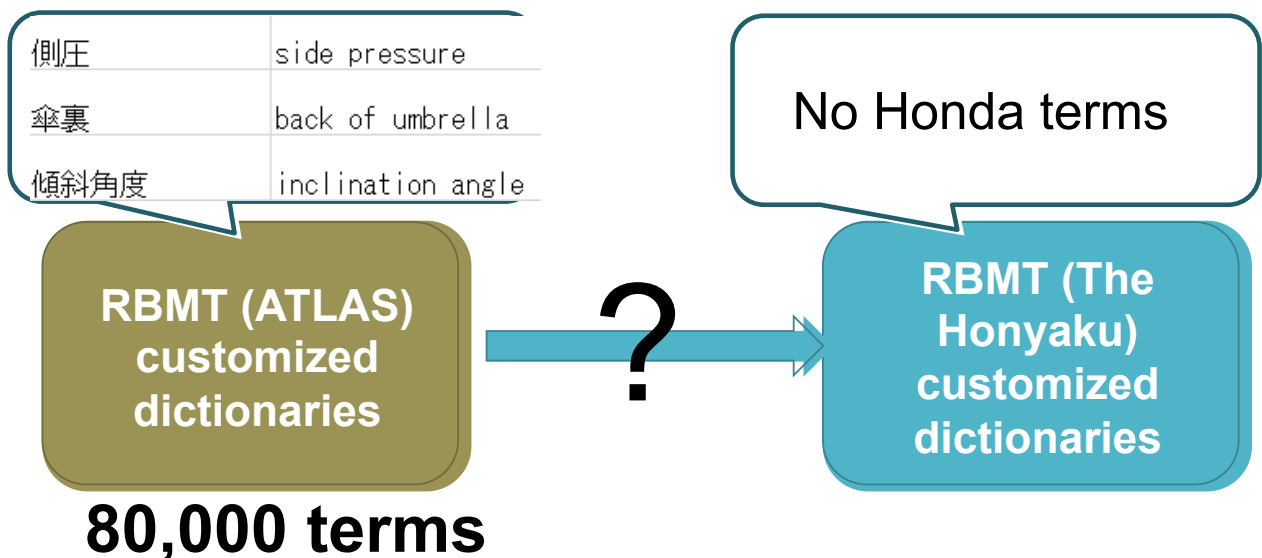
1. **80,000 terms, including many Honda-only terms**
2. **Many incomplete fragmental phrases/very few fixed phrases**
3. **File formats: complicated slides**
4. **No two documents are alike**

Why is neural/statistical MT not used at Honda R&D?

1. Term-level accuracy and consistency are poor in NMT
2. Human-translated corpus is too small
Because the majority of translations are lightly post-edited machine translations
3. Protection of intellectual property and secrecy
4. Higher cost
5. Most documents do not repeat

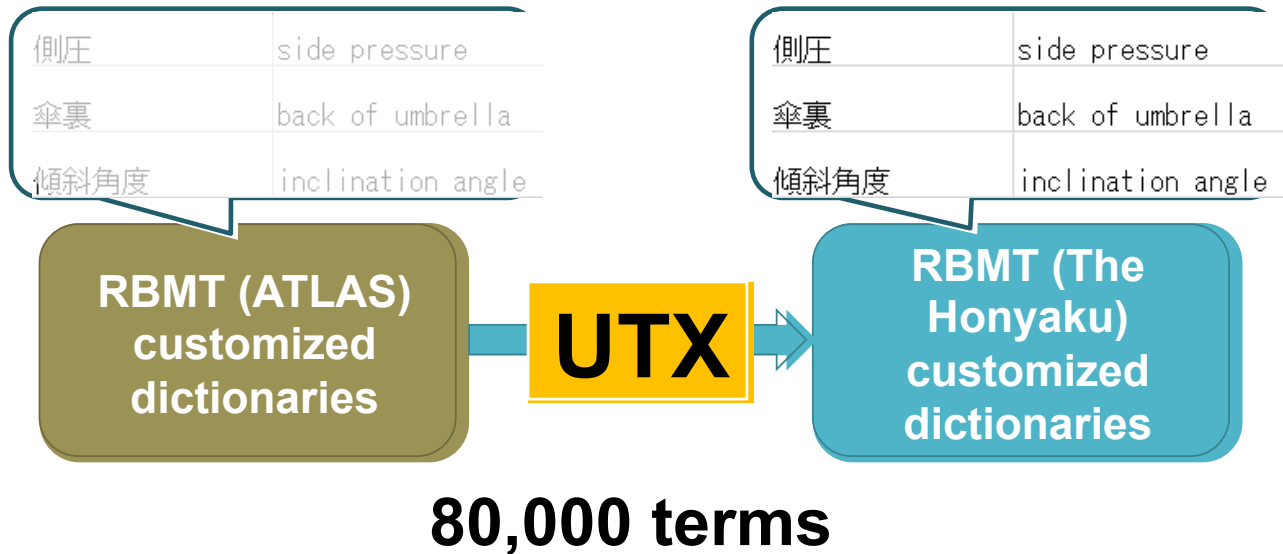
Issue 1: MT migration (RBMT to RBMT)

80,000 Honda terms in Fujitsu ATLAS were needed to be imported into a new MT system, Toshiba's The Honyaku.



Solution 1: MT migration (RBMT to RBMT)

Solution: Conversion through UTX format. The customized dictionaries were transferred to the new MT.



UTX – glossary standard for sharing and reuse

UTX (Universal Terminology eXchange)

- Developed by AAMT, initially as RBMT dictionary data exchange format
- Later restructured as a structured glossary format
- Used by companies and organizations such as Japan Patent Office

<http://www.aamt.info/english/utx/>

Issue 2: term inconsistency

Identical Honda jargon was being translated inconsistently at various company sites around the world.

整合会



Correlation meeting?
Coordination meeting?
.....?

“整合会”

“correlation meeting”

“arrangement meeting”

“collaboration meeting”



“coordination meeting”

“adjustment meeting”

“alignment meeting”

Honda Terms were being translated inconsistently in 6 Regions Worldwide

Solution 2: term inconsistency

- Term statuses (approved, non-standard, forbidden etc.) were added.
- 1:n, n:1, n:n source/target term pair relationships are clearly defined.
- J to E glossary now also works as E to J.

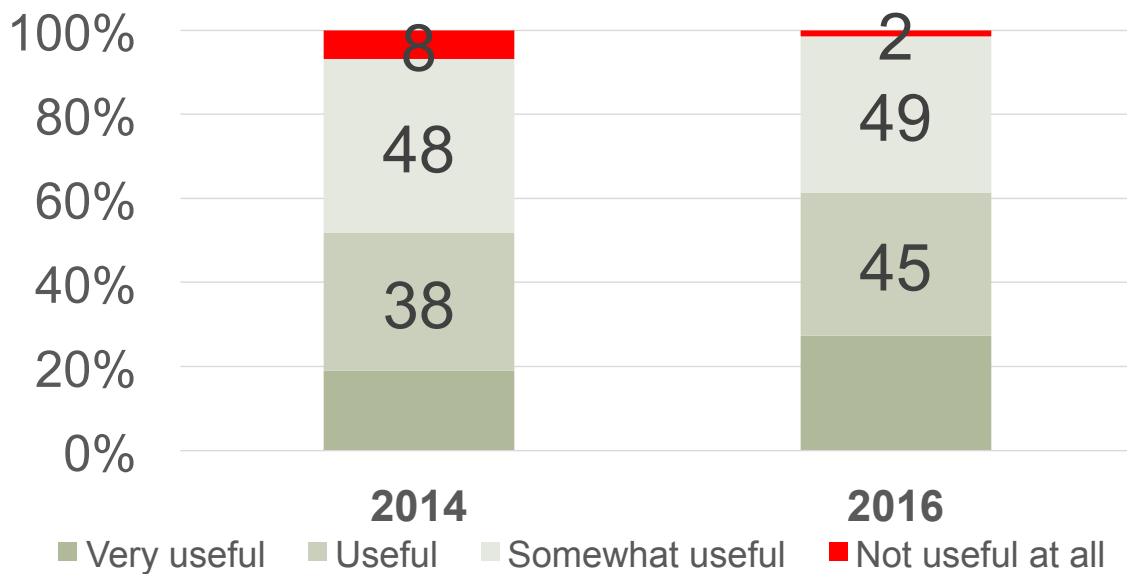
#term:ja	term:en	term status:ja	term status:en
整合会	correlation meeting	approved	approved
整合会	coordination meeting		forbidden
整合会	collaboration meeting		non-standard

Terminology management at Honda R&D

- 2013: UTX was introduced, transferring 80,000 terms from the old MT to a new one.
- 2014: a terminology committee was established to review existing/new terms to update the MT dictionaries monthly.

Reported useful by 98% of users

196 respondents: local US staff 80%, Japanese staff 20%



Terminology management continues

1. Review terms and term statuses
 2. Add new terms
 3. Delete unnecessary/obsolete terms
 4. Categorize terms
- ...to improve translation accuracy and efficiency

For fellow MT user companies

- **Glossaries control your company vocabulary**
Quality of human/machine translation can be improved with terminology management
- **Proper terminology management pays off!**

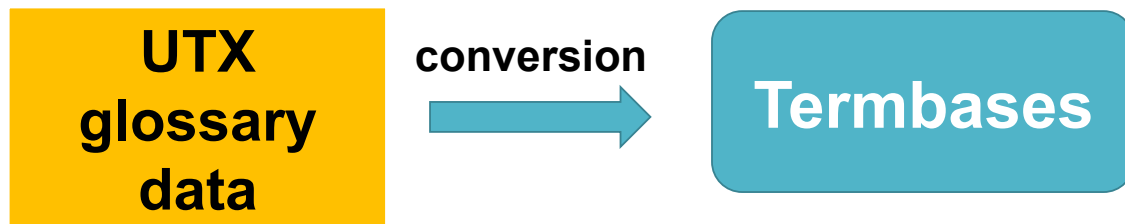
Solution 2: term inconsistency

- **Term statuses (approved, non-standard, forbidden etc.) were added.**
- **1:n, n:1, n:n source/target term pair relationships are clearly defined.**
- **J to E glossary now also works as E to J.**

#term:ja	term:en	term status:ja	term status:en
整合会	correlation meeting	approved	approved
整合会	coordination meeting		forbidden
整合会	collaboration meeting		non-standard

Future actions at Honda R&D

1. Further tuning of UTX glossary
2. UTX for terminology check
 - Can be used for post editing neural/statistical MT if necessary
 - Terminology tool training for translators



Take away

1. Glossaries are necessary for both humans and MT
2. UTX glossary management has been effective at Honda R&D
3. Neural MT may not be the only future – users are satisfied with RBMT

A detailed investigation of Bias Errors in Post-editing of MT output

Silvio Picinini

Localization, eBay Inc., San Jose, CA, USA

spicinini@ebay.com

Nicola Ueffing

Machine Translation Science Lab, eBay Inc., Kasernenstr. 25, Aachen, Germany

nueffing@ebay.com

Abstract

The use of post-editing of machine translation output is increasing throughout the language technology community. In this work, we investigate whether the MT system influences the human translator, thereby introducing "bias" and potentially leading to errors in the post-editing. We analyze how often a translator accepts an incorrect suggestion from the MT system and determine different types of bias errors. We carry out quantitative analysis on translations of eCommerce data from English into Portuguese, consisting of 713 segments with about 15k words. We observed a higher-than-expected number of bias errors, about 18 bias errors per 1,000 words. Among the most frequent types of bias error we observed ambiguous modifiers, terminology errors, polysemy, and omissions. The goal of this work is to provide quantitative data about bias errors in post-editing that help indicate the existence of bias. We explore some ideas on how to automate the finding of these error patterns and facilitate the quality assurance of post-editing.

1. Introduction

The use of machine translation (MT) for facilitating the work of translators is increasing throughout the language technology community. The human translator receives an automatically generated translation from the system, and then corrects the errors made by the system. This is called post-editing. As post-editing will gain even more importance, we believe that the quality of this work needs to be evaluated. Translations suggested by MT systems contain errors, and - for several reasons, such as time pressure - the posteditor might leave these MT errors uncorrected. We are calling this effect "bias", as in the posteditor being "biased" by the MT suggestion, and accepting translation errors.

In our work, we investigated whether the MT system influences the human translator, thereby introducing bias and potentially leading to errors in the post-editing. We analyzed how often a translator accepts an incorrect suggestion from the MT system. Furthermore, we explored the types of bias errors and performed a quantitative analysis.

Our analysis was carried out on translations of eCommerce data from English into Portuguese, consisting of 713 segments with about 15k words. In addition to the MT output and the post-editing, we carefully curated a golden post-editing reference. Using this golden reference, we calculated edit distances and related scores, and then classified and quantified the types of errors that emerged. We observed a higher-than-expected number of bias errors, about 18 bias errors per 1,000 words. Among the most frequent types of bias error we observed ambiguous modifiers, terminology errors, polysemy, and omissions.

The goal of this work is to provide quantitative data about bias errors in post-editing that helps indicate its existence. Additionally, we will provide data about certain types of error patterns that lead to bias. We explore some ideas on how to systematically find these error patterns

and facilitate the quality assurance of post-editing. Educating post-editors about bias and about these patterns can help improve the quality of the post-editing work, and therefore the final translation quality delivered to the user.

An early analysis of post-editing of machine translation output is presented in (Krings, 2001). This publication discusses the post-editing process and the quality of machine translations and post-editing, but does not have a quantitative analysis of errors. More recently, (Blain et al., 2011) presents a qualitative analysis of post-editing, focusing on reducing the post-editing effort. In addition to this analysis, the authors present methods for learning corrections from post-editings and improving the MT systems which generated the translations.

2. Analysis of Bias Errors

2.1. Data

We worked on translations from English into Portuguese in the eCommerce domain. The text are descriptions of items which are for sale on the eBay site. The English descriptions, consisting of 713 segments with 15k words in total, were automatically translated using the Microsoft statistical machine translation system, and were post-edited by a human translator, whom we will call post-editor 1 going forward. These post-editings were carefully reviewed by another language expert, whom we will call post-editor 2, who created perfect translations to be used as golden references.

2.2. Methodology

We performed a detailed manual analysis of the post-editings from post-editor 1, comparing them against the golden reference from post-editor 2, in order to detect bias errors. For each error corrected by post-editor 2, we analyzed source, machine translation, and post-editing for potential bias. We classified the errors into certain groups which will be described in section 3.

We used edit distance (Word Error Rate – WER) in two significant ways. First, the distance calculated between the **machine translation and the post-editing**. This is an indication of where post-editing happened and how much. Based on those data, we developed a process (described in a section below) to identify instances of lack of post-editing:

- If the post-editor does not post-edit a segment (for example, by skipping it), the edit distance is zero. This could look like all MT errors were accepted and there was bias, but in reality the posteditor just missed the entire segment. We wanted to find and exclude these instances from the bias analysis.
- If the post-editor rushes through the task and make just one change in a segment, and there were others to make, this will result in a low edit distance. This would look like bias when it is not bias, it is just lack of proper post-editing. We also wanted to find these instances and exclude them.

Second, we used the edit distance between the **golden reference and the post-editing**. The primary use of it was to triage the segments to be analyzed. If the edit distance was zero, this meant that the golden reference agreed in full with the post-editing, so this segment should not be part of the analysis.

The edit distance between post-editing and golden reference can indicate:

- If the edit distance is low, this is an indication that the post-editing was generally good and not many changes were needed.
- If the edit distance is significant:
 - There could be a lack of knowledge – the post-editing made changes and they were wrong, so the golden reference corrected this. This

could appear as high PE-MT distance and also high Golden-PE distance.

- There could be bias – the post-editing accepted the MT and the golden reference changed it. This could appear as lower PE-MT distance and higher Golden-PE distance.

We looked into the numbers for the edit distance through the WER scores, see [Table 1](#). The results are consistent with our expectations: The PE-Golden is higher with bias than without it, which means that there were more corrections for bias segments, as expected. The PE-MT is slightly lower with bias compared to without it, which means that there were fewer changes by post-editors in segments with bias, and therefore they left more errors in them.

avg. WER	All	without bias	with bias
PE vs. golden	0.12	0.09	0.20
PE vs. MT	0.25	0.26	0.21

Table 1. Average WER of post-editing (PE) vs. golden reference and vs. MT output

Finding and excluding content with lack of post-editing

The bias that we are trying to identify happens when the post-editor looks at the machine translation and makes a conscious decision to accept it, and the machine translation is wrong. However, it could happen that the post-editor would skip working on a segment, or could make one change at the beginning of a segment and leave the rest untouched. These would not be example of bias, they would be examples of lack of complete post-editing. In order to try to identify this phenomenon, and exclude it from our analysis of bias, we went through the steps described below.

1. Generated WER scores for each segment, between the post-edited version and the initial MT version (PE-MT).
2. With numbers for each segment, we plotted these numbers on a chart (Figure 1):

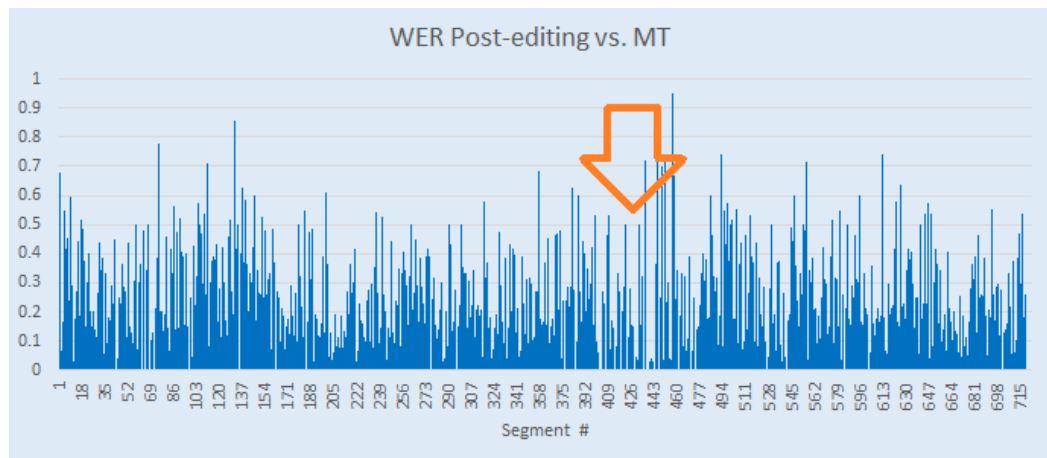


Figure 1. Segment-level WER Post-editing vs. MT

This chart shows regions of data where the volume of post-editing seems lower than the rest of the chart. Further investigation showed that the post-editor indeed failed to do a complete post-editing on segments in this region.

3. We looked for a different chart display that would make this phenomenon more visible than plotting the scores. Therefore, we calculated the average of the WER for the past 30 segments, and plotted this rolling average of distances (shown in Figure 2). The orange line is the average for the file.

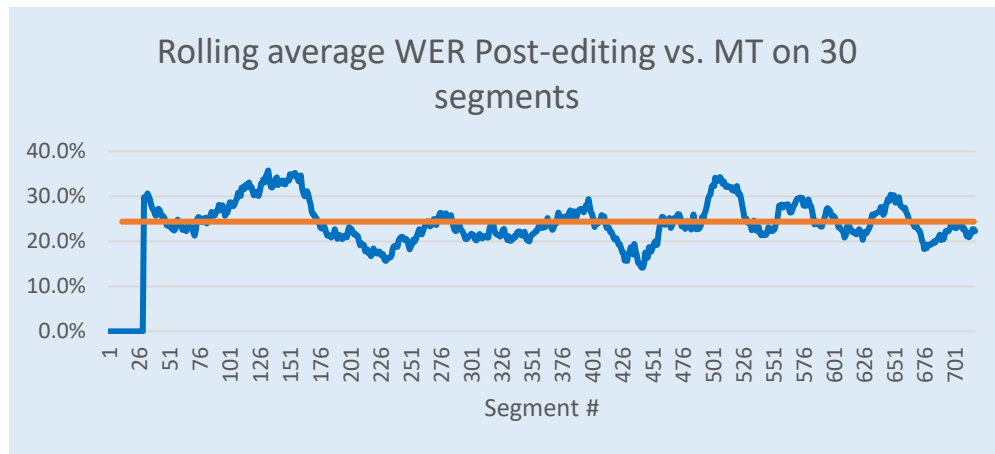


Figure 2. Rolling average WER Post-editing vs. MT on 30 segments

This type of chart shows the amount of post-editing effort progressing through the file. If the post-editor, for example, rushes the work towards the end of the file because of a deadline, this will be reflected in a lower WER/edit distance in a series of segments in that region of the file. This lowering will appear in the chart, as the rolling average will go down for that region. This visualization showed two regions of interest (where the chart shows the lowest values), one region around segment number 221 and the other region around segment number 441. After investigating these regions, we confirmed that the second one had segments lacking post-editing.

4. We looked for one more type of chart and we plotted the “Rolling % of zero-WER in 30 segments” and “Rolling % of low WER (<4%) in 30 segments”. In this chart (shown in Figure 3) we tracked the % of zeros in the past 30 segments. A concentration of segments with no post-editing would start to increase the percentage as we move through them, so regions in this chart with peaks are our regions of interest. We did the same for “% of lows” (shown in Figure 4), tracking not only zero changes but also low % of changes up to 4 %. These are segments that could have changed one character, for example.

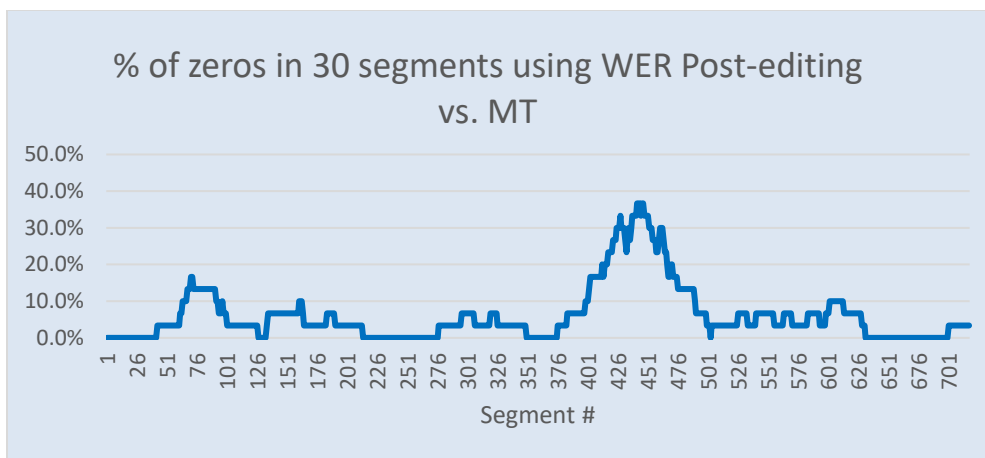


Figure 3. Percentage of zeros in 30 segments using WER Post-editing vs. MT

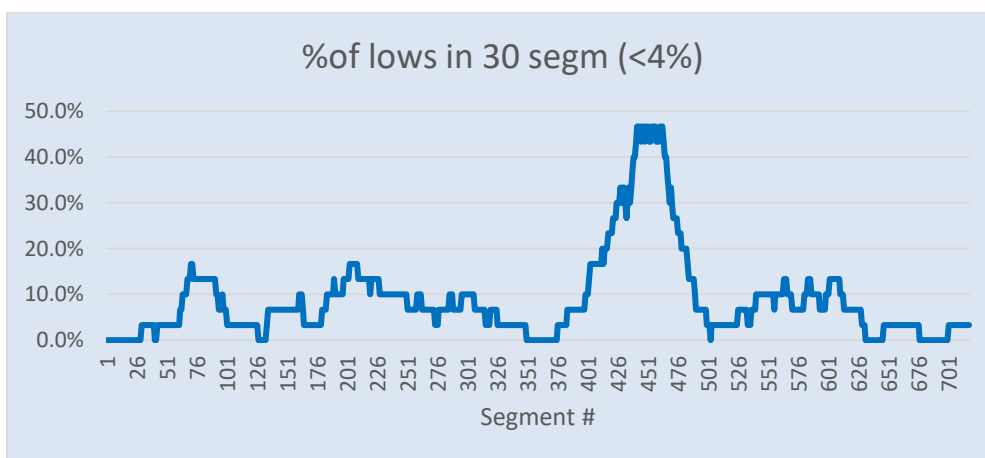


Figure 4. Percentage of low WER (<4%) in 30 segments

Both charts were very effective in pointing out regions of interest (highest values on the chart, around segment # 441 as before. While at first it may seem counter-intuitive to look for the highest numbers when talking about low edit distance, it takes just a few seconds to realize that we are looking for “high concentrations” of low scores, and then the peaks on the charts make sense.

2.3. Findings

General observations

“Is there a significant volume of bias?”, that was the question that we wanted to explore for this particular case. While a “Yes” answer can’t be easily generalized to other cases, we hope that there is value in concluding that (1) bias happens and (2) this is an issue that needs attention when thinking about improving the post-editing quality.

Types of errors/causes found

Our analysis did not start with a defined set of standard error types. Instead, as error patterns emerged, they became a type. We are used to error typologies, but the classification used in this work is trying to look at **causes of MT errors**. Some of these descriptions below will look more clearly like a cause, such as “Modifiers to Multiple Words” or “Multiword expressions” and others may look like a traditional error type, but there is still a cause behind it. Whatever the causes are, we should just keep in mind that these causes created an MT error, and then bias occurred when that MT error was not changed.

- Multiple Modifiers or Words (MMoW) – this pattern describes situations where a modifier may or may not apply to several words around it. This ambiguity is difficult for the MT to resolve. This more frequently applies to nouns and adjectives, but we opted for a more general name because there are some examples of those, and the same principle applies. Examples of this situation are shown in Figure 5 and 6:



Veteran musicians, DJs, and public speakers
are taken aback by the sensitivity and reliability of the Samson QMIC.

Figure 5. Example 1 of Multiple Modifiers or Words

We as humans intuitively know that this is talking about veteran musicians but also veteran DJs and veteran public speakers. However, the MT engine does not know that, and will produce a translation that says, “DJs, public speakers and veteran musicians are taken...”, and the cause of the error is a modifier adjective that applies to multiple nouns. Another example:

The hypercardioid capsule is built to match frequency and sensitivity of
lectern, choir, and boundary mics.




Figure 6. Example 2 of Multiple Modifiers or Words

In this situation, we have three modifiers applied to one word. We as humans use the context to understand that mics (microphones) probably have frequency and sensitivity and therefore lectern, choir and boundary are three different types of microphones. So this sentence actually means “...lectern mics, choir mics and boundary mics.” The MT does not know that and will produce a translation that sounds like “... sensitivity of boundary mics, and choir and lectern.”, and the cause of the error is multiple modifier (you can see them as nouns or adjectives) that apply to one noun.

This pattern appeared a significant number of times and tends to be difficult for MT. We decided to explore further this pattern in two ways, explained later in this paper:

- Can we find this pattern more systematically?
- Does this issue occur also for Neural MT?
- Multiword expressions (MWE) – these are issues where a sequence of words has a completely different meaning than the individual words. Examples of this pattern are idioms and phrasal verbs. It is a difficult construction for the MT to handle because of the change in meaning, so it is a cause of errors made by MT.

Examples include “makes an impression”, “cut short”, “built in”, “turn over to”.

- Polysemy - Polysemous words are words with multiple meanings and therefore multiple translations. In our case, we look at all issues related to polysemous words that have two competing meanings that are popular in the corpora and confuse the MT engine. This is a cause of errors for MT.

Consider, for example, “a choice of restaurants to eat”. The more common meaning of “choice” is probably “to make a choice” but in this example, you are not actually making a choice, and instead the meaning is “variety of options”. If this meaning is less common in the corpora, the MT may make an error. In “performance-conscious photographer”, “conscious” means “photographer concerned with the performance”, but it was translated literally as “did not lose conscience”. So the translation ended up sounding like “performance-did-not-pass-out photographer”.

Other examples include:

- In “Enter a new world of creativity”, “enter” was translated as “insert” as in “enter a password” instead of “walk into” a new world.
- “fleece-lined compartments” had “lined” translated as “aligned” instead of “covered with fleece”
- In “Publishes...materials of benefit to the bar”, “bar” refers to lawyers and was translated as the place to go for drinks.
- In “Washer...including cycles for active wear”, “wear” refers to clothing and was translated with the meaning as in “wear and tear”.
- Mistranslation - In general, “Mistranslation” represents causes that made the MT engine produce a mistranslation. However, every error can be considered a mistranslation. In this work, we classified all possible issues into specific categories. The issues left to be classified as mistranslation are the ones where the translation is wrong but the cause can’t be easily identified. The example of “parent and child” translated as “father and son” should illustrate this category well. We don’t know exactly why the translation is wrong, we only know that it is. This goes into a “Mistranslation” category.

Other Mistranslation examples include:

- “allow concentration to be focused elsewhere” had “elsewhere” translated literally as a location, when “elsewhere” here means focused on “something else”
- “reduces eye fatigue and neck pain” had “neck pain” translated as “throat pain”
- “overcooking” translated as “burned meals”
- Do Not Translate terms – brands and other terms that should not be translated are a cause of errors for MT when the engine has to decide if the term is a brand or a common word. Generic examples would be brands like Gap, Guess or Coach. In our case, examples include the brand Philosophy and a product name called JBL Venue Stadium.
- Terminology – this cause of errors appears when the MT does not know the proper terminology for a certain subject matter. Examples include “focal length” for cameras, “devices” for heraldry, “refrigerator” and “green gas”.

- Part of Speech (POS) –we were interested in this specific cause of error, when the MT would translate a word using the wrong POS for it. Some examples we found showed significant ambiguity that would cause MT errors, some of them difficult even for humans to resolve. Examples include:
 - “Nuts & Bolts component utilities include...” where “component” is an adjective meaning “utilities that compose the Nuts and Bolts...”. The translation treated it as a noun.
 - “The dual apertures of the Vivitar MC Macro Focusing Zoom allow for more flexibility in varying light”, where varying is an adjective meaning “light that can vary”. The translation meant “... more flexibility in the ability to vary light”, treating “varying” as a verb.
 - “One example was gilt -- a process presumably done after striking...”, where gilt is a noun (the action named “gilt”) and it as translated as the adjective “gilt”. In sentences where the structure is “<subject> was xxxx”, the xxxx is usually an adjective, but in our example, it was not.
- Omission of the initial article – it is a common style in English to omit an article at the beginning of a sentence. Examples with and without article include:
 - "AutoCAD LT 2D CAD design software simplifies tasks" vs. “*The* AutoCAD LT 2D CAD design software simplifies tasks"
 - "Zeus IOPS eliminates the wait time" vs. “*The* Zeus IOPS eliminates the wait time"
 - "Familiar six-button configuration provide direct access" vs. “*The / A* Familiar six-button configuration provide direct access"
 - "CenterFlex technology helps enable [...]" vs. “*The* CenterFlex technology helps enable [...]"

While readers of English are used to this construction, the MT notices that pattern and consistently produces translations that miss the initial article in several languages. The impact of that is very different from English because readers of those languages are used to the article being present virtually all the time. This is a systematic cause of MT inadequacy. The bias in post-editing consists of not adding the initial article on the target language.

- Untranslated words – we found instances of words that should have been translated and were not. Examples include: “POI” (acronym for Points of Interest), "non-resonant", "adaptogen", "dot inlay". These issues may be linked to these words being out of vocabulary.
- Omission – we tracked omissions made by MT and not corrected by post-editing. Examples include: card in “card printers”, sharp.
- Addition – same as omission for words added by MT and not removed. Example: added “obtain”.
- Prepositions – significant changes of meaning can be caused by a preposition. A different preposition or its omission in the translation may have an impact. Example:
 - In “save consumer’s money by reducing the operating costs”, the preposition “by” is what defines the meaning as “the reduction of operating costs is what will save consumers money”. The translation sounded as “save consumers money, reducing the operating costs” and actually reversed the meaning as if “save consumers money” would “reduce the operating costs”.

- “The sole in the Callaway Wedge” was translated as “The sole in wedge shape of the Callaway” changing the meaning just with one preposition.
- Gender agreement – we tracked when the MT made the wrong agreement and was not corrected
- Number agreement – same as gender for singular/plural
- Word order – MT created wrong word orders and they were not corrected.
- Grammar - MT created wrong word orders and they were not corrected.
- Verb tense – MT used the wrong tense for a verb and this was not corrected. Examples include:
 - In “Getting a camera with a greater number of megapixels means cropping and enlarging won't adversely affect picture quality”, the gerund “getting” actually does not mean that you are actually already getting the camera at the moment. This would be the meaning for the gerund. Instead, it means the hypothetical action of the infinitive, something like “To get a camera... means cropping... won't affect...”. This infinitive is the tense that is required to appear in the translation. The MT will translate as a gerund and the post-editing needs to change to an infinitive. If this does not happen, there is a bias.
 - English has almost no difference between the subjunctive mode and the indicative. The construction “If I were invisible” instead of “If I was invisible” may be the most visible instance of differences in the mode. Yet, there is a popular song that uses the “was” form. This similarity will cause the MT to make errors translating subjunctives as indicatives. If this is not corrected by post-editing, there is a bias. Examples include:
 - “so that they can be lifted” in “the rockets are fitted with magnets so that they can be lifted and loaded with cranes”
 - “(would) freeze herself” and “would cause” in “It seems as if Hilda, while trying to scare up a dancing partner, accidentally freezes herself and causes objects to fly throughout the house”
 - “to be found” in “Usually the puzzles require items to be found and then executed”. It means “the puzzle requires that items be found”, and not “requires items that will be found”.
 - Spelling (including language rules) – Brazilian Portuguese had a spelling reform. Therefore, there are new language rules in place. The corpora used to train the MT engine contains content created before the reform. Therefore, there are spelling errors training the MT, and they will appear in the translation. If they are not corrected, there will be bias.

Spelling reforms and corpora for MT will pose a certain challenge for MT systems. Many languages use corpora from different flavors of the language, such as European Portuguese and Spanish versus those used for Brazil and Latin America. The spelling of Portuguese varies a little bit between Brazil and Portugal, so the MT ends up making a few Portugal suggestions for Brazil and Spain suggestions for Mexico or Colombia. The advantages of having more corpora by doing this outweigh the downsides of it. If the corpora will not be fixed, the role of post-editing will include correcting these “cross-border” spelling issues.

Errors per 1k words

The main number that we obtained was the number of bias errors per 1k words. Although there is no official standard for the industry, we typically consider translations as high quality if the

number of errors per 1,000 words does not exceed 2, based on our own personal experience. The total number that we found in this work is given in Table 2.

Total Number of words	14,986
Total Number of bias errors	270
Errors per 1k words	18.02

Table 2. Number of bias error vs. number of words in post-editing

This number is about nine times a reference for quality, indicating that the total number of bias errors is significant. We should keep in mind that these are only bias errors. In addition to these, there are regular non-bias errors, where the post-editor makes changes and they are still not correct. The entire picture of quality is comprised of bias + non-bias errors.

Numbers for each type of error

	Polysemy	Mis-translation	Multiple Modifiers or Words	Multi-word Expressions	Omission initial article	Do Not Translate terms	Untranslated	Omission	Addition
Number of errors	54	56	22	14	10	9	15	16	4
Errors per 1k words	3.60	3.74	1.47	0.93	0.67	0.60	1.00	1.07	0.27

	Terminology	Gender agreement	Number agreement	Prepositions	Word order	Spelling (incl Lang Rules)	Grammar Verb tense	Part-of-Speech	Total
Number of errors	14	7	12	8	8	4	9	8	270
Errors per 1k words	0.93	0.47	0.80	0.53	0.53	0.27	0.60	0.53	18.02

Table 3. Numbers for each type of error

Table 3 lists the frequency of each of the error types which we defined during our analysis. The breakdown per type shows that several types of causes of errors (described previously) are deserving of further attention:

- Polysemous words
- Multiple modifiers or words
- Multiword expressions
- Terminology
- Omissions
- Untranslated words

- Other causes of mistranslation

Detailed Analysis for Errors caused by Multiple Modifiers or Words (MMoW)

We analyzed the error caused by Multiple Modifiers or Words in more detail. We found 22 instances of this error, indicating about 1.5 errors per 1k words. This is a significant number for just one type of error.

Next, we were interested in the question whether this error occurred with the same frequency for different types of MT systems. Therefore, we compared the output from 2 different types of MT systems for these 22 segments to find out whether these errors stem from inherent complexity of the source segment. We used this issue to have a sense of the impact that Neural Machine Translation may have on the quality. The hypothesis is that if NMT produces an output that contains less causes of errors, there will be less errors and therefore less bias. We wanted to see if this happened.

We looked into the 22 issues identified originally on Microsoft Statistical MT and created a NMT output from Microsoft Neural MT for them. We then evaluated to see how many of the original 22 errors were still present in the NMT output. We found that 10 out of 22 times, the NMT system corrected the error, meaning that it improved over the SMT system in 45% of the cases. However, in the remaining 10 segments, we still observed the same type of error in the NMT output. These results indicate that NMT tends to produce less errors than SMT for the post-editing corrections. This leads to better post-editing quality. However, 55% of the errors in the SMT output were still present in the NMT, indicating that the issues that a significant portion of the issues that are difficult for SMT remain an issue for NMT. This seems to indicate that the work on patterns that we started here would be a worth pursuit in improving NMT, and in evaluating it.

Once we identify that Multiple Modifiers or Words was an issue worth our attention, we thought of how we could find these expressions in a more semi-automated way. The process that we used can be described in these general steps:

1. Run a POS tagging of the source content
2. List tokens and tags and simplify the POS tags to a minimum; see Table 4 for examples. Create patterns indicating errors and find these patterns in the tagged content

We applied a simple formula to find a pattern: adjective-noun-noun and found the example “enhanced telephony capability” above. We also looked for another pattern: adjective or noun-noun-“and”-noun. We found examples such as “cook time and temperature” with this pattern. Analyzing the data, we found that:

- Using only these two narrow formulas we already found 7 out of 22 issues (32%). This indicates that a few formulas could find the majority of the patterns.

Token	Tag	Simplified Tag
the	DT,B-NP-plural	DT
enhanced	JJ,enhance/VBD,enhance/VBN,I-NP-plural	JJ
telephony	NN:U,I-NP-plural	NN
capability	NNS,E-NP-plural	NN

Table 4. Examples of POS tags and simplified tags for English tokens

- We manually analyzed the 22 MMoW issues to find out how many were suitable to be found with formulas/patterns. Out of 22, 20 of them could be found. This indicates that most of the MMoW issues are findable with patterns, and that there is potential to semi-automate the harvesting of these terms from a content tagged with POS.

3. Conclusions

1. We found significant bias in the post-editing of MT. This cannot be generalized to all cases, but it shows that the bias exists and is an issue to be considered as part of improving post-editing.
2. We found patterns that cause MT errors and can cause significant bias. These patterns should be considered for improvement of post-editing and for measurement of post-editing quality.
3. We found that it is possible to apply some automation in detecting the error patterns that cause errors on MT.
4. We found that Neural MT is likely to reduce the errors from bias by eliminating the original MT error. However, a significant percentage of the issues that cause errors on MT are not resolved by Neural MT and remain of interest for improving and measuring the quality of Neural MT.

4. Future Work

1. The semi-automated finding of patterns should be explored further. Once a representative number of instances of patterns is obtained, different metrics can be calculated. For example, we could find that there are 100 instances of MMoW in the content. If, upon reviewing them, we find, for example, 43 errors, this indicates that this type of error is produced by MT 43% of the time.
2. We think that there is potential in measuring the quality of the MT output based on difficult issues instead of a random sample. If a system 1 performs better than another system 2 on polysemous words, multiple modifiers or words, and multiword expression, it is likely that this system 1 will perform better on any translation than system 2. The same reasoning of measuring difficult words can be applied to measuring post-editing quality. We would like to create a measurement method that is not based on random sampling nor error typology, that targets difficult words, that is not subjective (make simple binary decisions), that is fast, cost-effective and suitable for crowdsourcing (with bilingual people and not professional linguists). We are working on this topic.

References

- Blain, Frédéric, Senellart, Jean, Schwenk, Holger, Plitt, Mirko and Roturier, Johannes (2011). Qualitative analysis of post-editing for high quality machine translation. In *Proceedings of MT Summit XIII: the Thirteenth Machine Translation Summit*, pages 164-171, Xiamen, China.
- Krings, Hans P. (2001). *Repairing texts: empirical investigations of machine translation post-editing processes*. Vol. 5. Kent State University Press.

Terminology post-editing of neural MT by UTX glossary data

MT Summit 2017

YAMAMOTO Yuji

UTX team leader, AAMT

<http://www.aamt.info/english/utx/>

Presenter: Yamamoto Yuji

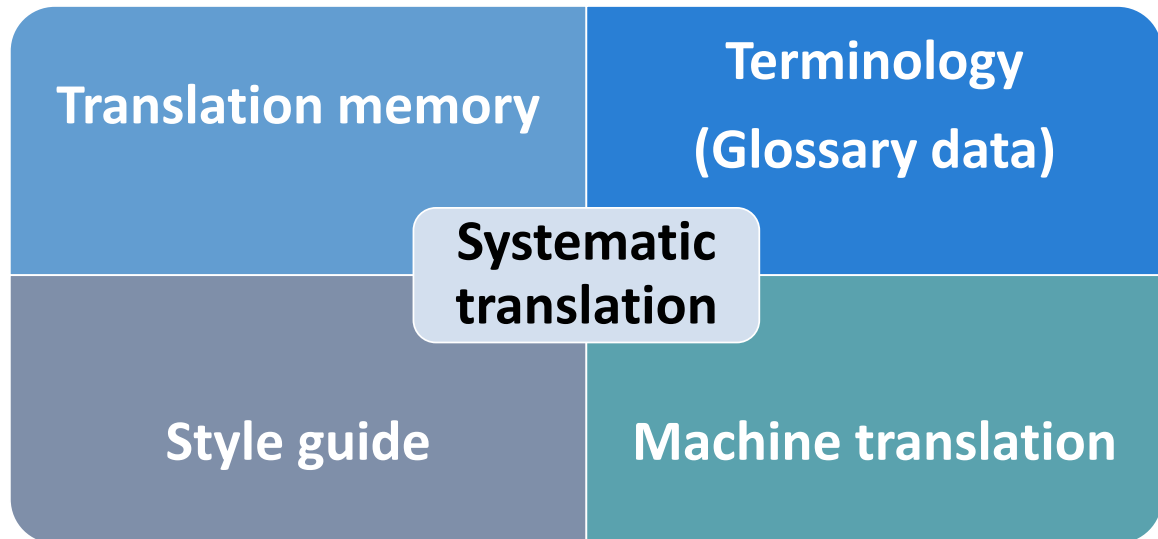
- CosmosHouse Founder/Representative
- Language/translation consultant
- AAMT UTX team leader
- ISO/TC37 (terminology) committee member
- Contact at <http://cosmoshouse.com/mail.htm>

Agenda

1. Background – terminology and NMT
2. UTX – a structured glossary format
3. Terminology post-editing
4. Conclusion

Background – terminology and NMT

Terminology is an essential part of systematic translation



Commercial translation requirements

- Glossaries are established by client companies.
e.g. Microsoft glossaries
- Use of a company vocabulary is not optional.
You are required to use certain terms for translation.

NMT problems

1. NMT mistranslates low-frequency words
2. NMT cannot reflect an existing glossary
3. NMT lacks terminological consistency

Problem 1: NMT mistranslates less-frequent terms

- such as proper names and technical terms
 - e.g. auxiliary verb→*補助動詞 (助動詞 is correct)
 - response rate→*奏功率 (回答率 is correct)
- Missing or repeated translation

Problem 2: NMT cannot reflect an existing glossary

- e.g. “liaison” in ISO context
- A glossary is not an issue for general MT users
- A glossary is essential in a systematic translation
- Many companies are not managing glossaries in an organized manner
- Translation problems are hidden in such an environment

Problem 3: NMT lacks terminological consistency

- e.g. International Standard→国際規格、国際標準
 - resource→資源、リソース
- Terminology consistency is not an issue for general MT users
- But terminology consistency is important in systematic translation

Prevalence of RbMT in Japan

- Strong demands for translation.
- EN-JA bilingual market.
- Early MT commercialization since 1990s.
- Many commercial RbMT packages are sold.

Toshiba



Fujitsu



Cross
Language

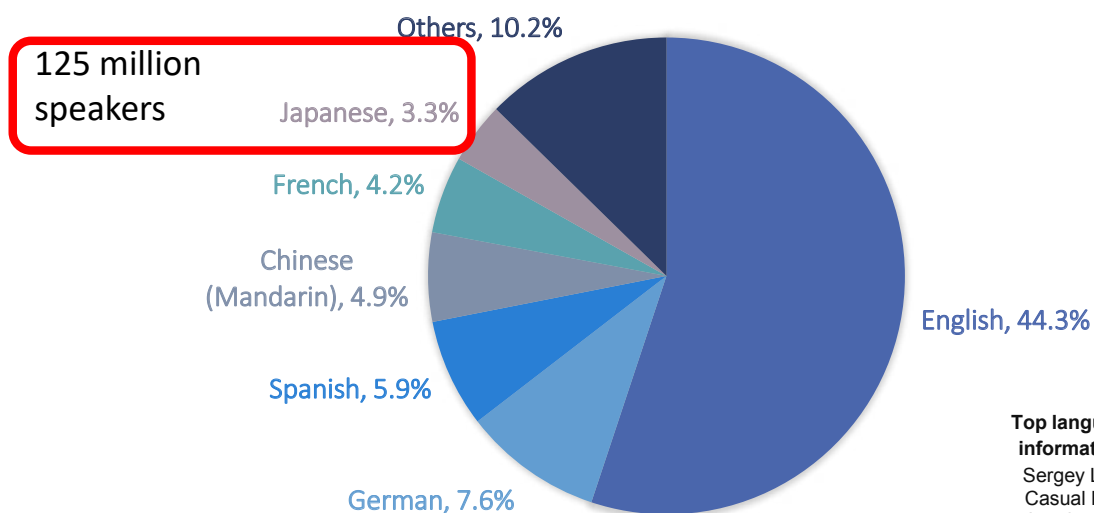


Kodensha



Japanese is an influential language, but its market is bilingual

INFORMATION PRODUCTION



Top languages in global information production
Sergey Lobachev
Casual Reference Librarian
London Public Library

https://journal.lib.uoguelph.ca/index.php/perj/article/view/826/1358#.WY_eh1GrSHs

Bilingual or multilingual scenario?

- Japan – Japanese and English
- Europe, Americas, Africa, etc. - multilingual

Terminology management must be simplified

- Or it will never be implemented.
- Multilingual complexity is not necessary for a bilingual environment.
- A simple Excel sheet is too simple.

UTX – a structured glossary format

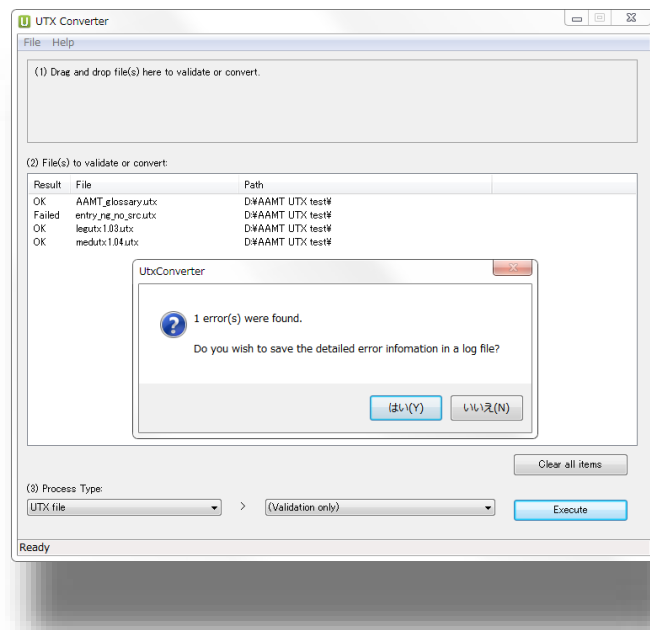
What is UTX (Universal Terminological eXchange)?

Simple but structured glossary data format

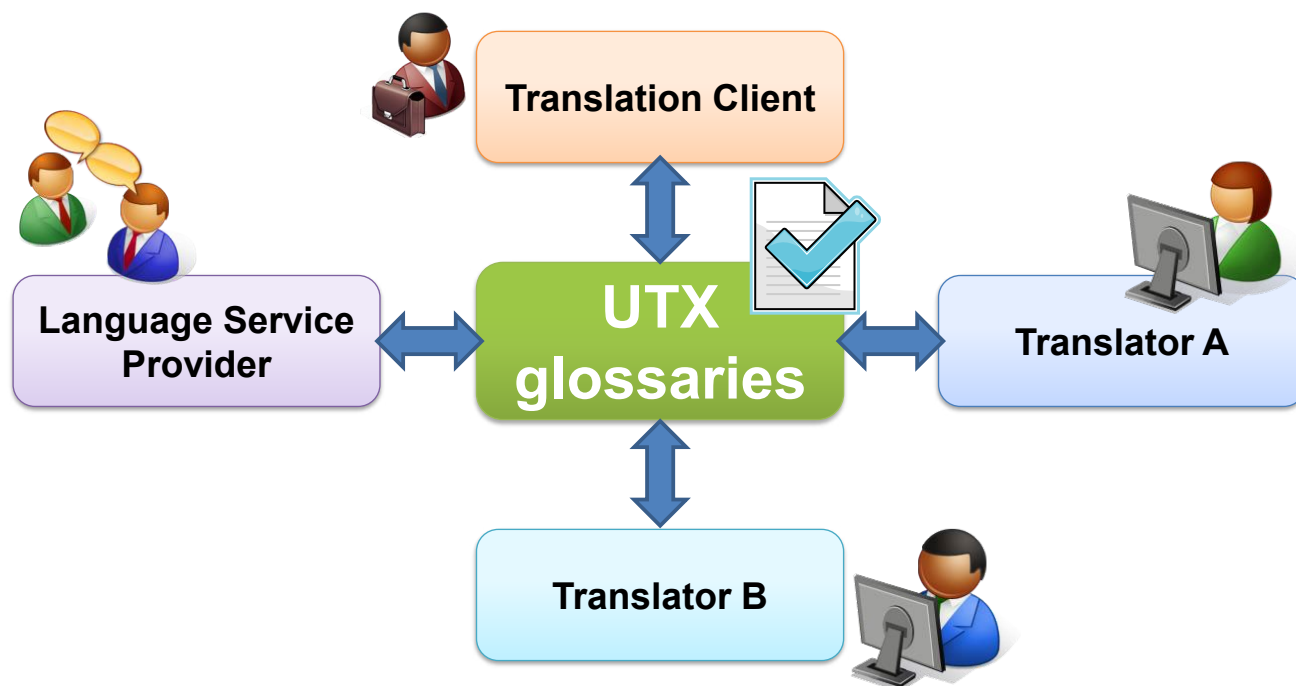
for terminology tools and MT

UTX is free to use

- UTX specification is free
- Many UTX glossaries are free
- UTX Converter is free
- (open source)

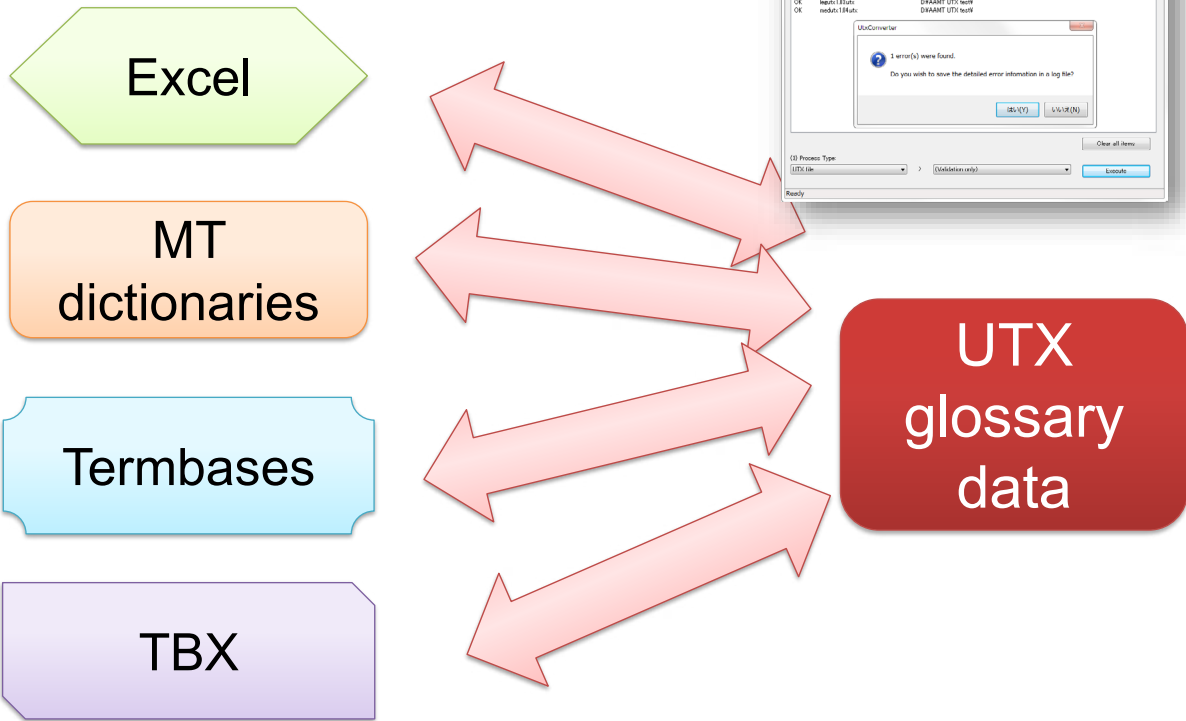


UTX facilitates sharing and reusing of glossaries



Conversion to/from UTX

UTX Converter



UTX glossary sample

Manage essential glossary data in a standardized format

Information about the glossary (creation date, license, etc.)

#UTX 1.11; en-US/zh-CN; 2014-09-25; copyright: AAMT (2012); license: CC-by 3.0

#src	tgt	src:pos	term status
Asia-Pacific Association for Machine Translation	亚洲太平洋机器翻译协会	properNoun	approved
dictionary administrator	字典管理员	noun	approved
contributor	用语提交者	noun	provisional
domain	领域	noun	
glossary	词汇表	noun	
bidirectional	双向	adjective	approved
merge	合并	verb	approved
Source term (American English)	Target term (Chinese)	Part of speech	Term status

Term status provides **reliability**

JPO (Japan Patent Office) UTX glossary

- Created by JPO, converted by AAMT
- Available for free
- Japanese to English
- 130 thousands entries
- Only rare technical terms
 - User-defined terms not included in technical dictionaries shipped with the package

JPO glossary covers all IPC sections

IPC (International Patent Classification)

- A human necessities
- B performing operations; transporting
- C chemistry; metallurgy
- D textiles; paper
- E fixed constructions
- F mechanical engineering; lighting; heating; weapons; blasting
- G physics
- H electricity

Examples of entries

Domain (semantic feature)	Entries	Example	
Plants (common names, species, scientific names, etc.)	5498	白いぼキュウリ	white spine cucumber
		メラレウカ・アルテルフォリア	Melaleuca Alternifolia
		いらくさ科植物	urticaceous plant
Animals (common names, scientific names, etc.)	3025	ヤブカ	striped mosquito
		モンシロチョウ	Pieris rapae
		ユーグレナ	Euglena
People (personal names, titles, etc.)	1316	昌聰	Yoshiaki
		調香士	perfumer
		登壇者	presenter
Companies and organizations	7340	日本醸造協会	Brewing Society of Japan
		猟友会	hunters' association
		インド技術研究所	Indian Institute of Technology
Others	46975	Chemistry, medicine, machine, engineering, and other technical terms.	
		オキシジフタル酸二無水物	oxydiphthalic dianhydride

Patent documents characteristics

1. Extremely long sentences
2. Ambiguous sentence structure
3. Peculiar writing style
- 4. Many technical terms (obfuscation)**

Terminology post-editing

What is “terminology post-editing”?

- post-editing method focused on terminology checking
- requires structured glossary data that has **strong correlation** with the source documents



Terminology post-editing: merits and limitations

■ Merits

- Fully- or partially-automated check
- Check with no lingual knowledge

■ Limitations

- Accuracy is insufficient
(requires other criteria for a full quality assessment)

Quick check or post-editing

Terminology check in SDL Trados

The screenshot displays the SDL Trados interface during a terminology check. A 'Term Recognition' window is open, showing a glossary with the following entries:

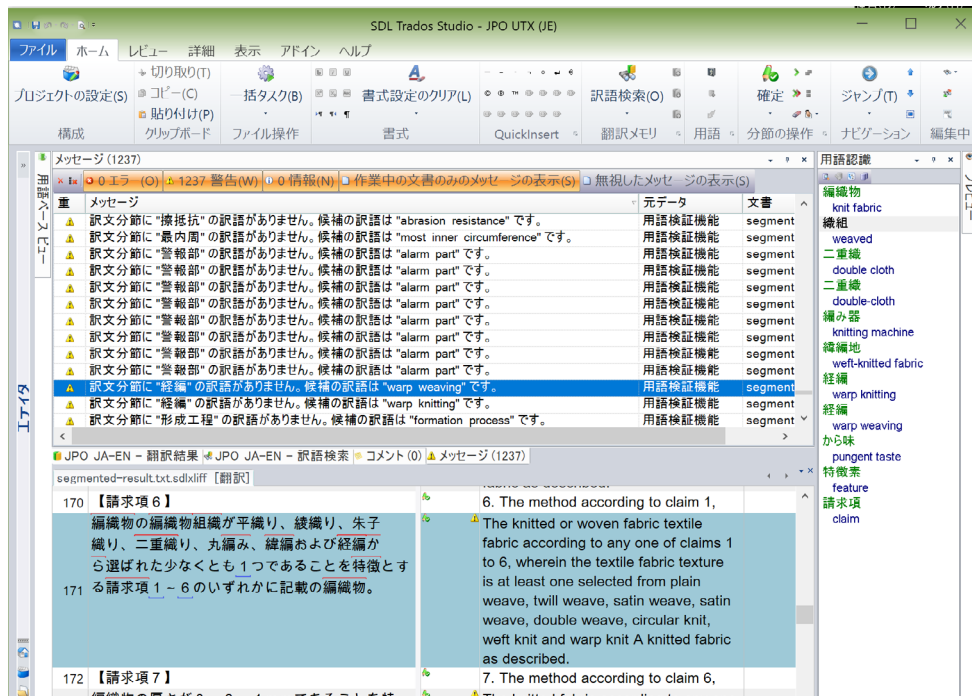
Term	Status
用語集	approved
グロッサリー	forbidden

Below the main interface, a 'Messages' panel shows the following information:

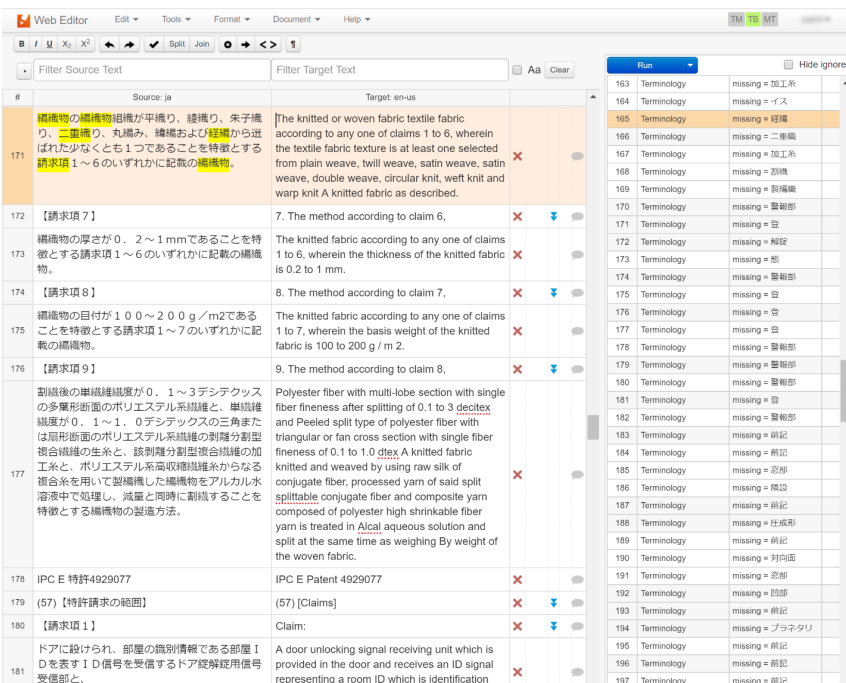
- 92% Translation Details: Status: Translated, Origin: Interactive, Score: 92%
- 100% Before Interactive Editing: Origin: Translation Memory, System: UTX, Score: 92%
- Messages:
 - Information: Your termbase contains the source term "glossary" without translation in the current target language.
 - Error: Wrong usage of the term "グロッサリー" - this term is defined as forbidden.

Red dashed boxes highlight the 'Term Recognition' window and the 'Messages' panel.

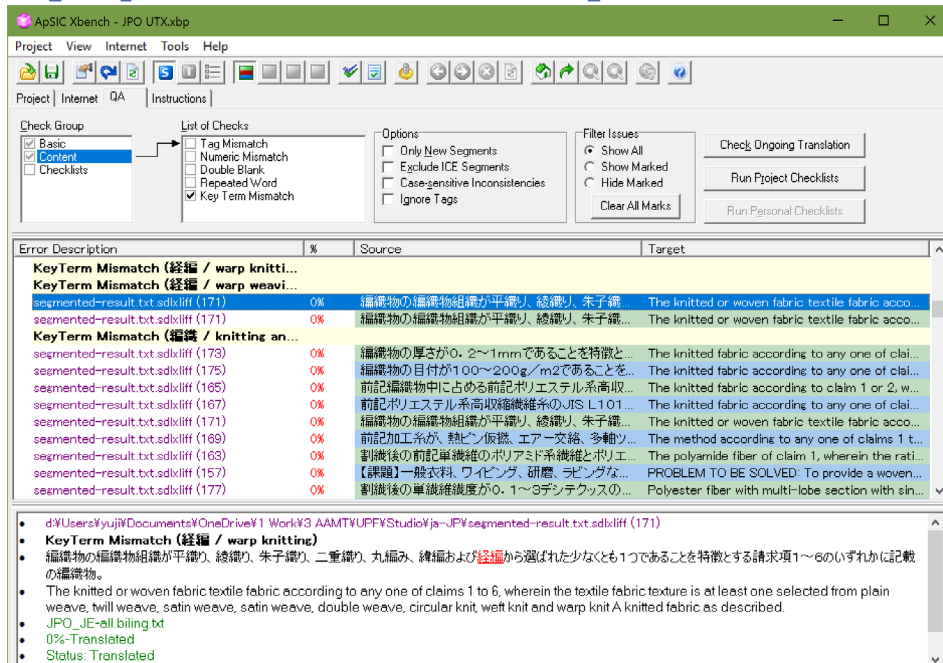
Patent NMT translation checked by UTX glossary data (SDL Trados)



Patent NMT translation checked by UTX glossary data (Memsource)



Patent NMT translation checked by UTX glossary data (ApSIC Xbench)



Result: potential term errors

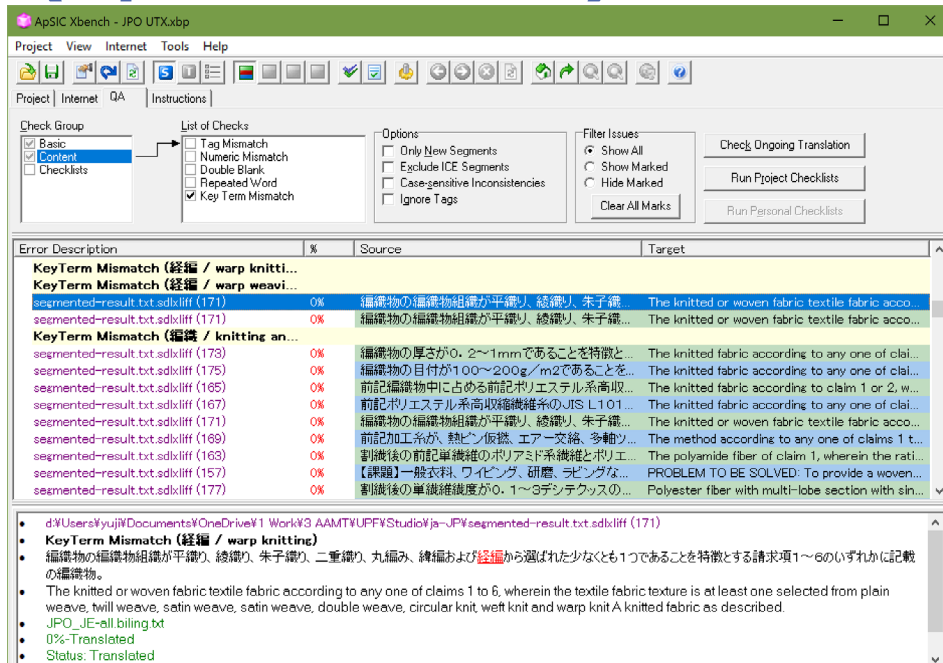
391 segments (sentences). More error detection is not necessarily better.

	Detected potential term errors
SDL Trados	1237
Memsources	372
ApSIC Xbench	603

Examples of incorrectly translated terms:

- 請求項/claim
- デシテックス/decitex
- 経編/warp weaving
- センサシステム/sensor system

Patent NMT translation checked by UTX glossary data (ApSIC Xbench)



Result: potential term errors

391 segments (sentences). More error detection is not necessarily better.

	Detected potential term errors
SDL Trados	1237
Memsources	372
ApSIC Xbench	603

Examples of incorrectly translated terms:

- 請求項/claim
- デシテックス/decitex
- 経編/warp weaving
- センサシステム/sensor system

Conclusion

1. NMT has many terminological flaws.
2. Glossary data and terminological check can find potential term errors.
3. To do so, you need a simple but structured glossary data format (such as UTX).
4. The UTX format was proved to be effective in finding potential errors.

Special thanks to Akimoto Kei (AAMT)

More info

- Visit <http://www.aamt.info/english/utx/> for the specification and glossary data (free)
- Search for **“UTX glossary”**
- Contact at <http://cosmoshouse.com/mail.htm>
- We welcome your feedback!



Harvesting Polysemous Terms from e-commerce Data to Enhance QA

Silvio Picinini

Localization, eBay Inc., San Jose, CA, USA

spicinini@ebay.com

Abstract

Polysemous words can be difficult to translate and can affect the quality of Machine Translation (MT) output. Once the MT quality is affected, it has a direct impact on post-editing and on human-assisted machine translation. The presence of these terms increases the risk of errors. We think that these important words can be used to improve and to measure quality of translations. We present three methods for finding these words from e-commerce data, based on Named Entity Recognition, Part of Speech and Search Queries.

1. Introduction

Polysemous are words or sets of words with multiple meanings. For this work, we consider a broader definition that reflects what polysemy causes in Machine Translation (MT). A polysemous word here is “a word that can have different translations”. This means that the MT engine can be confused about which translation is the correct one; this in turn affects the quality of the machine translation. Instead of “polysemous”, we could call these “polytranslation” terms, and just invent this word.

For example, a word that can be assigned multiple POS (parts-of-speech) tags may have a similar meaning, but it will be translated differently if it is a noun, a verb, or an adjective. Example: Print a report (verb), print magazine (adjective), this fabric has a nice print (noun).

A brand that is also a common word (e.g. Gap, Guess, Coach) will be left untranslated when referring to a brand and will be translated when used as a common word.

Also, a word like “mixer” may have a generic meaning of “a device that mixes”, but in the real world, it can refer to very different products, such as a kitchen mixer or a sound mixer for music. These are two very different devices with very different translations. Also, it can be a party (singles mixer), a very different meaning.

This work presents three new processes that leverage eBay e-commerce data to harvest polysemous words, so that these can be used for different applications, such as the ones described in this paper.

Before going into the methods, we make two general points about data below.

2. Leveraging semantic value and relevance added by the public

We think that it is important to make a point about the importance of capturing semantic meaning from user behavior. It is a massive and no-cost source of information, therefore, we should be interested in using it. One of the methods described uses information from buyer behavior on eBay. By entering a query and then going into a certain category, the buyer is associating meaning to the query, which is comprised of one or two words. The same happens when a seller describes an item for sale and chooses a category for it, giving the words of the

item additional context and meaning. It is also important to capture relevance from user behavior. If we capture how frequent certain meanings are, we are capturing how relevant they are. This “quantification” is something that traditional dictionaries cannot do, and is a significant difference compared to dictionaries.

All of this can be seen as a “public semantic annotation” work that is being done without cost for the companies. While companies collect vast amounts of data, we are all constantly looking for ways to enhance the meaning of this data, and the examples in this work are in line with that overall effort.

3. Harvesting relevant words in context

Another important point about polysemous words and synonyms is these words are relevant because of their meaning relationship; polysemous words have a single form with multiple meanings and synonyms have multiple forms with a single meaning. There are also other types of relationships that can be of interest, such as hyponyms, hypernyms, and meronyms (a word that is a constituent part or a member of another word).

These words are useful in many ways, such as improving and measuring MT, but also improve search queries and possibly classification. The challenge is to find “applied” examples of these words in a specific context. While a dictionary or WordNet can tell us that words are synonyms, it will not tell us that “camcorder” is a synonym for “video camera” or that “flash drive” is the same as a “pendrive”. Finding these words applied to a context, an industry, or a subject matter should be more useful than generic words.

4. Applications of polysemous words

There are some possible applications for polysemous words in machine translation:

- Select data containing these words and create training and testing data for them

Since these words are more likely to be mistranslated, they are more likely to require more in-context training data to help the engine disambiguate different situations. So one application of polysemous words is to find examples of content with these words and have it translated/post-edited. This will allow the creation of training data for the engine to learn how to better handle these words, and the creation of testing data to evaluate the translation.

- Evaluate the quality of the MT of these words

The evaluation of the MT output quality is usually performed on the entire content (using automated metrics) or on a sample (if human evaluation is used). In both cases, there is usually no “selection” of certain segments matching some certain criteria which should be measured, the segments are randomly chosen. However, polysemous words could be used to provide an insight on the quality of the machine translation, using selected “more difficult” words. eBay has started collecting some data around this.

- Evaluate the quality of the post-editing of these words

Training and testing data may be created through post-editing. The quality of that post-editing work needs to be evaluated. Polysemous words are more likely to be mistranslated and

be wrong in the MT output. The post-editing process is supposed to correct those errors. If the error is corrected by the post-editor, this is an indication that the post-editing is of good quality. If the error is not corrected, this is an indication that the post-editing may not be of good quality, and may need further work before being used as training or testing data. The evaluation of post-editing is usually done by an evaluator on a random sample.

Looking at how polysemous words were post-edited is a way to assess the quality of the post-editing work and is also an indication of the final quality of the content that is going to become training or testing data.

5. Three processes to harvest polysemous words

5.1. From eBay search queries

This process is based on associating different categories to the same query. The premise is that if a word is associated with two very different categories, they are likely to have very different meanings, and there is a good chance that the word is polysemous.

Customers enter search queries on eBay. After seeing the results of their queries, they take an action that leads to a certain category. This is an indication of the meaning of the word that was entered as a query. Let's consider an example with the word "mixer". A query for that word does not clarify if the customer is looking for a sound mixer or a kitchen mixer. However, after the query display results, the customer takes action to look into one of these different devices. Once the customer acts, there is now a category that can be attached to the word in the query.

eBay creates a column called Leaf Category Histogram. It looks like Figure 1:

15709/69.108|6224/14.247|95672/6.274|6158/1.696

Figure 1. Leaf Category Histogram

This column contains the identification of the most frequent categories (in black above) accessed by the customer after entering a query. It also contains the % of instances where the customer went to a certain category (in red above). This number is an indication of the "intensity of the polysemy". If a word like "mixer" goes 60% of the time to a music category (for sound mixer) and 40% to a kitchen category, this is an indication that there is significant interest for both meanings. If another word has a 99.8% frequency and the second category is, for example, less than 0.1%, this is an indication that one meaning is nearly universal and the other is extremely rare. This can inform our harvesting of polysemous words.

Starting from that data, we find the higher level eBay categories associated with the word. We are interested in finding big differences in categories, which would be more likely to have different meanings. Once we manipulate the data, we arrive at information that looks like Table 1:

Word	Category 1	Frequency 1	Category 2	Frequency 2	Is Cat 1 diff from 2?	Is Freq 2 > 2%?
Mixer	Music	60%	Kitchen	40%	Yes	Yes

Table 1. Data after manipulation

The last two columns are formulas. With this data we can filter the last two columns and the result will be a list of polysemous candidates. Table 2 show some examples we found in our initial results:

Word	Category 1	Category 2	Comment
Vans	Clothing, Shoes & Accessories	eBay Motors	brand vs. type of car
notebook	Computers/Tablets & Networking	Books	computer vs. writing
Fossil	Jewelry & Watches	Collectibles	brand vs. actual fossil
mixer	Musical Instruments & Gear	Home & Garden	sound table vs. dough mixer
roadrunner	eBay Motors	Toys & Hobbies	car vs. character
Pebble	Cell Phones & Accessories	Pet Supplies	brand of watches
torch	Sporting Goods	Business & Industrial	flashlight vs. hot flame

Table 2. Results from queries

A quick human triage to validate which of these candidates are good produces our final list.

The initial results indicate that this process is efficient. A list of about 1900 queries yielded about 40 candidates. A human triage that took about an hour yielded 19 final terms, about 1% of the initial data.

5.2. From NER data

For Named Entity Recognition, we tag individual tokens, mapping them to different tags according to their meaning. The premise for finding polysemous words in this process is that the same word can be tagged with different tags, and if these tags indicate a significantly different meaning, there is a good chance that the word is polysemous.

This process leans on the concept of polysemous words being defined by “how words are translated”. The most benefit from this process comes from differentiating words that are not translated from words that are translated. The MT engine may be confused and translate brand names, or do not translate common words because they are commonly brand names. The word “charger” can refer to the car Dodge Charger. This is a product name and won’t be translated. But it can also refer to a charger for a cell phone. This is a common word and will be translated. Therefore, it is possible that there is “a charger in a Charger”, and the MT has to deal with this ambiguity.

We start with a list of tokens and tags for a certain category. Once we sort it by token, we will see that some tokens are tagged with different tags. Some NER tags indicate that the token should not be translated: Brand and Product Name. Other tags indicate that the meaning tends to be a common word: Type, Color. We organize the data with additional columns: Do Not Translate indicates when a token is tagged with Brand or Product Name. Translatable indicates when the token is tagged with a category that is usually a common word, and therefore translatable. Once the data is organized in this way, a few manipulations with sorting, filtering and formulas will produce the list of candidates that we are looking for.

Table 3 shows what the data looks like:

Word	Token	Do Not Translate?	Translatable?	Contains Translatable and DNT?
Charger	t		Yes	
Charger	p	Yes		Yes

Table 3. Data after manipulation

Table 4 shows some of our initial results:

Token	Contains DNT tag?	Contains translatable tag?	Comment
Black	b	c	Black and Decker brand vs. color
Case	b	n	Case Logic brand
Charger	md	ta	Dodge Charger car vs. device
RAM	md	ta	Dodge RAM pickup vs. RAM memory
Range	md	f	Range Rover brand vs. common word
Seat	ma	n	Car maker in Spain

Table 4. Results from NER

5.3. From Part of Speech data (POS)

This process is based on identifying when a word is used with different parts of speech in a certain content. It is very common for MT engines to make errors because of a word that is written in the same way, but can be a verb, a noun, or an adjective for example. While the English language does not have any difference for the usage of that word, other languages will have lots of variations for the different POS. Adjectives will have gender in Romance languages, and verbs will have a variety of forms. This brings again the concept that “translations will be different for the same word”, and this may confuse the MT engine and affect the MT quality.

The premise for this process is that if a word is associated with two different POS types, there is a good chance that the word is polysemous (will have different translations).

We run a POS tagger on the content, and the result looks like this:

<S> Loring[Loring/NNP,B-NP-singular|E-NP-singular] was[be/VBD,B-VP] a[a/DT,B-NP-plural] dedicated[dedicated/JJ,dedicate/VBD,dedicate/VBN,I-NP-plural] artist[artist/NN,E-NP-plural] whose[whose/WP\$,B-NP-plural] artistic[artistic/JJ,I-NP-plural] abilities[ability/NNS,I-NP-plural] and[and/CC,I-NP-plural] accomplishments[accomplishment/NNS,E-NP-plural] are[be/VBP,B-VP] beautifully[beautifully/RB,I-VP] shown[show/VBN,I-VP] in[in/IN,B-PP]

this[this/DT,B-NP-singular] book[book/NN,book/VB,book/VBP,E-NP-singular].[./,</S>,O]

With some manipulation, we create a list with two columns: word and POS tag. We sort that list by word, and secondarily by tag, and then we are on our way to identify words that have more than one POS tag, as shown in Table 5:

Word	Tag	Comment
Accessory	J	
Accessory	J	
Accessory	N	accessory tagged as N (noun) and J (adjective)

Table 5. Data after POS tagging and manipulation

Different POS taggers will have different tags, but this process only requires:

- Creating a vertical list of words and tags (usually simple introduction of CR characters)
- Identifying a different part of the tag for nouns, adjectives and verbs (sometimes the first letter of the tag will be enough, as above)

We can also subtotal the list by words and tag, and we will then have information about the frequency that each word and tag occurs. This number indicates the candidates with better potential. One word may have a 60%/40% ratio between noun and verb, while another word may have a 99%/1% ratio. If the same proportion appears in the training data, the first situation will more likely confuse the MT engine than the second situation.

Table 6 below show some of our initial results:

Accessory	N	accessory tagged as N and J
Acted	V	acted tagged as V and J
Adapted	V	adapted tagged as V and J
Added	V	added tagged as V and J
Adhesive	N	adhesive tagged as N and J
Adjusted	V	adjusted tagged as V and J
Adore	V	adore tagged as V and N
Affected	V	affected tagged as V and J

Table 6. Results from POS

6. Quantification effect enhances relevance

The processes presented have a “quantification” effect on the meaning. A term could be polysemous and one of the meaning could be very rare. In practical terms, this would not be a significant polysemy case, because there is no volume for that meaning. The eBay data helps indicating how often a term has one meaning versus another, by connecting the meaning to a frequency number.

In queries, the frequency is defined by the category that follows the term. In NER, the frequency indicates how often each meaning appears in one category, but we can also look across categories. In POS, this effect also appears. In absolute terms, a certain word can be

tagged with several parts of speech. However, one of these POS may be very rare in the context being analyzed, so this variation would not appear in the results.

These are positive effects, because they introduce the frequency/relevance into the analysis and results, as opposed to an analysis based just on the absolute existence of multiple meanings or POS in a dictionary.

7. Conclusion

The processes described here are finding words with limited human effort, indicating that they are efficient. These words are valuable for eBay because they take into account the eBay context. For example, Fossil is a noun and a brand, but a dictionary would not contain the brand. So these processes are finding words in a way that could be difficult to find with other resources. There is also value in the “quantification” of how frequent these words are.

The methods for harvesting polysemous words presented here are only possible due to the wealth of linguistic data that eBay has. We hope that other companies that have data will find these ideas useful, and those who do not have data will feel inspired to create data and use it.

Translation Dictation vs. Post-editing with Cloud-based Voice Recognition: A Pilot Experiment

Julián Zapata

julianz@intr.co

University of Ottawa & InTr Technologies, Ottawa-Gatineau, Canada

Sheila Castilho

sheila.castilho@adaptcentre.ie

ADAPT Centre/School of Computing, Dublin City University, Dublin, Ireland

Joss Moorkens

joss.moorkens@adaptcentre.ie

ADAPT Centre/School of Applied Language & Intercultural Studies, Dublin City University, Dublin, Ireland

Abstract

In this paper, we report on a pilot mixed-methods experiment investigating the effects on productivity and on the translator experience of integrating machine translation (MT) post-editing (PE) with voice recognition (VR) and translation dictation (TD). The experiment was performed with a sample of native Spanish participants. In the quantitative phase of the experiment, they performed four tasks under four different conditions, namely (1) conventional TD; (2) PE in dictation mode; (3) TD with VR; and (4) PE with VR (PEVR). In the follow-on qualitative phase, the participants filled out an online survey, providing details of their perceptions of the task and of PEVR in general. Our results suggest that PEVR may be a usable way to add MT to a translation workflow, with some caveats. When asked about their experience with the tasks, our participants preferred translation without the ‘constraint’ of MT, though the quantitative results show that PE tasks were generally more efficient. This paper provides a brief overview of past work exploring VR for from-scratch translation and PE purposes, describes our pilot experiment in detail, presents an overview and analysis of the data collected, and outlines avenues for future work.

1. Introduction

Machine translation (MT) post-editing (PE) and voice recognition (VR) technology are gaining ground in both translation technology research and the translation industry. Over 50% of international Language Service Providers now offer a PE service using dedicated MT engines integrated into translators’ computer-aided translation environments (Lommel and DePalma, 2016). In a recent survey of 586 translators in the UK, 15% responded that they use VR technology in their work (Chartered Institute of Linguists et al., 2017). These disparate technologies tend not to be deployed in tandem, although both offer translators the potential to increase productivity and reduce the technical effort usually required to translate from scratch when using conventional word-processing hardware and software.

We carried out a pilot experiment to investigate the effects on productivity and on the translator experience (TX) (Zapata, 2016a) of integrating PE with VR and translation dictation (TD) using a sequential mixed-methods design. In the quantitative phase, four

translators performed four translation tasks under four different conditions: (1) conventional TD (i.e., sight-translating using a digital dictaphone), (2) PE in dictation mode (PED) (i.e., dictating approved or amended segments into the same dictaphone), (3) TD with VR (TDVR) (using a cloud-based VR system on a tablet), and (4) PE with VR (PEVR) (using the same VR system as in task 3). The quantitative experiments consisted of three phases during which task times were measured and some input data were collected. Phase I consisted of dictating and post-editing with dictaphone or the VR system; phase II consisted of manually transcribing the recordings from tasks 1 and 2 on the researcher's laptop; and phase III consisted of revising/editing all four translations. As has been noted in a great deal of research about PE, productivity increases alone do not make a tool desirable for translators (see Teixeira, 2014; Moorkens and O'Brien, 2017). Translator attitudes and usability, the TX, are important factors in the adoption of any technology. For this reason, we have appended a follow-on qualitative phase, wherein the participants filled out an online survey, providing details of their perceptions of the task and of PEVR in general.

In this paper, we present our pilot experiment in detail. The paper is structured as follows: First, we provide a brief overview of past work exploring VR for from-scratch translation and PE purposes. Then, we describe the experimental setup, and present an overview and analysis of the quantitative and qualitative results. In the conclusion, we describe avenues for future work.

2. Related Work

2.1. TD and VR

The idea of using human voice to interact with computers and process texts is as old as the idea of computers themselves. For decades, and in recent years more than ever before, voice input has been widely used in a vast array of domains and applications, from virtual assistants on mobile phones to automated telephone customer services; from professional translation to legal and clinical documentation.

Simply put, VR (also known as voice/speech-to-text or automatic speech recognition) technology recognizes human-voice signals and converts them into digital data. The earliest experiments in VR suggested that voice input was expected to replace other input modes such as the keyboard and the mouse in full natural language communication tasks. However, it was soon discovered that speech often performed better in combination with other input modes such as the keyboard itself, as well as touch, stylus and gesture input on multimodal interfaces (Bolt, 1980; Pausch and Leatherby, 1991; Oviatt 2012).

In translation, there has been a long interest in speaking translations instead of typing them. First, in the 1960s and 1970s professional translators often collaborated with transcriptionists, and dictated their translations either directly to the transcriptionist or into a voice recorder (or dictaphone), before having them transcribed later (a technique often referred to as TD). In the 1990s and 2000s, researchers began to explore VR adaptation for TD purposes. Such developments focused mainly on reducing VR word error rates by combining VR and MT. Hybrid VR/MT systems are presented with the source text and use MT probabilistic models to improve recognition; translators simply dictate their translation from scratch without being presented with the MT output (Brousseau et al., 1995; Désilets et al., 2008; Dymetman et al., 1994; Reddy and Rose, 2010; Rodriguez et al., 2012; Vidal et al., 2006). More recently, further efforts have been made to evaluate the performance of translation students and professionals when using commercial VR systems for straight TD (Dragsted et al., 2009; Dragsted et al., 2011; Mees et al., 2013); to assess and analyze

professional translators' needs and opinions about VR technology (Ciobanu, 2014 and 2016; Zapata, 2012), and to explore TD in mobile and multimodal environments (Zapata and Kirkedal, 2015; Zapata, 2016a,b).

2.2. PE and VR

In recent years, the potential of using VR for PE purposes has also been investigated (García-Martínez et al., 2014; Mesa-Lao, 2014; Torres-Hostench et al., 2017). García-Martínez and her collaborators (2014) tested a VR system integrated into a PE environment (both research-level cloud-based systems). They argue that voice input is more interesting than the keyboard alone in a PE environment, not only because some segments may need major changes and therefore could be dictated, but also because, if the post-editor is not a touch typist, the visual attention back and forth between source text, MT text and keyboard adds to the complexity of the PE task.

Mesa-Lao (2014) surveyed student translators, 80% of which (n=15) reported that they would welcome the integration of voice as one of the possible input modes for performing PE tasks. Thus, voice input offers a third dimension to the PE task, making it possible to combine different input modes or to alternate between them according to the difficulty of the task and to the changing conditions of human-computer interaction. Some experiments have also suggested specifically that for certain translators, text types and language combinations, the benefits of VR and PE integration may not be the same (e.g. in terms of efficiency, productivity and cognitive effort) (see Carl et al. 2016a and 2016b).

Tests with VR within a mobile PE app were reported, first by Moorkens et al. (2016), then by Torres-Hostench et al. (2017). Participants were impressed by VR quality and found it useful for long segments. However, they mostly preferred to use the keyboard due to limitations of the software for making minor edits to MT output.

In the following section, we describe our pilot experiment more in detail: our participants' profile and our methodology.

3. Experimental Setup

3.1. Participants' Profile

This experiment included a sample of native (Latin American) Spanish speakers. All four participants are either pursuing or have recently completed a doctoral degree in translation studies. Participants had in common at least a minimum level of acquaintance with the notions of MT, PE and VR. Our sample includes two men and two women between the ages of 26 and 43. Participants reported 3 to 12 years of translation experience, two have training in interpreting, and both of those are regular users of VR (and were therefore familiar with voice commands and other specificities related to dictating with VR). All participants reported to be occasional post-editors.

3.2. Methodology

For this study, we applied a sequential, explanatory mixed-methods design, using the follow-up explanations model, in which the qualitative data is intended to expand upon the quantitative results (Creswell and Plano Clark, 2007:72). We chose this methodology to answer the following two research questions:

1. Can PEVR be as or more productive than comparable approaches, with or without MT and VR?

2. Does the participants' TX suggest that combining MT and VR is feasible for translation projects?

As mentioned in the introduction, four tasks were involved in the quantitative phase of this experiment, namely:

- 1) Conventional TD;
- 2) PED;
- 3) TDVR; and
- 4) PEVR.

A digital dictaphone was used for tasks 1 and 2. A commercial cloud-based speaker-independent VR system¹ was used on an Android tablet for tasks 3 and 4. (See Zapata and Kirkedal (2015) for a description of the different approaches to VR technology with respect to users (i.e. speaker-dependent, speaker-adapted and speaker-independent systems)).

Source texts were 20-segment sections of newstest 2013 data used in WMT² translation tasks. The test sets were analysed using the Wordsmith Wordlist³ tool to ensure that they were statistically similar, based on measurements for type/text ratio, average sentence length, and average word length. Table 1 shows the statistics of the test set.

Text file	Type/token ratio (TTR)	Mean word length (in characters)	Word length std.dev.	Sentences	Mean (in words)
Test Set 1	55.12	4.99	2.51	20	18.05
Test Set 2	55.73	4.80	2.63	20	19.65
Test Set 3	54.31	5.00	2.62	22	21.09
Test Set 4	54.20	5.18	2.69	20	17.25

Table 1. Test set statistics for source texts

A commercial-level MT system⁴ was used to translate the texts. All texts were printed out separately and presented to the participants in hard copy. Naturally, only in tasks 2 and 4 were participants presented with the segmented source and MT texts. The MT texts for tasks 1 and 3 were used only to calculate HTER scores (Snover et al., 2006); more details are provided in section 4.1.2.

Experiments were run individually (i.e. one participant at a time) over four days. A university study room was booked to perform the experiments.

Tasks were randomized as follows:

¹ Dragon Dictation, integrated in the Swype+Dragon app. See <http://www.swype.com/>.

² <http://www.statmt.org/wmt13/>

³ <http://lexically.net/wordsmith/>

⁴ Google Translate. See <https://translate.google.com/>.

Participant	Order of tasks			
ES1	1	2	3	4
ES2	3	4	1	2
ES3	4	3	2	1
ES4	2	1	4	3

Table 2. Participants and order of tasks

Before performing any of the experimental tasks, participants were briefly instructed how to use the digital dictaphone (for tasks 1 and 2) and the VR system on the tablet (for tasks 3 and 4) (i.e., they were given the opportunity to dictate while testing a few voice commands such as punctuation marks, etc.).

The quantitative experiments consisted of three phases during which task times were measured and some input data were collected:

- Phase I - dictating and post-editing with dictaphone or the VR system on the tablet,
- Phase II - manually transcribing the recordings from tasks 1 and 2 (for TD and PED) on the researcher's laptop; and
- Phase III - revising/editing all four translations on the researcher's laptop.

It is important to highlight that during phase II, participants were instructed not to edit the translation, only transcribe what they heard. The documents in which dictations were performed on the tablet for tasks 3 and 4 in phase I were automatically saved into a cloud-based drive⁵ after dictation, and therefore immediately synchronized and available to be edited/revised on the researcher's laptop in phase III.

In phase I, task times were measured using a stopwatch. In both phases II and III, Inputlog (Leijten and Van Waes, 2013) was used. Inputlog is a research-level program designed to log, analyse and visualize writing processes. The program provides data such as total time spent in the document, total time in active writing mode (i.e., of actual keystrokes), total time spent moving/clicking with the mouse, total number of characters typed, total switches between the keyboard and the mouse, etc. Beyond total task times alone, we were interested in collecting this kind of detailed input data, particularly for phase III. We are not reporting data other than task times here given the scope and limitations of this paper; we do consider, however, that input data analysis will be essential in larger-scale experiments.

Thereafter, in the qualitative phase, participants responded to a short online questionnaire, with socio-demographic questions, retrospective questions about the experiment, as well as questions providing insight on the TX with multimodal/mobile VR-enabled TD and PE applications (more details to be provided in section 4.4).

In the following section, some of the data collected is presented and analysed.

4. Results and Analysis

4.1. Task Times Measures (Quantitative Phase)

In order to investigate the effects on productivity of integrating PE with VR and TD in the quantitative phase of this research, we have conducted analysis of the task times as follows:

⁵ Dropbox. See <https://www.dropbox.com>.

1. Comparing tasks of the same nature with and without VR, that is, a) TD vs. TDVR (see 4.1), and b) PED vs. PEVR (see 4.2)
2. Comparing translation vs. PE within phases, that is: a) TD vs PED (4.3) and b) TDVR vs. PEVR (4.4).

We consider:

- a) Translation and/or PE time (phase I + phase II), that is, the time participants needed to translate and/or post-edit, as well as the transcription time (for TD and PED);
- b) Revision duration (phase III), that is, the total time participants needed to review/edit their translation/post-editing;
- c) Total task time (phase I + phase II+ phase III), that is, the total time the participants needed to perform each task.

TD versus TDVR

When comparing both TD tasks (Table 3), i.e. the one performed with a dictaphone (TD) and the one performed with a VR program (TDVR), we can see that the total translation time is always shorter when participants use VR. A reminder to the reader that the total translation time in the dictaphone task includes the time participants need to transcribe their translations (phase II).

Regarding revision duration, however, tasks performed with VR seem to take longer to be completed. We speculate that this is because during the revision time, participants do not only review their translation but also must correct errors produced by the VR program.

Participants	Task	Translation Time			Revision Time	Total Task Time
		Translation time	Transcription time	Total		
ES1	TD	537	716	1253	402	1655
	TDVR	796	n/a	796	656	1452
ES2	TD	688	1197	1885	405	2290
	TDVR	1330	n/a	1330	1191	2521
ES3	TD	846	1116	1962	227	2189
	TDVR	377	n/a	377	722	1099
ES4	TD	700	1432	2132	454	2586
	TDVR	460	n/a	460	1046	1506

Table 3. TD vs TDVR (in seconds)

Overall, when considering all phases, total task time seems to be lower for TDVR, apart from participant ES2, who shows lower time when performing TD.

PED versus PEVR

Results for both PE tasks (PED and PEVR) were also compared (table 4). We notice that the PE time (total) is lower for all participants in the VR condition. As for revision, the time is higher in PEVR, which we assume is for the same reason described in above: that participants also need to correct errors produced by the VR application. However, when considering all phases, participants were still faster post-editing with VR than with the dictaphone.

To compare how much PE was performed for each task, we have calculated the translation edit rate (HTER) (Snover et al. 2016). The HTER score is a measure that compares the raw MT output and the post-edited version, and goes from 0 to 1, where the higher number, the more modifications were made in the raw MT output. We can see in table 4 that most of the participants have an average score of 0.2 – which indicates that little post-editing was performed. However, participant ES3 displays more post-editing performed for the PED task (0.52).

Participants	Task	PE Time			Revision Time	Total Task Time	HTER
		PE time	Transcription time	Total			
ES1	PED	633	692	1325	238	1563	0.24
	PEVR	623	n/a	623	776	1399	0.23
ES2	PED	822	604	1426	537	1963	0.24
	PEVR	910	n/a	910	606	1516	0.17
ES3	PED	612	1366	1978	270	2248	0.52
	PEVR	344	n/a	344	475	819	0.25
ES4	PED	396	1725	2121	654	2775	0.26
	PEVR	1176	n/a	1176	1007	2183	0.14

Table 4. PED vs PEVR (times are in seconds)

TD versus PED

As mentioned above, we also decided to consider the differences between translation and PE when both were performed in the same manner; that is TD and PED; and TDVR and PEVR.

Table 5 compares the results for TD and PED. When looking at the results for translation and PE translation time (total task time; last column), we notice that the results are mixed: while participants ES1 and ES2 were faster with TD, the other two participants (ES3 and ES4) were faster with PED. Interestingly, the transcription time is inversely higher, that is, participants ES1 and ES2 had higher transcription time for the TD tasks, whereas ES3 and ES4 had higher transcription time in PED. Now, when considering the total translation/PE time, we can see that the results are very close, the more visible differences lying for ES1 and ES2, where the former is faster with TD and the latter with PED.

In sum, when looking at the different time measures across phases, we notice no trend in the results. This indicates that, in general, there were not many differences between TD and PED.

Participants	Task	Translation/PE Time			Revision Time	Total Task Time
		Translation/PE time	Transcription time	Total		
ES1	TD	537	716	1253	402	1655
	PED	633	692	1325	238	1563
ES2	TD	688	1197	1885	405	2290
	PED	822	604	1426	537	1963
ES3	TD	846	1116	1962	227	2189
	PED	612	1366	1978	270	2248
ES4	TD	700	1432	2132	454	2586
	PED	396	1725	2121	654	2775

Table 5. TD vs PED (in seconds)

Table 6 compares the results for TDVR and PEVR. We can see that total task times are lower for the first three participants when post-editing with VR than translating from scratch. Only participant ES4 was faster in the translation task. Interestingly, participant ES4 displayed close times for revision, whereas participant ES1 showed lower times to revise the translation. In sum, only participant ES4 showed higher times when post-editing than when translating from scratch, which suggests that PE with the help of VR could generally lead to higher productivity.

Participants	Task	Translation/PE Time	Revision Time	Total Task Time
ES1	TDVR	796	656	1452
	PEVR	623	776	1399
ES2	TDVR	1330	1191	2521
	PEVR	910	606	1516
ES3	TDVR	377	722	1099
	PEVR	344	475	819
ES4	TDVR	460	1046	1506
	PEVR	1176	1007	2183

Table 6. TDVR vs PEVR (in seconds)

4.2. TX Analysis (Qualitative Phase)

In the follow-on, qualitative phase of this experiment, participants responded to an online questionnaire with sociodemographic questions (see *Participant's profile* in section 3.1 above) and retrospective questions about the experiment, as well as questions providing insight on the TX with multimodal/mobile VR-enabled TD and PE applications. The notion of TX is inspired from the notion of user experience (UX) – extensively investigated in the field

of human-computer interaction – and is defined as “a translator’s perceptions of and responses to the use or anticipated use of a product, system or service” (Zapata, 2016a).

In this section, we report on the results of our questionnaire.

Subjectively Experienced Productivity

The questionnaire included an item to ask participants to indicate which one of the four translation tasks they *felt* made them most productive, and which one made them least productive. Three participants believed that TDVR made them most productive when in fact they had performed the PEVR task faster. Two participants felt that they were slowest in the PED condition. This perception of slower pace when MT has been introduced, contradicting quantitative measurements that recorded increased speed, has been seen elsewhere by Plitt and Masselot (2010) and Gaspari et al. (2014). When compared to their actual productivity times, we note that apart from ES1 regarding TD (where he/she is least productive), the other participants perceive it differently from the actual numbers. Table 7 below shows the perceived productivity against the actual productivity, where l/L = least, m/M = most, lower-case letters are for the perceived productivity and capital letters for the actual productivity.

Participant	TD	PED	TDVR	PEVR
ES1	l/L		m	M
ES2		l	m/L	M
ES3		l	m/L	M
ES4	m	L	l/M	

Table 7. Subjectively experienced productivity against actual productivity

Subjectively Perceived Quality

The questionnaire also included an item to ask participants to indicate which one of the four translation tasks they *felt* would result in the best quality, and which one would result in the worst quality (that is, quality of the final target text). Table 8 shows that two of the four participants were confident enough in the PEVR process, that they expected the output texts from that process to be of high quality.

Participant	TD	PED	TDVR	PEVR
ES1	worst			best
ES2			worst	best
ES3		worst	best	
ES4	best			worst

Table 8. Subjectively perceived quality

Challenges for VR-enabled TD and PE

A further question asked participants to elaborate on what they thought are the challenges of VR, on the one hand, and of MT, on the other hand, to provide translators with a useful VR-enabled TD and PE tool.

Participants found VR to be reasonably accurate, but with room for improvement, particularly regarding “proper names and figures”. Participants preferred translation without the ‘constraint’ of MT as they considered the suggestions artificial. Participant ES2 wrote that “the Spanish translation sounded more like a transliteration of a technical text in English, and this is not translation as far as I understand”. The added cognitive load when MT is added to source and target texts may be initially off-putting for translators, and may add to the perception of decreased speed when MT is introduced to the workflow. They recognized that VR and MT could aid productivity, but would prefer to add MT electively. Participant ES1 wrote that “a translator or post-editor should have the option to translate from scratch by default, and request the help from the machine only when needed”. Participant ES2 agreed: “For quality purposes, I prefer the [VR] translation from scratch or post-editing from [translation memories] where you have more leeway.” In the opinion of participant ES4, “MT makes work faster but not necessarily better. It somehow guides the work towards the paradigmatic level. I think the overall cohesion of the document is affected.”

Advantages and Disadvantages of Mobile *versus* PC-based TD and PE

Finally, participants were asked to elaborate on the perceived advantages and disadvantages of using a mobile TD and PE tool (i.e., on a mobile device such as a smartphone or a tablet) *versus* a laptop- or PC-based tool. Several mentioned the flexibility of a mobile device, and participant ES2 suggested that “it may help translators to develop interpreting strategies; such as segmentation, quick thinking, anticipation, short-term memory, etc.” Two participants mentioned the difficulties of working in a noisy environment and of speaking translations in a public place. Participant ES3 felt that, although PEVR felt fast to him/her, it was difficult to edit retrospectively. He/she added that if there was “a way to make it more seamless between the keyboard and the mic, a balance so to say, then that’d be amazing.”

5. Conclusion and Future Work

We have reported a pilot experiment on the use of a cloud-based voice recognition (VR) application for translation dictation (TD) and post-editing (PE), using both quantitative and qualitative methods.

In answer to our first research question, based on this small-scale pilot experiment, PE with VR can be as or more productive than comparable approaches, with or without machine translation (MT) and VR. When looking at quantitative data alone, our results showed that, in general, PE with the aid of a VR system was the most efficient method, being the fastest for three of the participants. Interestingly, PE in dictation mode (PED) was the slowest for two participants, followed by TD and TD with VR (TDVR). In the quantitative data, however, we observe that most participants perceived productivity to be higher in the TDVR condition, and expressed a preference to translate/dictate from scratch and have PE added as an option.

One of the issues we identified in our experiment is high revision/editing times in the VR tasks; transcriptions by the VR system were far from flawless, leading to higher revision/editing times. VR applications may produce errors due to translators’ lack of familiarity with TD and insufficient training in how to speak to a VR system, especially for properly adding punctuation using the appropriate commands. Trainers and researchers in translation have explicitly affirmed that training in sight translation, TD, and VR will be essential to succeed with (mobile) voice-enabled tools and devices (Mees et al. 2013; Zapata

and Quirion, 2016). We noted also that some foreign-language words (e.g. Russian names) in the source texts caused a few misrecognitions in Spanish VR. Moreover, we noticed that some participants would often wait until the software had transcribed a sentence or chunk of a sentence onto the word processor page to continue speaking, which tends to confuse the system (as opposed to when the dictation is continuous). Lastly, if the user pauses for several seconds, the VR system “stops listening” and disconnects, which also causes both the system and the user to lose the flow of the dictation.

Another point to highlight is that the participants’ typing skills may considerably affect translation times. If our time task measures excluded the transcription time in TD and PED, the whole productivity picture would change. Considering this and the issues described in the previous paragraph, the ideal scenario would be one in which translators do not need to transcribe their dictation, either in TD or PE. Instead, they would have a VR system with human-like transcription capabilities, keeping dictation, transcription, and editing/revision times (as well as recognition errors) to a minimum.

In answer to our second research question, participants’ TX suggests that combining MT and VR is indeed feasible for translation projects, with some caveats. When asked about their experience with the tasks, our participants seem to have preferred translation without the ‘constraint’ of MT as they considered the suggestions artificial, though the quantitative results show that the PE task was more efficient than that of translation from scratch. The results of this small-scale experiment suggest that PE with VR (PEVR) may be a usable way to add MT to a translation workflow, and is worth testing at a larger scale.

For future work, we intend to carry out experiments with more participants and language pairs. Further experimentation will include input logging, as well as eye-tracking technologies to collect empirical data on cognitive effort when using VR for TD and PE. We also seek to evaluate the impact of training translators in TD and VR over a period of time before performing TDVR and PEVR tasks. Also, we will include objective measures of quality (with the participation of expert evaluators) to compare it with the participants’ perceived quality of the target texts. Another avenue for future work is to investigate a collaborative scenario in which translators/post-editors collaborate with transcriptionists and/or revisers who would take part in the different phases of the experiment. This list of ideas for future work is of course non-exhaustive; the possibilities seem endless.

The unprecedented robustness of VR technology and its availability on mobile devices via the cloud opens a world of possibilities for human-aided MT and human translation environments. By keeping human translators at the core of research, with strong consideration of their perceptions and preferences for new technologies and applications, we can advance towards finding the right balance in translator-computer interaction (O’Brien, 2012), towards establishing what it is that the machine can do better than humans, and what it is that humans can do better than the machine.

Acknowledgement

We would like to thank our anonymous participants for their time and involvement in this pilot experiment. This work was supported by the ADAPT Centre for Digital Content Technology, funded under the SFI Research Centres Programme (Grant 13/RC/2106) and co-funded under the European Regional Development Fund.

References

- Bolt, R. A. (1980). “Put-that-there”: Voice and gesture at the graphics interface. In *Proceedings of the SIGGRAPH’80*, pages 262–270. ACM Press.

- Brousseau, J., Drouin, C., Foster, G., Isabelle, P., Kuhn, R., Normandin, Y., & Plamondon, P. (1995). French speech recognition in an automatic dictation system for translators: The TransTalk project. In *Proceedings of Eurospeech '95*, pages 193-196, Madrid, Spain.
- Carl, M., Aizawa, A., & Yamada, M. (2016a). English-to-Japanese Translation vs. Dictation vs. Post-editing: Comparing Translation Modes in a Multilingual Setting. In *The LREC 2016 Proceedings: Tenth International Conference on Language Resources and Evaluation*, pages 4024–4031, Portorož, Slovenia.
- Carl, M., Lacruz, I., Yamada, M., & Aizawa, A. (2016b). Comparing spoken and written translation with post-editing in the ENJA15 English to Japanese Translation Corpus. In *The 22nd Annual Meeting of the Association for Natural Language Processing (NLP2016)*, Sendai, Japan.
- Chartered Institute of Linguists, European Commission Representation in the UK, and the Institute of Translation and Interpreting. (2017). *UK Translator Survey: Final Report*. Technical Report. Chartered Institute of Linguists (CIOL), London, UK.
- Ciobanu, D. (2014). Of Dragons and Speech Recognition Wizards and Apprentices. *Revista Tradumàtica*, 12: 524–538.
- Ciobanu, D. (2016). Automatic Speech Recognition in the Professional Translation Process. *Translation Spaces*, 5(1): 124–144.
- Désilets, A., Stojanovic, M., Lapointe, J.-F., Rose, R., and Reddy, A. (2008). Evaluating Productivity Gains of Hybrid ASR-MT Systems for Translation Dictation. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 158-165, Waikiki, USA.
- Dragsted, B., Hansen, I. G., and Selsøe Sørensen, H. (2009). Experts Exposed. *Copenhagen Studies in Language*, 38: 293–317.
- Dragsted, B., Mees, I. M., and Hansen, I. G. (2011). Speaking your translation: students' first encounter with speech recognition technology. *Translation & Interpreting*, 3(1): 10-43.
- Dymetman, M., Brousseau, J., Foster, G., Isabelle, P., Normandin, Y., and Plamondon, P. (1994). Towards an Automatic Dictation System for Translators: the TransTalk Project. In *Fourth European Conference on Speech Communication and Technology*, page 4, Yokohama, Japan.
- Garcia-Martinez, M., Singla, K., Tammewar, A., Mesa-Lao, B., Thakur, A., Anusuya, M. A., Bangalore, S., Carl, M. (2014). SEECAT: ASR & Eye-tracking Enabled Computer-Assisted Translation. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, pages 81–88, Dubrovnik, Croatia.
- Gaspari, F., Toral, A., Kumar Naskar, S., Groves, D., Way, A. (2014). Perception vs Reality: Measuring Machine Translation Post-Editing Productivity. In *Proceedings of AMTA 2014 Workshop on Post-editing Technology and Practice*, pages 60-72, Vancouver, Canada.
- Leijten, M., and Van Waes, L. (2013). Keystroke Logging in Writing Research: Using Inputlog to Analyze and Visualize Writing Processes. *Written Communication*, 30(3): 358–392.

- Lommel, A. R and DePalma, D. A. (2016). *Europe's Leading Role in Machine Translation: How Europe Is Driving the Shift to MT*. Technical Report. Common Sense Advisory, Boston, USA.
- Mees, I. M., Dragsted, B., Hansen, I. G., and Jakobsen, A. L. (2013). Sound effects in translation. *Target*, 25(1): 140–154.
- Mesa-Lao, B. (2014). Speech-Enabled Computer-Aided Translation: A Satisfaction Survey with Post-Editor Trainees. In *Workshop on Humans and Computer-assisted Translation*, pages 99-103, Gothenburg, Sweden
- Moorkens, J., and O'Brien, S. (2017). Assessing user interface needs of post-editors of machine translation. In *Human Issues in Translation Technology: The IATIS Yearbook*, pages 109-130. Taylor & Francis.
- Moorkens, J., O'Brien, S., and Vreeke, J. (2016). Developing and testing Kanjingo: a mobile app for post-editing. *Tradumàtica*, 14: 58-65.
- O'Brien, S. (2012). Translation as human–computer interaction. *Translation Spaces*, 1(1): 101–122.
- Oviatt, S. (2012). Multimodal Interfaces. In J. A. Jacko (Ed.), *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications* (3rd ed., pages 415-429). Lawrence Erlbaum Associates.
- Pausch, R., and Leatherby, J. H. (1991). An Empirical Study: Adding Voice Input to a Graphical Editor. *Journal of the American Voice Input/Output Society* 9(2): 55-66.
- Plitt, M., Masselot, F. (2010). A productivity test of statistical machine translation post-editing in a typical localization context. *Prague Bulletin of Mathematical Linguistics* 93: 7-16.
- Reddy, A., and Rose, R. C. (2010). Integration of Statistical Models for Dictation of Document Translations in a Machine Aided Human Translation Task. *IEEE Transactions on Audio, Speech and Language Processing*, 18(8): 1-11.
- Rodriguez, L., Reddy, A., and Rose, R. (2012). Efficient Integration of Translation and Speech Models in Dictation Based Machine Aided Human Translation. In *Proceedings of the IEEE 2012 International Conference on Acoustics, Speech, and Signal Processing*, 2: 4949-4952.
- Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223-231, Cambridge, USA.
- Teixeira, C. S. C. (2014). Perceived vs. measured performance in the post-editing of suggestions from machine translation and translation memories. In *Proceedings of AMTA 2014 Workshop on Post-editing Technology and Practice*, pages 45-59, Vancouver, Canada.
- Torres-Hostench, O., Moorkens, J., O'Brien, S., and Vreeke, J. (2017). Testing interaction with a Mobile MT post-editing app. *Translation & Interpreting*, 9(2):138-150.

- Vidal, E., Casacuberta, F., Rodríguez, L., Civera, J., and Martínez Hinarejos, C. D. (2006). Computer-assisted translation using speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 14(3): 941-951.
- Zapata, J. (2012). *Traduction dictée interactive : intégrer la reconnaissance vocale à l'enseignement et à la pratique de la traduction professionnelle*. M.A. thesis. University of Ottawa.
- Zapata, J. (2016a). Translating On the Go? Investigating the Potential of Multimodal Mobile Devices for Interactive Translation Dictation. *Tradumatica*, 14: 66-74.
- Zapata, J. (2016b). *Translators in the Loop: Observing and Analyzing the Translator Experience with Multimodal Interfaces for Interactive Translation Dictation Environment Design*. PhD thesis. University of Ottawa.
- Zapata, J., and Kirkedal, A. S. (2015). Assessing the Performance of Automatic Speech Recognition Systems When Used by Native and Non-Native Speakers of Three Major Languages in Dictation Workflows. In *Proceedings of the 20th Nordic Conference of Computational Linguistics*, pages 201-210, Vilnius, Lithuania.
- Zapata, J., and Quirion, J. (2016). La traduction dictée interactive et sa nécessaire intégration à la formation des traducteurs. *Babel*, 62(4): 531-551.

TOIN

WILL NEURAL MT BE A BREAKTHROUGH IN ENGLISH-TO-JAPANESE TECHNICAL TRANSLATION?

Tsunao Mikasa and Nobuko Kasahara, TOIN Corporation
September 2017

AGENDA

- **Question:** Is MT really usable for English-to-Japanese translation?
- **Pilot project** we carried out for assessing quality and productivity
 - Overview
 - MT engines examined
 - Methodology and assumptions
 - Results
- **Conclusion**



QUESTION

Is MT usable for **English-to-Japanese (E2J)** translation services where the **required quality** is at the same level as **Human Translation (HT)**?

- Until recently, the answer was **NO**; to obtain certain productivity gains in post-editing, quality of final translation needed to be compromised
- In other words, only “Light PE” was worth considering, and “real” translation was achievable only by human translators with no help of “machine translators”

QUESTION (CONTD.)

Is MT usable for **English-to-Japanese (E2J)** translation services where the **required quality** is at the same level as **Human Translation (HT)**?

- We claim that the answer will be **YES** if using the latest MT technologies, in particular **neural** engines (under some reasonable assumptions about content types)
- In other words, MT will enable most E2J translators to achieve the **same quality** without compromise at **higher productivity** (except for some special content types, such as marketing materials)

PILOT PROJECT

To examine our claim, we carried out a simple pilot project for accessing **quality** and **productivity** in Human Translation (HT) and Post-Editing (PE)

Key Assumptions:

- We focused on **Technical** documents, as this sector accounts for the largest portion of many language service providers in Japan
- PE quality was required to be **the same level as HT**, since our interest was in examining whether HT quality can be achieved by PE without any compromise in quality (not “Light PE”)

MT ENGINES EXAMINED

We examined two engines which are recognized as ones of the best **Neural** and **Statistical English-to-Japanese** MT engines:

- **Google NMT**—Neural
- **NICT みんなの自動翻訳@TexTra®** —Statistical

(NICT: National Institute of Information and Communications Technology 情報通信研究機構)

Note: NICT has recently also released its Neural engine

METHODOLOGY AND ASSUMPTIONS

Content translated: A typical technical document, User Manual of a major PLM software product

- **Not too technical**, easy-to-understand for the average user (and for translators!)

Volume:

- **5k** words for PE/HT productivity evaluation
- Additional **10k** words for MT quality evaluation

METHODOLOGY AND ASSUMPTIONS (CONTD.)

Sample segments:

Segm	Source Segment	MT Target Segment	Translator KH			Translated Target
			MT Engine	PE-er	Post-edited Target	
245	The action is only available when creating or editing a change task.	この操作は、変更タスクを作成または編集するときに使用できるようになります。	NICT	KH	この操作は、変更タスクを作成または編集するときのみ使用できます。	
246	The action is only available when you access the Resulting Objects table from the change task information page.	操作は、変更タスクの情報ページの「結果オブジェクト」(Resulting Objects)テーブルにアクセスした場合にのみ使用できます。	NICT	KH		この操作は、変更タスク情報ページから「結果のオブジェクト」テーブルにアクセスするときのみ使用できます。
247	Open a new window to edit the change task.	新しいウィンドウが開き、変更タスクを編集します。	NICT	KH	新しいウィンドウを開き、変更タスクを編集します。	
248	Set effectivity on an object.	オブジェクトのエフェクティビティを設定します。	NICT	KH		オブジェクトで有効性を設定します。
249	View effectivity on an object.	オブジェクトのエフェクティビティを表示します。	NICT	KH	オブジェクトの有効性を表示します。	

METHODOLOGY AND ASSUMPTIONS (CONTD.)

Resources—Linguists (Translators/Post-Editors) who worked in the pilot:

- **Four** senior-level linguists with 10+ year-experience in E2J technical translation
- Past experience in PE was **not** required (though two of them did have some PE experience)
- Each of them translated/PE'd the same 5,000-word sample document
- They focused on achieving sufficient (HT-level) quality in PE; never forced to use MT outputs or “hurry up” in PE

METHODOLOGY AND ASSUMPTIONS (CONTD.)

Linguistic reference:

Made linguistic reference as **simple** as possible to see the pure impact of MT on quality and productivity:

- No Translation Memory (TM)
- No Terminology Database (TD)
- No Style Guidelines (SG)

METHODOLOGY AND ASSUMPTIONS (CONTD.)

Pilot project for Productivity evaluation

- Each linguist produces a translation of each segment, either by
 - **HT**: translating the source segment without referring to any MT outputs, or
 - **PE**: editing MT output of the source segment
 - To do HT or PE is randomly chosen by the system so that the total # of HT/PE'd segments will be equal

METHODOLOGY AND ASSUMPTIONS (CONTD.)

Pilot project for Productivity evaluation (contd.)

- For **PE**, either **GNMT** or **NICT** engine applied
 - Randomly chosen by the system so that the total # of the segments from each engine will be equal
 - Not make it visible to the linguist which engine was used (to avoid any bias)
- We used **TAUS DQF tools** for productivity evaluation

METHODOLOGY AND ASSUMPTIONS (CONTD.)



Post-editing on TAUS DQF tools:

TAUS EVAL The Industry's Benchmark

Hi Evaluator's Name

Home

Project-Name

Information
Required Level of Quality: *Good Enough*
Content Type: Website Content
Filename:
Segment: 1 of 8

Source: English (United States)
Start
Current Himeji Castle (Himeji-jo) is the largest, most perfectly designed original castle that remains in Japan.
Next It was also called the Egret Castle as the shape of the castles layout centered on the five-tiered donjon resembled an egret about to take flight.

Target: Japanese
Start
Current 姫路城(姫路城)は、日本に残っている最も大きくて完全に設計されたオリジナルの城です。

PAUSE

NEXT
Or Press Enter

Please write to us with any questions at dqf@taus.net.
Copyright TAUS 2014

[PAUSE] ボタン
中断時にクリック

[NEXT] ボタン
(クリック後は戻れない)



METHODOLOGY AND ASSUMPTIONS (CONTD.)

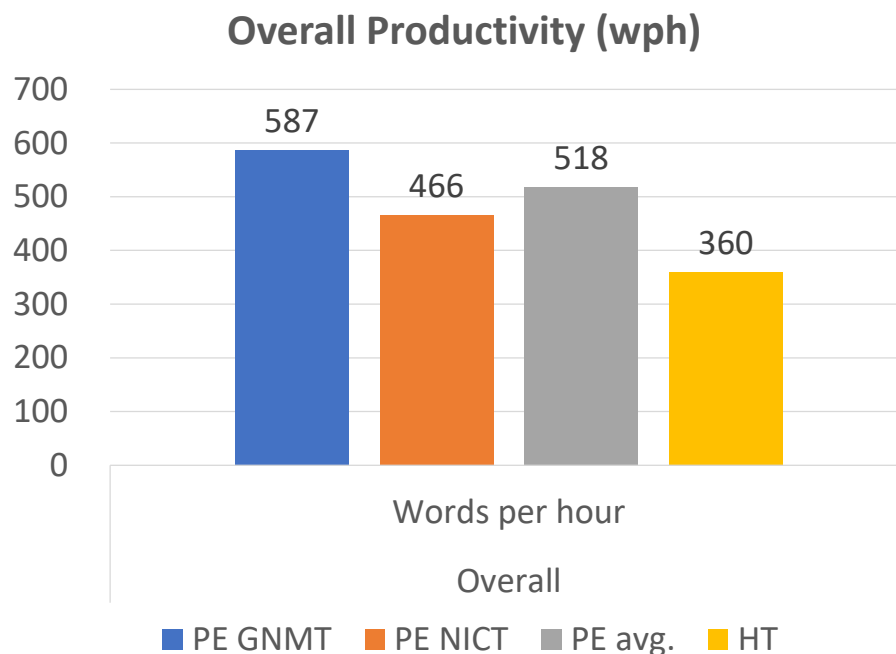
Quality evaluation of raw MT outputs

- Evaluated quality of raw MT outputs of **GNMT** and **NICT** engines
 - Randomly chosen by the system so that the total # of the segments from each engine will be equal
 - Not make it visible to the evaluator which engine was used (to avoid any bias)

METHODOLOGY AND ASSUMPTIONS (CONTD.)

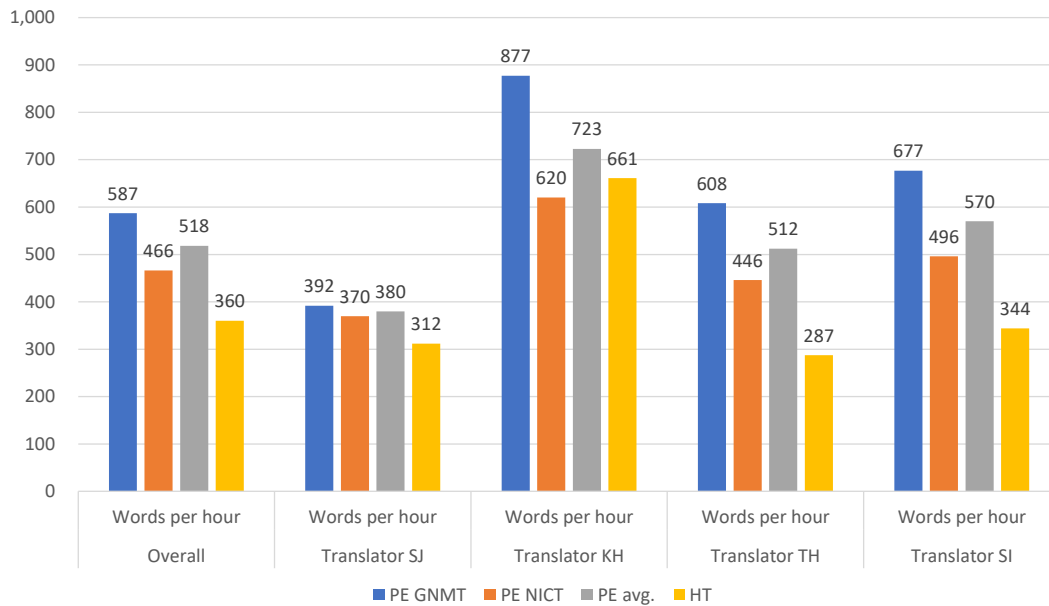
- We used **TAUS DQF tools** and their evaluation criteria for quality evaluation
 - **Fluency**
 - **Flawless (4)** —refers to a perfectly flowing text with no errors.
 - **Good (3)** —refers to a smoothly flowing text even when a number of minor errors are present.
 - **Disfluent (2)** —refers to a text that is poorly written and difficult to understand.
 - **Incomprehensible (1)** —refers to a very poorly written text that is impossible to understand.
 - **Adequacy**
 - **Everything (4)**—All the meaning in the source is contained in the translation, no more, no less.
 - **Most (3)**—Almost all the meaning in the source is contained in the translation.
 - **Little (2)**—Fragments of the meaning in the source are contained in the translation.
 - **None (1)**—None of the meaning in the source is contained in the translation.

RESULTS—PRODUCTIVITY



RESULTS—PRODUCTIVITY (CONTD.)

Productivity (wph) by Translator



RESULTS—PRODUCTIVITY (CONTD.)

Key findings:

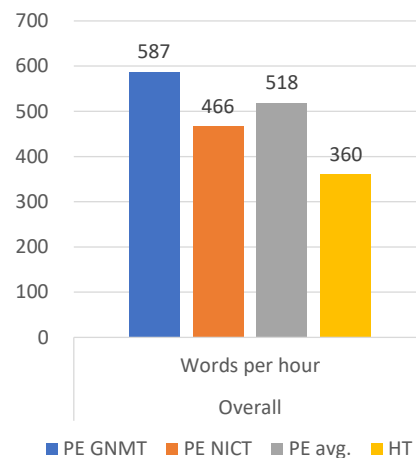
○ PE w/ GNMT

- Highest productivity
- 63% faster than HT on average

○ PE w/ NICT

- 30% faster than HT on average

Overall Productivity (wph)

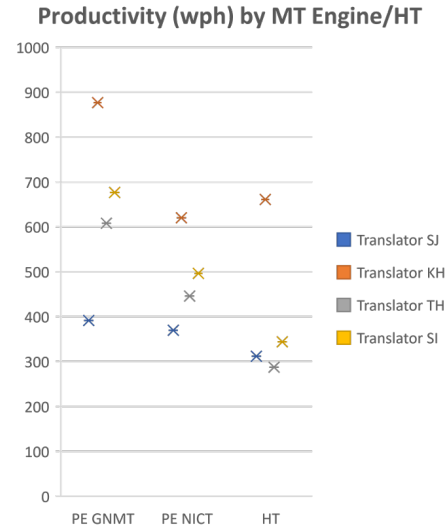


RESULTS—PRODUCTIVITY (CONTD.)

- PE GNMT > PE NICT > HT
—The same tendency observed almost **independent of the translator**
- **Correlation ratio** between **Productivity** and **MT Engine/HT**:
 $\eta = 0.57$

$$\eta := \sqrt{\frac{\sum_{i=1}^n n_i (\bar{x}_i - \bar{x})^2}{\sum_{i=1}^n \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2}}$$

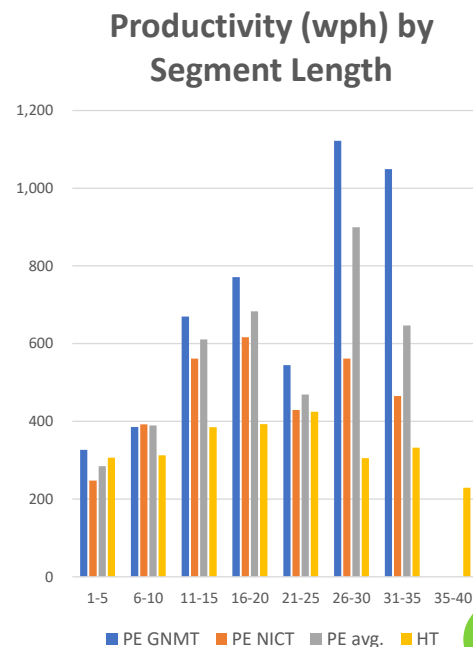
($0 \leq \eta \leq 1$)



RESULTS—PRODUCTIVITY (CONTD.)

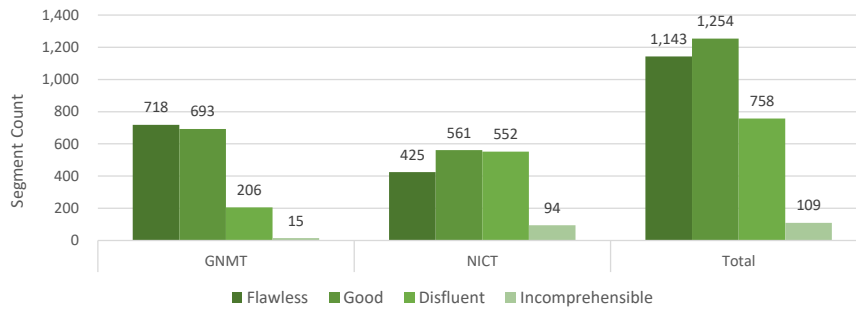
Other observations:

- **No** apparent correlation observed between **Productivity** and **Segment Length** (word count of each segment)
- In particular, in **HT**, SL does not seem to affect Productivity at all
- **GNMT** seems to show a slight tendency that the **longer SL**, the **higher productivity**, but it's not significant

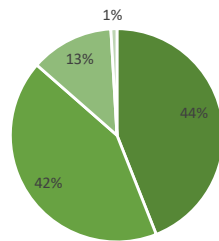


RESULTS—QUALITY: FLUENCY

Fluency Evaluation Results

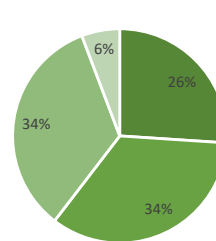


Fluency: GNMT



Flawless Good Disfluent Incomprehensible

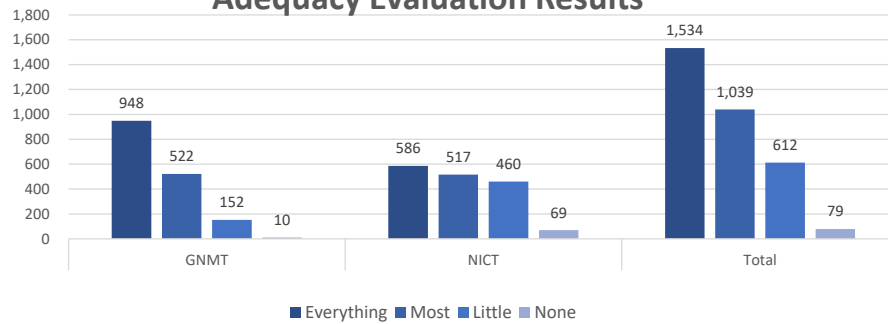
Fluency: NICT



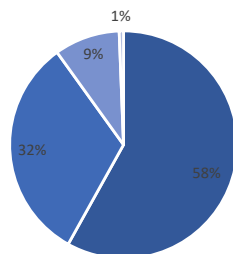
Flawless Good Disfluent Incomprehensible

RESULTS—QUALITY: ADEQUACY

Adequacy Evaluation Results

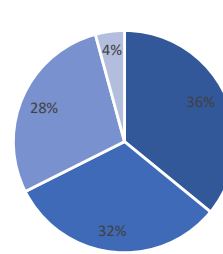


Adequacy: GNMT



Everything Most Little None

Adequacy: NICT



Everything Most Little None

RESULTS—QUALITY

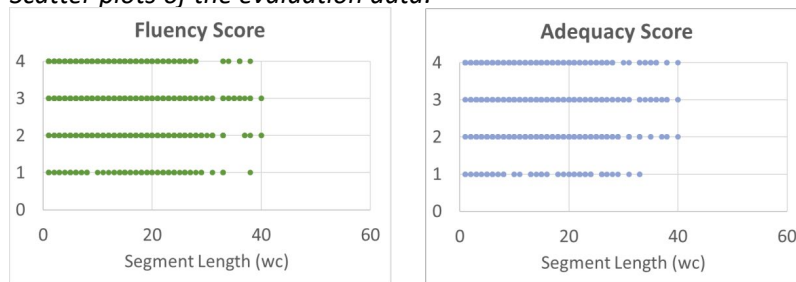
○ Key findings

- **GNMT** had better scores overall, where
 - +85% segments had **Flowless (4) or Good (3) fluency**
 - +90% segments had **Everything (4) or Most (3) adequacy**

○ Other observations

- Almost **no** correlation observed between **Segment Length** and **Quality** of MT outputs in our pilot:

Scatter plots of the evaluation data:



CONCLUSION

Productivity gains

- We observed **63%** average productivity gains in **PE w/ GNMT** as well as **30%** gains in **PE w/ NICT**.
- This strongly suggests that significant improvement in efficiency can be achieved in most E2J technical localization projects by utilizing the latest MT engines, in particular, GNMT, in the translation process.

Other findings

- In our pilot, we didn't observe the tendency "Longer sentences give worse MT outputs, thus result in lower PE productivity", which may be a myth.

The Impact of MT Quality Estimation on Post-Editing Effort

Carlos Teixeira **Sharon O'Brien**
carlos.teixeira@dcu.ie sharon.obrien@dcu.ie

Centre for Translation and Textual Studies (CTTS)
ADAPT Centre for Digital Content Technology
Dublin City University (Ireland)

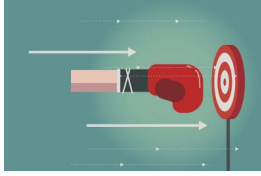


MOTIVATION

- Professional translators edit suggestions coming from translation memories (TM) and machine translation (MT)
- Handling those two **types of linguistic support** requires different strategies
- TM suggestions incorporate metadata to increase efficiency and quality (e.g. Fuzzy Match scores)
- QE scores are an attempt to provide **relevant metadata** for MT suggestions

Novelty: Despite recent advances in QE research, little is known about the **real impact of QE scores** on the translation process.

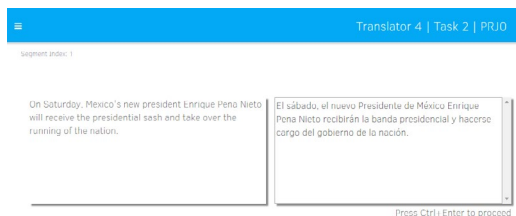
POTENTIAL IMPACT



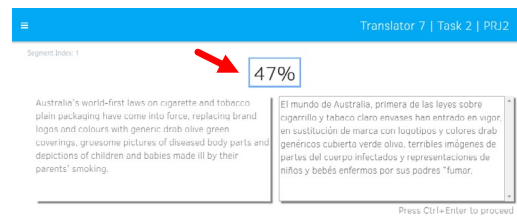
- Improve translators' efficiency when working with MT
- Reduce cognitive strain on translators
- Increase translators' trust in MT output by offering accurate QE
- Reduce their need to search for validation from additional, external resources (cf. Bundgaard 2017, Daems et al. 2016)

EXPERIMENT DESIGN

- Online post-editing tool (HandyCAT)



Only source text and MT displayed



Quality estimation (QE) scores displayed

- Participants: 20 professional translators
- Materials: 4 texts (WMT13 news material)
- Languages: English → Spanish
- Four different QE modes (more details below)

EXPERIMENT DESIGN (cont.'d)

QE mode consists of two parts:

- **Score Type:**
 - No QE: the QE box is hidden in HandyCAT
 - Accurate QE: scores obtained from the automatic scoring system that ranked best in the WMT13 shared task (automatic, accurate)
 - Inaccurate QE: 'random' scores (automatic, inaccurate)
 - Human QE: scores obtained using a human evaluation method (human, accurate) (Graham et al. 2015)
- **Score Level:** Percentage (between ~20% and 99 %)

EXPERIMENT DESIGN (cont.'d)

Research question:

- What is the impact of the different modes of QE scores on:
 - temporal effort (time spent)
 - physical effort (number of keystrokes)
 - cognitive effort (gaze behaviour)

Data collection:

- activity logging
- screen recording
- eye tracking

EXPERIMENT DESIGN (cont.'d)

Full range of variables being considered:

Role	Name	Type	Measurement / Levels
Dependent	Temporal – Translation time	numeric	seconds per word
	Physical – Amount of typing		keys per word
	Fixation count		n per word
	Cognitive – Fixation duration		seconds per word
	Pupil dilation		mm (variance)
Independent (Fixed effects)	Primary	QE score type	No_QE, Acc_QE, Inacc_QE, Human_QE
		QE score level	N/A (No_QE condition) L0: 0.1 to 19.9% L2: 20 to 39.0% L4: 40 to 59.9% L6: 60 to 79.9% L8: 80% to 100%
	Secondary	Document	SRC1, SRC2, SRC5, SRC7
		Task order	T01, T02, T03, T04

RESULTS - Temporal effort

(time spent per word)

Fixed Effects

Target:log_Time

Source	F	df1	df2	Sig.
Corrected Model ▼	20.880	12	1,027	.000
Score_Type	0.035	2	1,027	.965
Score_Level_Ordinal	14.049	3	1,027	.000
Document	34.544	3	1,027	.000
Task_Order	1.950	3	1,027	.120

Primary variables

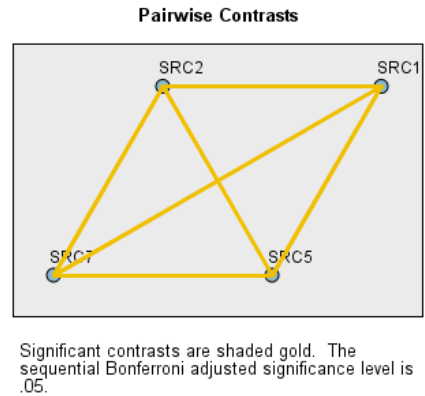
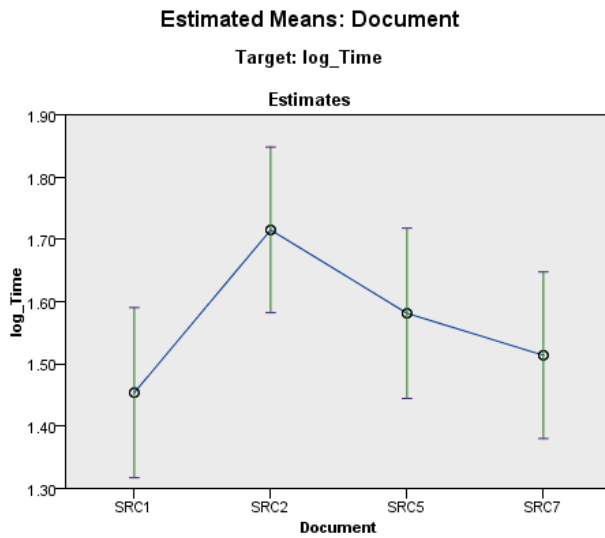
Secondary variables

Probability distribution:Normal
Link function:Identity



RESULTS - Temporal effort

Effects found for Document

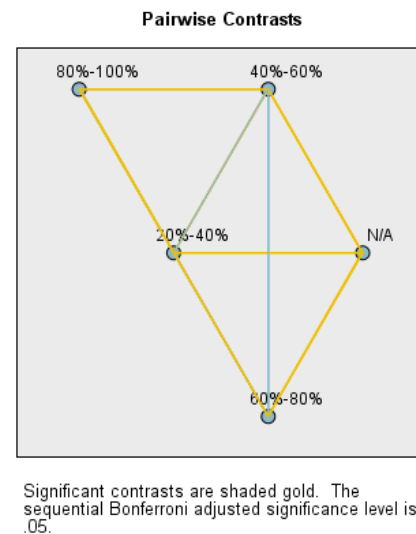
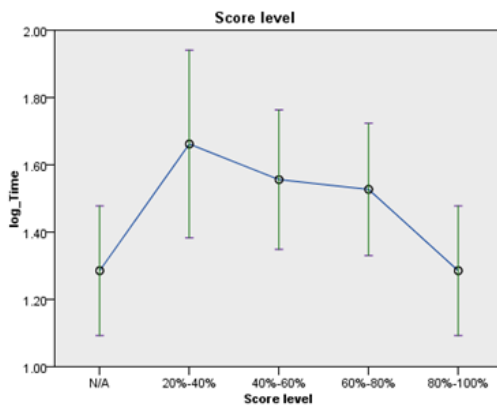


RESULTS - Temporal effort

Effects found for Score Level

Estimated Means: Significant Effects
Target: log_Time

Estimated means charts for significant effects ($p < .05$) are displayed. Up to ten effects are displayed, beginning with the top three-way effects. Effects shown contain categorical predictors only.



RESULTS – Physical effort

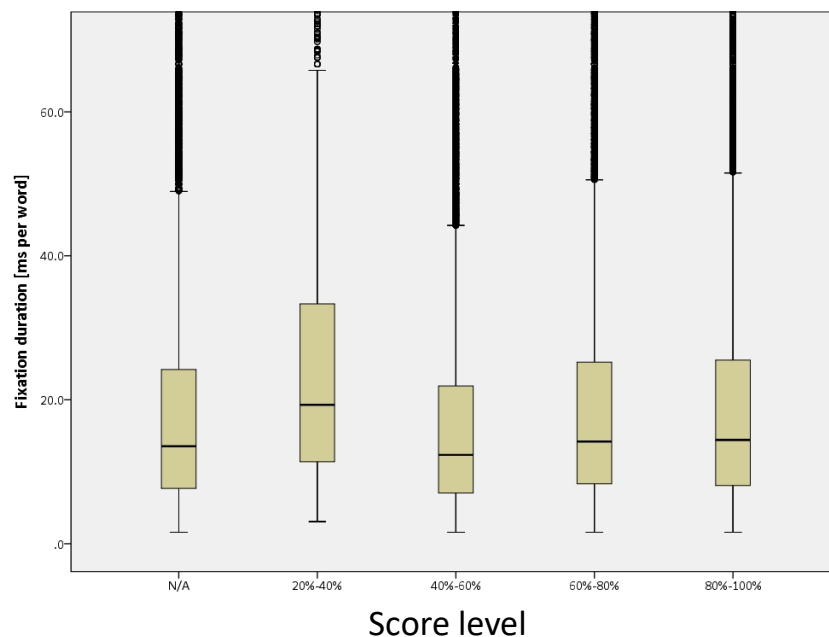
(# of keys typed per word)

Results are similar to the ones found for Temporal effort:

- **No significant** differences in average # of keys according to **Score Type**
- **Significant** differences in average # of keys according to **Score Level**

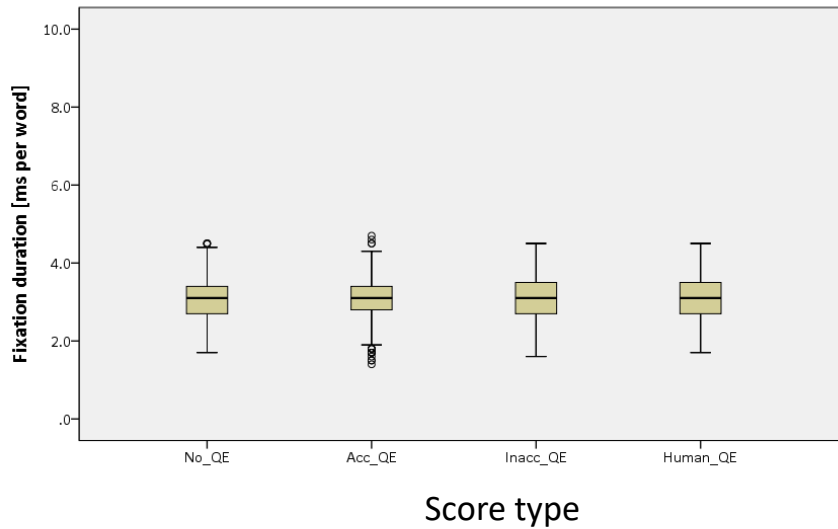
RESULTS – Cognitive effort

Fixation duration per Score Level – No significant effects found



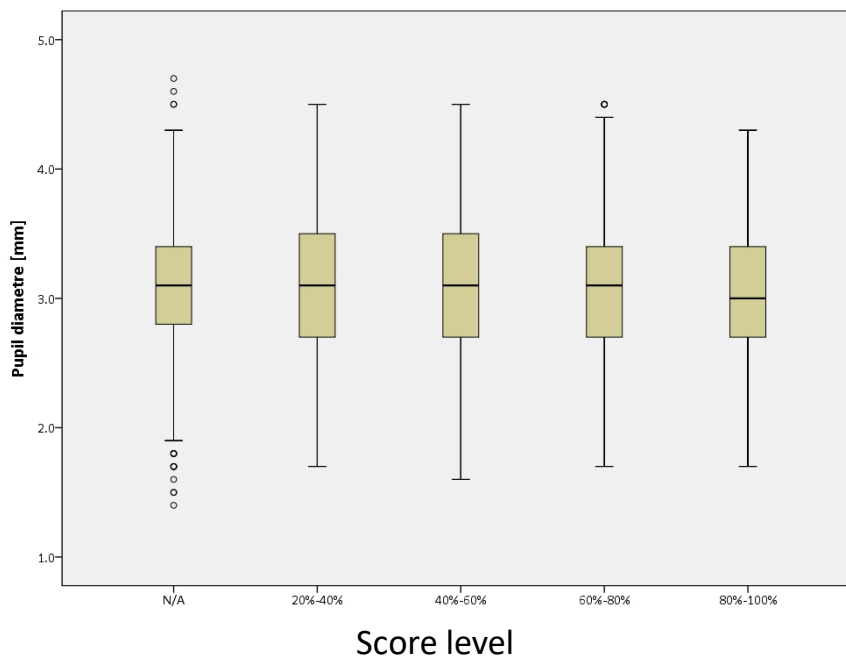
RESULTS – Cognitive effort

Fixation duration per Score Type – No significant effects found



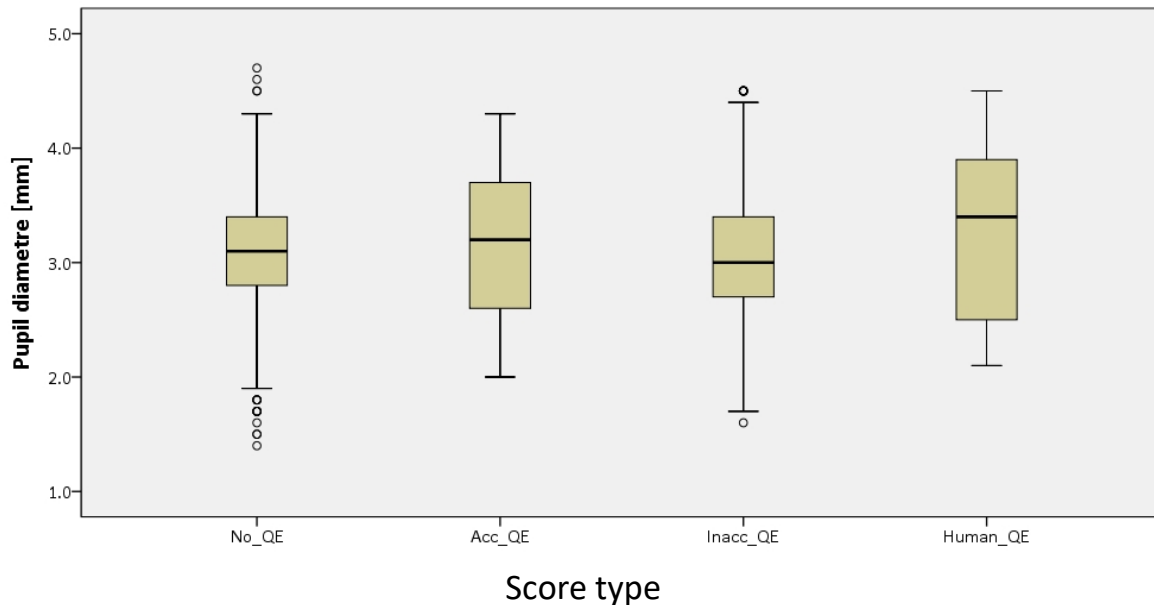
RESULTS – Cognitive effort

Pupil diameter per Score Level – No significant effects found



RESULTS – Cognitive effort

Pupil diameter per Score Type – No significant effects found



SUMMARY

Our results indicate:

- No significant effect of **Score Type** on either time or edits.
- A significant effect of **Score Level** on both time and edits:
The higher the score level the less time is spent and the fewer keys are typed (regardless of how the scores were calculated!)
- Displaying QE scores (even if they are accurate) is not necessarily better than displaying no scores.
- No significant variations in the number of fixations, fixation duration or pupil size that could be associated with the display of QE scores.

DISCUSSION

- In our experiment, only a QE percentage was displayed.
- Perhaps the same results would have been found for TM if we had removed the *diff* indication and just left the Match percentages?

DISCUSSION (cont.'d)

The screenshot shows the SDL Trados Studio interface. The main window displays a translation comparison between English and German. The English text is "Finding a location for your photo printer" and the German text is "Geeigneten Aufstellungsort für Ihren Fotodrucker finden". The match percentage is 91%. A red circle highlights the 91% match percentage, and a blue circle highlights the question mark in the English text. The interface also shows a "Term Recognition" panel on the right with the term "photo printer" and its German equivalent "Fotodrucker".

DISCUSSION (cont.'d)

- Our results point toward the need to combine QE scores with the display of phrase-level or word-level QE indication.

This is what we displayed:

The screenshot shows a translation interface with a blue header bar containing a menu icon and the text "Translator 7 | Task 2 | PRJ2". Below the header, the text "Segment Index: 1" is visible. A blue box displays "81%". The source text on the left is "The Army intelligence analyst, arrested in June 2010, is accused of stealing thousands of classified documents while serving in Iraq." The target text on the right is "El ejército analista de inteligencia, detenido en junio de 2010, es acusado de robar miles de documentos clasificados aunque sirven en Iraq." At the bottom right, there is a prompt: "Press Ctrl+Enter to proceed".

DISCUSSION (cont.'d)

- Our results point toward the need to combine QE scores with the display of phrase-level or word-level QE indication.

This might be the way forward to make QE more effective:

The screenshot shows a translation interface similar to the previous one. The source text is "The Army intelligence analyst, arrested in June 2010, is accused of stealing thousands of classified documents while serving in Iraq." The target text is "El ejército analista de inteligencia, detenido en junio de 2010, es acusado de robar miles de documentos clasificados aunque sirven en Iraq." In this version, the phrases "El ejército analista de inteligencia" and "aunque sirven" are highlighted in orange in the target text. The quality score "81%" is shown in a blue box. The prompt "Press Ctrl+Enter to proceed" is at the bottom right.

FUTURE RESEARCH

- Test the effect of word-level or phrase-level QE indicators
- Test different layouts for the presentation of QE information
- Try more fine-grained buckets of QE score levels to identify ideal cut-off point
- Assess the effect of QE on the Quality of the final translations
- Study the impact of QE if translators learned to trust the information (longitudinal)

REFERENCES

- Bundgaard, Kristine. 2017. *(Post-)editing - A Workplace Study of Translator-Computer Interaction at Textminded Danmark A/S*. Doctoral thesis: Aarhus University.
- Daems, Joke, Michael Carl, Sonia Vandepitte, Robert Hartsuiker, and Lieve Macken. 2016. "The Effectiveness of Consulting External Resources During Translation and Post-editing of General Text Types". In *New directions in empirical translation process research: Exploring the CRITT TPR-DB*. [New frontiers in translation studies], Michael Carl, Srinivas Bangalore and Moritz Schaeffer (eds.) Cham: Springer.
- Graham, Yvette, Nitika Mathur and Timothy Baldwin. 2015. "Accurate evaluation of segment-level machine translation metrics." In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics Human Language Technologies*, Denver, Colorado.
- Hokamp, Chris and Qun Liu. 2015. "HandyCAT: The Flexible CAT Tool for Translation Research." Demo presented at EAMT 2015, May 15-19, Istanbul, Turkey.

Thank you!

ありがとうございました

Carlos Teixeira
carlos.teixeira@dcu.ie

Sharon O'Brien
sharon.obrien@dcu.ie



Utilizing Neural MT Engines in Industrial Translation

Toru Shishido

HUMAN SCIENCE

Human Science Co., Ltd.
WWW.SCIENCE.CO.JP



Agenda

- ◆ Evaluation
 - ◆ Raw Output Quality
 - ◆ Throughput PE vs HT
- ◆ Challenges / Best Practices
- ◆ Key Takeaways
- ◆ Q&A

HUMAN SCIENCE
WWW.SCIENCE.CO.JP



Raw Output Quality Evaluation – Details of sample

- ◆ Human assessment
- ◆ Language pair: English-Japanese
- ◆ Translation volume: 3786 words
- ◆ Content type: Software manual



Raw Output Quality Evaluation – Human assessment / how to score

Better



Worse

Meaning and Accuracy

1. Perfectly Understandable

2. Fully Understandable

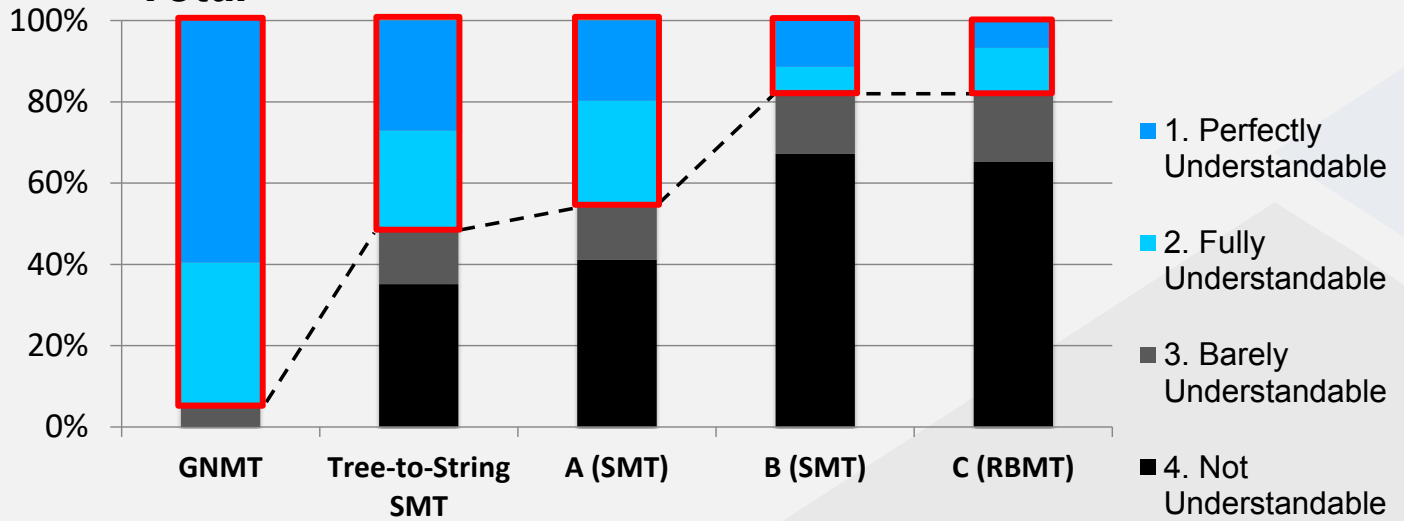
3. Barely Understandable

4. Not Understandable



Raw Output Quality Evaluation – Results

-Total-

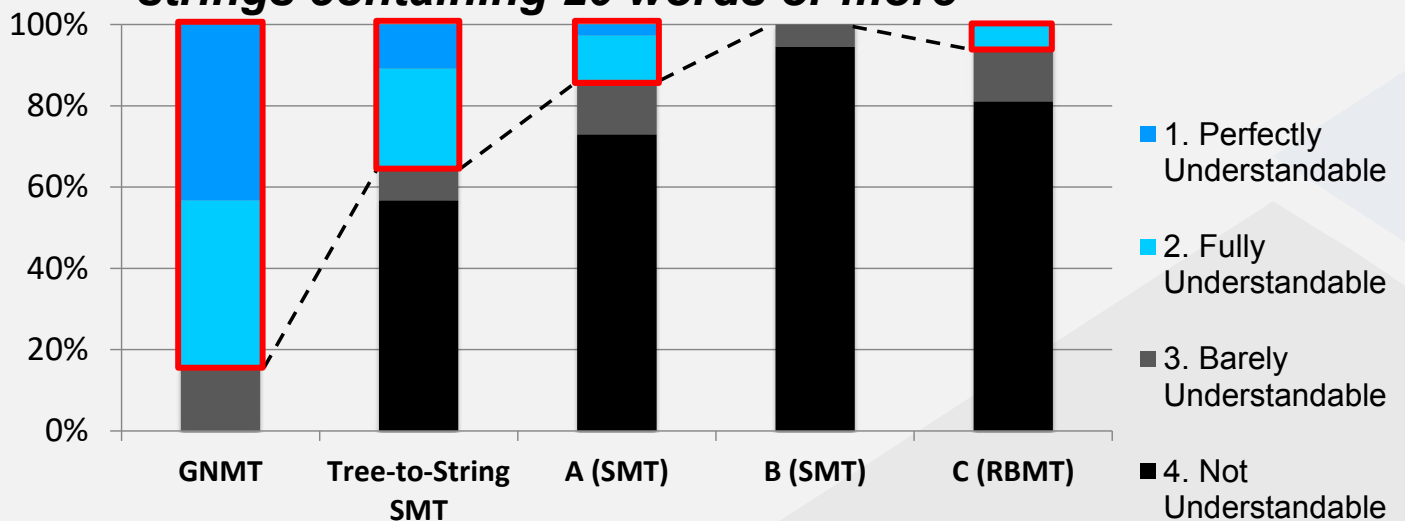


HUMAN SCIENCE
WWW.SCIENCE.CO.JP



Raw Output Quality Evaluation – Results

-strings containing 20 words or more-



HUMAN SCIENCE
WWW.SCIENCE.CO.JP



Raw Output Quality Evaluation – Analysis

- ◆ GNMT is extremely good in translating software manuals
- ◆ The quality of GNMT is high even with long sentences
- ◆ Reasons (speculations):
 - ◆ There are large amount of software manuals on the Internet
 - ◆ Google crawls the Internet for its training corpus
 - ◆ GNMT is like an MT system with a huge translation memory from multiple software vendors



Throughput Evaluation - Details

- ◆ Localization projects of various document types
- ◆ Not in production but completely simulated

(As of July 28, 2017)	Source volume	Number of projects
English-Japanese	49,883 weighted words	36
Japanese-English	10,057 weighted characters	2



Throughput Evaluation - Context levels

- ◆ Translation depends on the information outside the sentence
 - ◆ Other sentences in the document
 - ◆ Basic knowledge of the products / services
 - ◆ Common sense
- LOW: A sentence provides all the information for translation.
- MEDIUM
- HIGH: Information from other sources is needed.



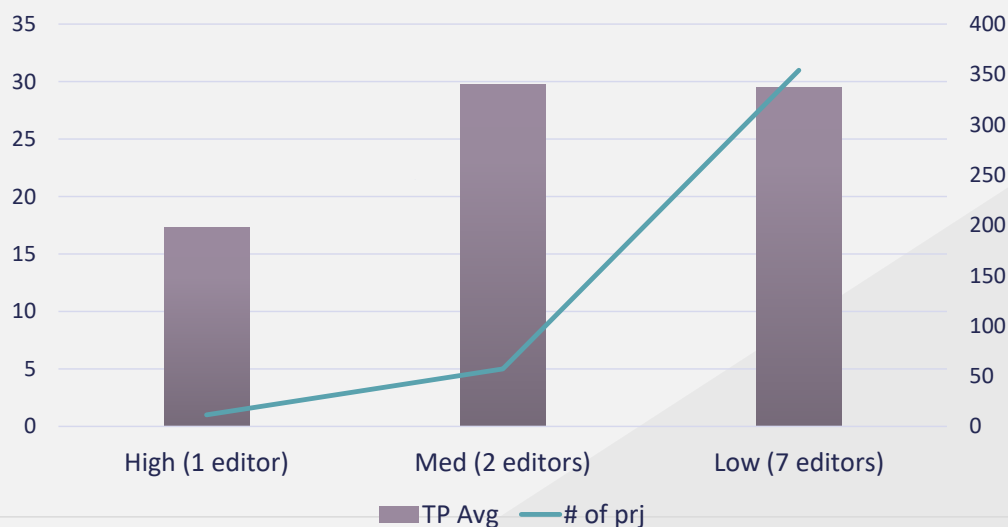
Throughput Evaluation - Results: English-Japanese

Weighted word count	PostEdit time (hr)	Speed (w/hr)	Content type	Context level
2174	11	197.6	Training role play scripts	HIGH
1229	4	307.3	Resource file	Faster than human translation (~250w/hr)
1108	3.5	316.8	FAQ (web services)	
3175	10	317.5	Product information	
1682	4	420.5	FAQ (web services)	LOW
1482	3	494.0	Service description	LOW
1023	2	615.0	Software manual	LOW



Throughput Evaluation - Results : English-Japanese

Throughput average by context level



Throughput Evaluation - Results: Japanese-English

Weighted char count	PostEdit time (hr)	Speed (ch/hr)	Content type	Context level
9352	4	2338.0	Whitepaper	LOW
705	0.33	2350.0	Developer page (UGC)	MEDIUM

Needs to collect more data, but
much faster than manual translation (~500ch/hr)
+ good for UGC



Challenges – Fun Fact

- ◆ There are 24 spelling patterns for the translation of **User Interface**:

ユーザーインターフェース	ユーザーインタフェース	ユーザーインターフェイス	ユーザーインタフェイス
ユーザインターフェース	ユーザインタフェース	ユーザインターフェイス	ユーザインタフェイス
ユーザー▲インターフェース	ユーザー▲インタフェース	ユーザー▲インターフェイス	ユーザー▲インタフェイス
ユーザ▲インターフェース	ユーザ▲インタフェース	ユーザ▲インターフェイス	ユーザ▲インタフェイス
ユーザー・インターフェース	ユーザー・インタフェース	ユーザー・インターフェイス	ユーザー・インタフェイス
ユーザ・インターフェース	ユーザ・インタフェース	ユーザ・インターフェイス	ユーザ・インタフェイス

▲ stands for a single-byte space.

HUMAN SCIENCE
WWW.SCIENCE.CO.JP



Challenges (1) – Following rules in style guides (1)

- ◆ Most of companies have their own style guides and the rules are slightly different, such as spacing rules, brackets, long vowels (*cho-on*), etc.

Spacing rules	Company A	Company B	Company C
Katakana words	User interface ユーザー▲インターフェイス	User interface ユーザインタフェース	User interface ユーザー・インターフェイス
Between single-byte and double-byte characters	From Sept. 19 to 21 9▲月▲19▲日～▲21▲日	From Sept. 19 to 21 9月19日～9月21日	From Sept. 19 to 21 9▲月▲19▲日～21▲日

HUMAN SCIENCE
WWW.SCIENCE.CO.JP



Challenges (1) – Following rules in style guides (2)

- ◆ Most of companies have their own style guides and the rules are slightly different, such as spacing rules, brackets, long vowels (*cho-on*), etc.

	Company A	Company B	Company C
Brackets	Use [] (single-byte) for user interface terms Use 『 』 for book titles and use 「 」 for chapter/section titles	Use [] (double-byte) for user interface terms Use 『 』 for book, chapter and section titles	Use 「 」 (double-byte) for user interface terms Use 『 』 for book, chapter and section titles
Long vowels (<i>cho-on</i>)	User ... ユーザー Printer ... プリンター Programmer ... プログラマー (depends of numbers of syllables)	User ... ユーザー Printer ... プリンター Programmer ... プログラマ	User ... ユーザ Printer ... プリンタ Programmer ... プログラマ

Challenges (2) – Tone (*de-arū* vs *desu-masu* (常体/敬体))

- ◆ There are two major writing styles in Japanese, *de-arū* style vs *desu-masu* style. These styles should be applied appropriately to match the context.

	Source	Raw MT	Post edited
<i>de-arū</i> style (常体)	(This course helps you to:) • Use new services and features from the ABC product to learn about modern technologies.	• ABC 製品の新しいサービスと機能を使用して、最新の技術を学ぶことができます。	• ABC 製品の新しいサービスと機能を使用して、最新の技術について学習する。 (e.g., bullet items)
<i>desu-masu</i> style (敬体)	Use new services and features from the ABC product to learn about modern technologies.	ABC 製品の新しいサービスと機能を使用して、最新の技術を学ぶことができます。	ABC 製品の新しいサービスと機能を使用すると、最新の技術を学ぶことができます。 (e.g., normal texts)

Challenges (3) – Glossary (UI terms / client specific / titles of references)

- ◆ Most companies have UI glossaries and terminologies so the post editors need to apply the appropriate terms.

	Source	Raw MT	Post edited
Example 1	Click on the “Continue” button.	「続行」ボタンをクリックします。	[Continue (続行)] ボタンをクリックします。
Example 2	a getting started guide	スタートガイド	入門ガイド
Example 3	詳細については、「APIを使用した展開」を参照してください。	For details, see “Deployment using API”.	For details, see “Deploying with API”.



Challenges (4) – General terms (Contexts/Inconsistencies)

- ◆ Post editors need to apply the correct translations to context-sensitive terms.
- ◆ Even the translations are correct, they must be consistent.

	Source	Raw MT	Consider when post editing
Example 1	available	利用可能 (<i>able to use</i>) ご利用いただけます (<i>polite “able to use”</i>) あります (<i>exists / be in stock</i>)	Post editors must consider the context of the text since the MT engines do not see the context.
Example 2	question	質問 (<i>an act of asking</i>) 問題 (<i>a problem</i>) 疑わしいこと (<i>a doubt</i>)	
Example 3	server-side	サーバーサイド (<i>server side</i>) サーバー側 (<i>server side</i>)	Both translations are correct, but inconsistent.



Challenges (4) – General terms (new words/buzzwords)

- ◆ Some new words may not be translated correctly sometimes.

	Source	Raw MT	Post edited
deep dive	The XXX Conference is a one-day deep dive into new technology.	XXX Conference は、新たな技術についての 深いダイビング です。 (a recreational diving)	XXX Conference は、新たな技術について考える 1 日間の ディープダイブ (or 分析ワークショップ) です。 (an extensive analysis)
DevOps	DevOps focuses on improving automation.	開発部門 は自動化の改善に重点を置いています。 (Development Dept.)	DevOps では自動化の改善に重点を置きます。

Challenges (5) – Tags / variables

- ◆ In most cases, tags are not properly treated. Also, tags can cause poor translation.

	Source	Raw MT	Post edited
 tag	Cover letter	Coverletter (the tag is omitted)	カバー レター
variable tag	Please ¥{0¥} to try again.	再試行するには ¥ ▲ {0 ▲ ¥} してください。 (unnecessary spaces)	¥{0¥}して、もう一度お試しください。

Challenges and solutions

Issues	Solutions	Can be fixed automatically?
Client-specific style specifications	Apply the rules with regular expression	Some yes, others no
Tone	Check and replace manually in Post Edit	No
Terminology (UI / client-specific / ref mat titles)	Apply some translations from terminology file automatically, and then replace manually in Post Edit (if necessary)	Some yes, others no
Terminology (general/new terms)		
Tags / variables	Delete before MT and insert manually in Post Edit	No

Best Practices

- ◆ Decide the content type to be machine-translated
 - Manuals, user interface, FAQ, UGC, marketing contents
- ◆ Align the final expectations between client and LSP
 - Final quality of translation, TATs, costs, content cycles
- ◆ Then, support and train post editors
 - Appropriate allocation of post editors by content type and final quality expectation, pre-process with SW components, continuous feedback loop

Takeaways - Neural MT for Commercial Use

- ◆ NMT makes the translation hours 1.36x faster and the productivity 1.48x higher (evaluation average)
- ◆ Usable in production both in English-Japanese and Japanese-English pairs in IT localization (incl. UGC)
- ◆ There are issues to be solved manually in Post Edit, but some can be automatically processed with software components

HUMAN SCIENCE
WWW.SCIENCE.CO.JP



Q&A



t-shishido@science.co.jp



www.science.co.jp



+81-3-5321-3111

HUMAN SCIENCE
WWW.SCIENCE.CO.JP



Comparative Evaluation of NMT with Established SMT Programs

Lena Marg
Naoko Miyazaki
Elaine O'Curran
Tanja Schmidt

welocalize 
doing things differently

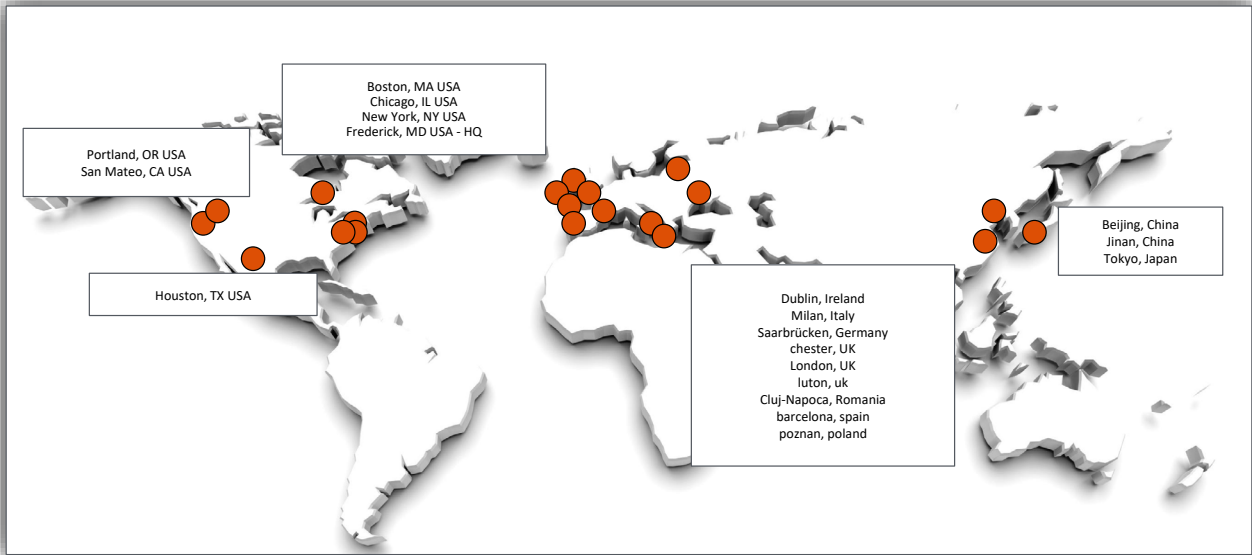
AGENDA

1. Objective
2. Scope of the evaluation
 - Language pairs
 - Content types
 - Size and integrity of the test sets
3. Evaluation methodologies
 - Human evaluations
 - Automatic scoring
4. Results
5. Conclusions

welocalize 
doing things differently

WHO + WHERE WE ARE

WORDS TRANSLATED 2015: 1.15 BILLION
LANGUAGES TRANSLATED: 175+
EMPLOYEES: 1000+
GLOBAL OFFICES: 21
7TH LARGEST PROVIDER IN THE WORLD
4TH LARGEST LSP IN THE US
*2016 Common Sense Advisory



Objective



Objective

- Compare the performance of two public NMT systems with a customized SMT solution that is applied in production for two enterprise-level clients.
- Evaluate how generic NMT performs out-of-the-box for different languages and content types that are in high demand in our industry.
- Enable us to make well-founded business decisions as we move forward with our MT strategy.
- Provide data-driven advice and support to our clients.



Scope of the Evaluation



Sampling and Sample Size

Evaluation Type	Sample Size (TUs)	Sample Origin
Autoscoring (HT)	Approx. 2500	This is the randomized, blind test set taken from the customized SMT engine. The segments in the test set are not included in the engine's training data and originate from production TMs.
Side-by-side engine ranking	200	The 200 segments for human evaluation are randomly selected from the 2500 TU test set described above
Adequacy and Fluency scoring	100	From the 200 segments above, we randomly selected 100 segments for the more detailed human analysis and post-editing sample
Strength and Weaknesses Assessment	100	Same sample as above
Autoscoring (PE)	100	Same sample as above is post-edited and scored



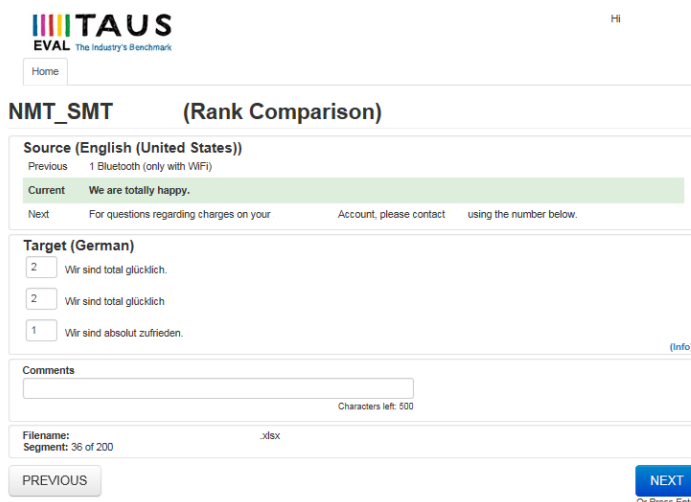
Scope Overview

Evaluation Type	MT Systems	Content Type	Language Pairs	Evaluators
Autoscoring (HT)	Customized SMT, Generic1 NMT, Generic2 NMT	Light Marketing, Technical Documentation	de-DE, fr-FR, ja-JP, pt-BR, ru-RU, zh-CN	Proprietary scoring tool (wescore)
Side-by-side engine ranking	Customized SMT, Generic1 NMT, Generic2 NMT	Light Marketing, Technical Documentation	de-DE, fr-FR, ja-JP, pt-BR, ru-RU, zh-CN	Two evaluators: one account translator, one experienced MT evaluator
Adequacy and Fluency scoring	Customized SMT, Generic2 NMT	Light Marketing, Technical Documentation	de-DE, ja-JP, pt-BR	One evaluator: account translator
Strength and Weaknesses Assessment	Customized SMT, Generic2 NMT	Light Marketing, Technical Documentation	de-DE, ja-JP, pt-BR	One evaluator: account translator
Autoscoring (PE)	Customized SMT, Generic1 NMT	Light Marketing	de-DE, ja-JP, pt-BR	One evaluator: account translator



Side-by-Side Engine Ranking

- The TAUS DQF tool used for this evaluation randomizes the order in which the target segments from the engines being compared are presented. This means the evaluator(s) do not get conditioned into giving anticipated rankings
- Ranking (1,2,3) of the 3 engines, from best to worst
- Allows equal ranking of two or three outputs



The screenshot shows the TAUS EVAL interface for a rank comparison task. The header includes the TAUS logo and the text "Hi". Below the header is a "Home" button. The main title is "NMT_SMT (Rank Comparison)". The interface is divided into two main sections: "Source (English (United States))" and "Target (German)".

Source (English (United States))
Previous: 1 Bluetooth (only with WiFi)
Current: We are totally happy.
Next: For questions regarding charges on your Account, please contact using the number below.

Target (German)
2: Wir sind total glücklich.
2: Wir sind total glücklich
1: Wir sind absolut zufrieden. [\(Info\)](#)

Comments: Characters left: 500

Filename: .xlsx
Segment: 36 of 200

Navigation buttons: PREVIOUS, NEXT (Or Press Enter)



Adequacy and Fluency Scoring

Adequacy Score Evaluation Criteria	
5	All meaning expressed in the source appears in the translation. You do not need to refer to the source to understand the meaning.
4	Most of the source meaning is expressed in the translation. You can understand most of the meaning without referring to the source.
3	Much of the source meaning is expressed in the translation. Roughly half the MT output can be understood without referring to the source.
2	Little of the source meaning is expressed in the translation. Although you can guess fractions of the MT output, you cannot understand it without referring to the source.
1	None of the meaning expressed in the source is expressed in the translation. You cannot make any sense of the MT output alone AND/OR the MT output says exactly the opposite of the source.

Fluency Score Evaluation Criteria	
5	Native language fluency. No grammar errors, good word choice and syntactic structure. No PE required.
4	Near native fluency. Few terminology or grammar errors which don't impact the overall understanding of the meaning. Little PE required.
3	Not very fluent. About half of translation contains errors and requires PE.
2	Little fluency. Wrong word choice, poor grammar and syntactic structure. A lot of PE required.
1	No fluency. Absolutely ungrammatical and for the most part doesn't make any sense. Translation has to be re-written from scratch.



Ranking Strengths and Weaknesses

WHICH TRANSLATION IS BETTER WITH REGARD TO:
accuracy (accurate rendition of source meaning)
fluency & style
general domain terminology
client-specific terminology & instructions
completeness (all key information from source is rendered)
redundancy (translation contains additional information not contained in the source)
syntax
grammar
localization (correct format of punctuation; spacing; dates & time, units measurement)
tags & placeholders
spelling
Other



Autoscoring

- BLEU
- NIST
- METEOR
- GTM
- Precision
- Recall
- TER
- PE Distance*

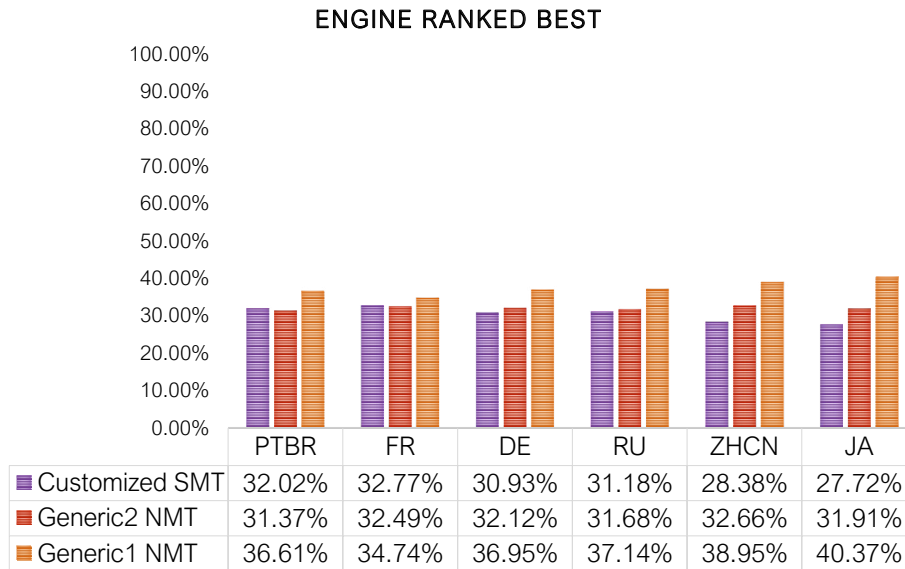
*In our analysis we focus on PE distance, which applies the Levenshtein algorithm and is character-based. Compared to word-based scoring, this method captures morphological post-edits, such as fixing word forms, and we have found it to correlate well with human judgment.



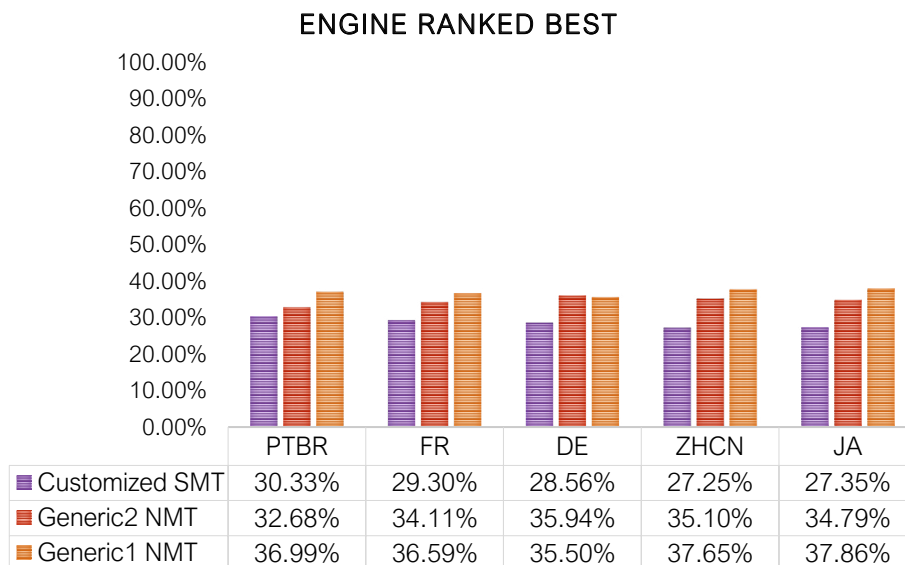
Results



Engine Ranking Results for Light Marketing



Engine Ranking Results for Technical Documentation



German Results

Locale	Evaluation	Light Marketing Content				Technical Documentation Content				
		Generic1 NMT	Generic2 NMT	Customized SMT	Diff Best NMT & SMT	Generic1 NMT	Generic2 NMT	Customized SMT	Diff Best NMT & SMT	
de-DE	Ranking	√	2	3	6.02 pp	2	√	3	7.38 pp	
	Adequacy			√	0.06		√		0.08	
	Fluency		√		0.07		√		0.45	
	Accuracy						√			
	Fluency & Style		√				√			
	Syntax		√				√			
	Grammar		√				√			
	Terminology			√						
	Completeness			√			√			
	Localization			√						
	Edit Distance (HT)		2	3	√	3.32 pp	√	3	2	1.12 pp
	Edit Distance (PE)		2		√	1.55 pp				



welocalize 
doing things differently

Japanese Results

Locale	Evaluation	Light Marketing Content				Technical Documentation Content				
		Generic1 NMT	Generic2 NMT	Customized SMT	Diff Best NMT & SMT	Generic1 NMT	Generic2 NMT	Customized SMT	Diff Best NMT & SMT	
ja-JP	Ranking	√	2	3	12.96 pp	√	2	3	10.51 pp	
	Adequacy		√		0.32		√		0.76	
	Fluency		√		0.2		√		0.49	
	Accuracy		√				√			
	Fluency & Style		√				√			
	Completeness		√				√			
	Syntax		√				√			
	Grammar		√				√			
	Terminology			√				√		
	Spelling							√		
	Edit Distance (HT)		√	3	2	8.17 pp	√	3	2	5.79 pp
	Edit Distance (PE)		√		2	21.07 pp				



welocalize 
doing things differently

Brazilian Results

Locale	Evaluation	Light Marketing Content				Technical Documentation Content				
		Generic1 NMT	Generic2 NMT	Customized SMT	Diff Best NMT & SMT	Generic1 NMT	Generic2 NMT	Customized SMT	Diff Best NMT & SMT	
pt-BR	Ranking	√	3	2	4.59 pp	√	2	3	6.65 pp	
	Accuracy		√		0.09		√		0.26	
	Fluency		√		0.45		√		0.28	
	Accuracy						√			
	Fluency & Style						√			
	Completeness		√				√			
	Redundancy		√							
	Syntax		√				√			
	Grammar		√				√			
	Terminology			√						
	Localization			√						
	Tags & Placeholders			√						
	Edit Distance (HT)		2	3	√	1.68 pp	√	3	2	0.28 pp
	Edit Distance (PE)		2		√	3.62 pp				



welocalizeQ
doing things differently

French, Russian, Simplified Chinese Results

Locale	Evaluation	Light Marketing Content				Technical Documentation Content			
		Generic1 NMT	Generic2 NMT	Customized SMT	Diff Best NMT & SMT	Generic1 NMT	Generic2 NMT	Customized SMT	Diff Best NMT & SMT
fr-FR	Ranking	√	3	2	1.97 pp	√	2	3	7.29 pp
	Edit Distance (HT)	2	3	√	2.02 pp	2	3	√	0.62 pp
zh-CN	Ranking	√	2	3	10.57 pp	√	2	3	10.40 pp
	Edit Distance (HT)	√	3	2	5.87 pp	√	3	2	3.12 pp
ru-RU	Ranking	√	2	3	5.95 pp				
	Edit Distance (HT)	2	3	√	1.58 pp				



welocalizeQ
doing things differently

SUMMARY

- All evaluators prefer generic NMT during side-by-side ranking, the first evaluation task.
- NMT also wins Adequacy & Fluency scoring with the exception of German Adequacy for Light Marketing.
- Evaluators for JA, DE, PTBR overall prefer customized SMT for terminology and localization-related issues, but NMT for fluency, style, grammar and syntax. JA also prefers NMT for accuracy.
- NMT outperforms SMT more consistently on Technical Documentation than on Light Marketing.
- For Technical Documentation the autoscores favor NMT, while they show mixed results for Light Marketing.
- After completing the post-editing task on Light Marketing, the German and Brazilian translators had a slight preference for SMT, contradicting the previous human evaluation results and indicating that the autoscores may be more accurate.



welocalizeO
doing things differently

SUMMARY

- The most significant quality improvement with NMT are for Chinese and Japanese
- For the other languages, the quality differences between NMT and SMT are less pronounced

Locale	Evaluation	Light Marketing				Technical Documentation			
		Generic NMT1	Generic NMT2	Customized SMT	Diff Best NMT & SMT	Generic NMT1	Generic NMT2	Customized SMT	Diff Best NMT & SMT
de-DE	Ranking	√	2	3	6.02 pp	2	√	3	7.38 pp
	Accuracy			√	0.06		√		0.08
	Fluency		√		0.07		√		0.45
	Edit Distance	2	3	√	3.32 pp	√	3	2	1.12 pp
	Edit Distance (PE)	2		√	1.55 pp				
fr-FR	Ranking	√	3	2	1.97 pp	√	2	3	7.29 pp
	Edit Distance	2	3	√	2.02 pp	2	3	√	0.62 pp
ja-JP	Ranking	√	2	3	12.96 pp	√	2	3	10.51 pp
	Accuracy		√		0.32		√		0.76
	Fluency		√		0.2		√		0.49
	Edit Distance	√	3	2	8.17 pp	√	3	2	5.79 pp
	Edit Distance (PE)	√		2	21.07 pp				
pt-BR	Ranking	√	3	2	4.59 pp	√	2	3	6.65 pp
	Accuracy		√		0.09		√		0.26
	Fluency		√		0.45		√		0.28
	Edit Distance	2	3	√	1.68 pp	√	3	2	0.28 pp
	Edit Distance (PE)	2		√	3.62 pp				
zh-CN	Ranking	√	2	3	10.57 pp	√	2	3	10.40 pp
	Edit Distance	√	3	2	5.87 pp	√	3	2	3.12 pp
ru-RU	Ranking	√	2	3	5.95 pp				
	Edit Distance	2	3	√	1.58 pp				



welocalizeO
doing things differently

Conclusions



welocalize 
doing things differently

Conclusions

- Generic NMT is a suitable alternative for generic domains across all the language pairs.
- In the technology domain, generic NMT is a suitable alternative for some language pairs, such as Chinese and Japanese, where we see a substantial increase in performance compared to customized SMT.
- Because most of our enterprise-level programs rely on accurate terminology, we recommend waiting for customized NMT for the remaining language pairs.
- Post-edit distance on actual post-edited content proved to be the most reliable metric in our evaluation. Ranking and Adequacy & Fluency scoring from the same resource was not always consistent. Autoscores (HT) did not correlate with human evaluations in several cases.



NEXT STEPS

We are running several follow-up pilots:

- 1) Comparing the performance of customized NMT against customized SMT.
- 2) Comparing Post-edit distance in live production using customized SMT and generic NMT. We would like to see if more extensive production data will confirm our initial findings.



Machine Translation Summit XVI

Journey around Neural Machine Translation quality

Marco Ganci

Sr. Software Engineer, Globalization Solutions

marco.ganci@autodesk.com

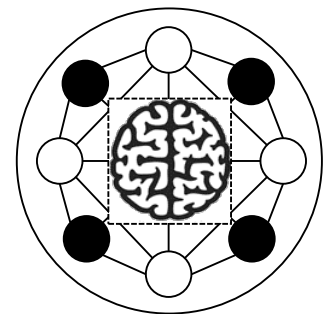
[linkedin.com/in/marcoganci](https://www.linkedin.com/in/marcoganci)

 AUTODESK.

© 2017 Autodesk | Localization Solutions

Neural Machine Translation

- **Neural Machine Translation (NMT)** is an end-to-end learning approach for automated translation, with the potential to overcome many of the weaknesses of conventional phrase-based translation systems
- The strength of NMT lies in its ability to learn directly, in an end-to-end fashion, the mapping from input text to associated output text



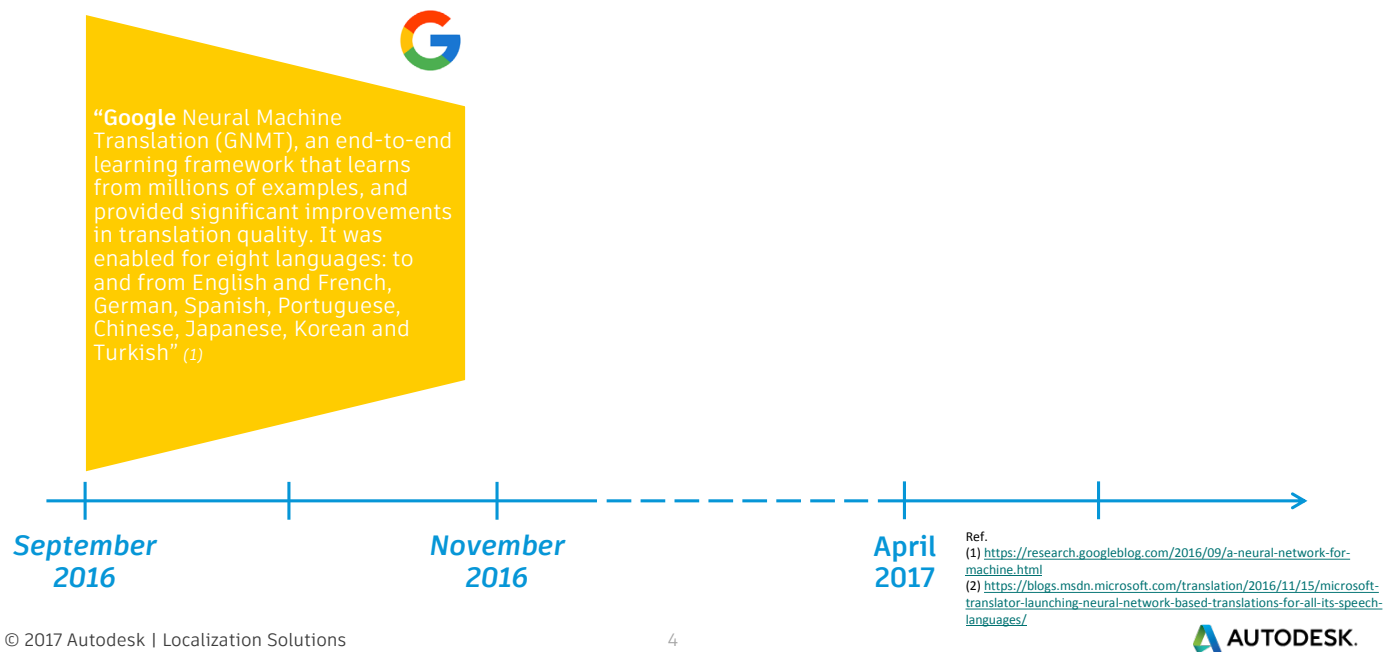
Ref. <https://arxiv.org/abs/1609.08144>

 AUTODESK.

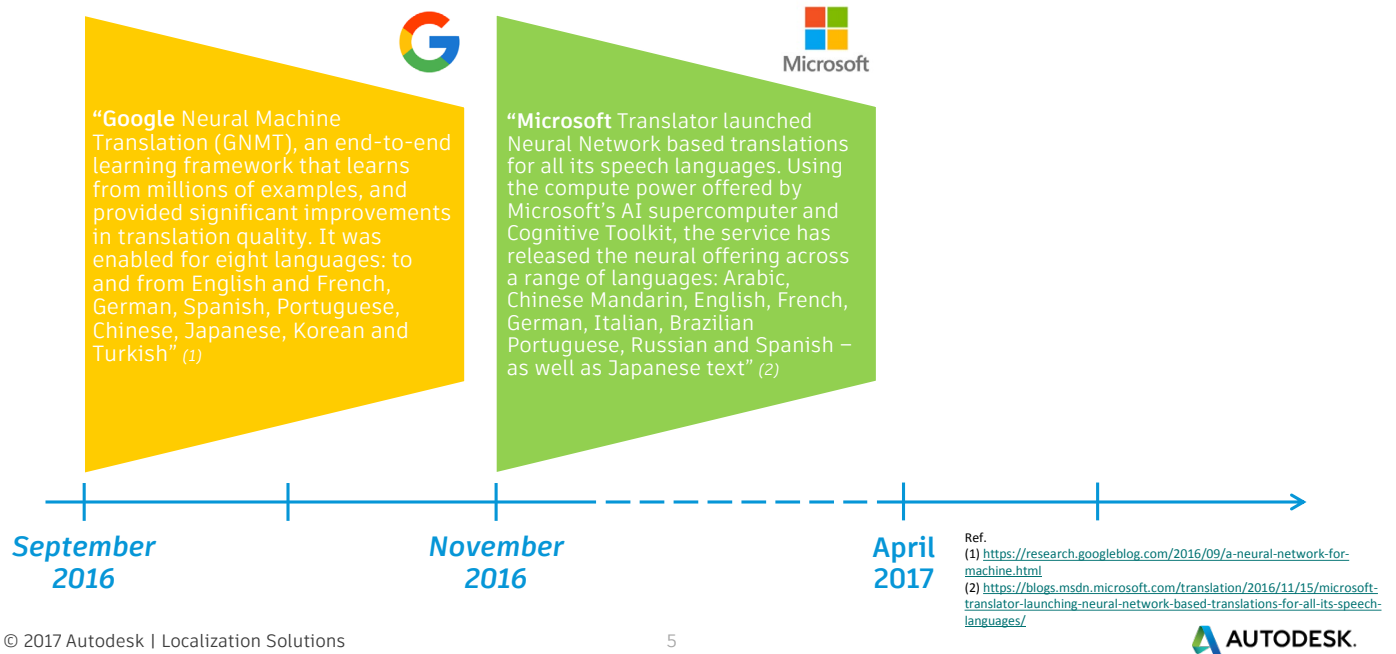
Neural Machine Translation timeline



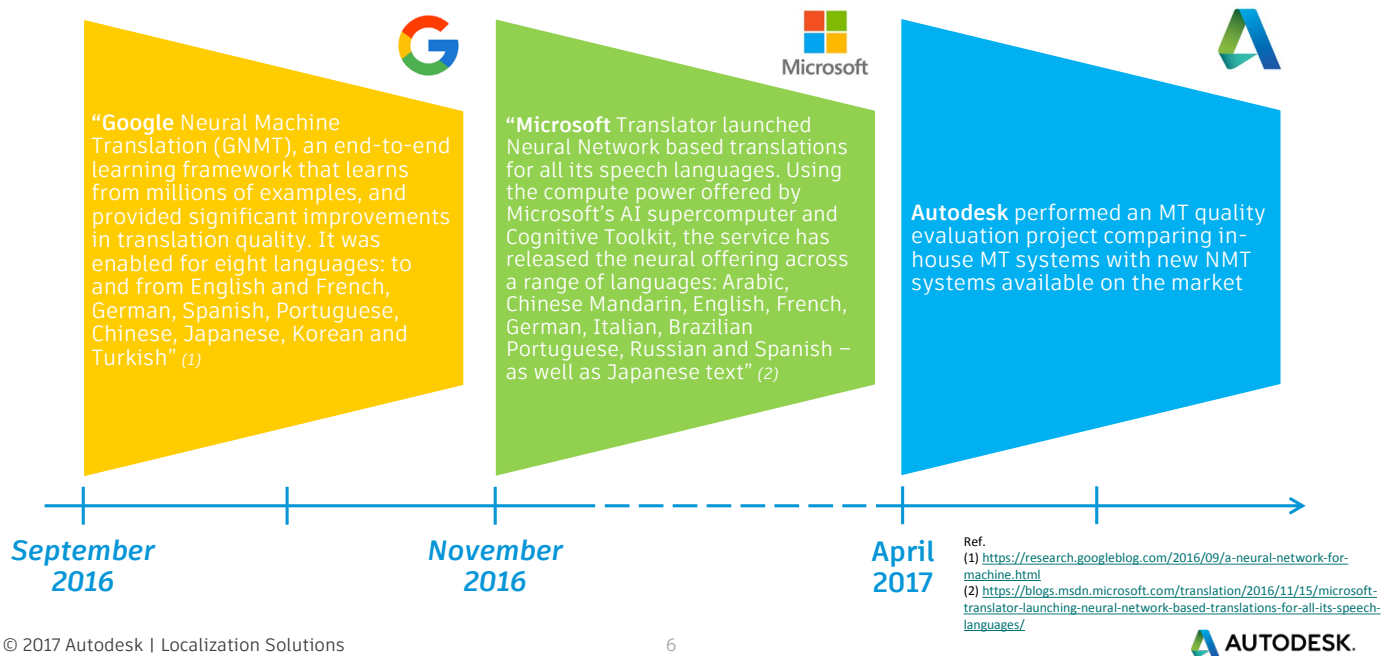
Neural Machine Translation timeline



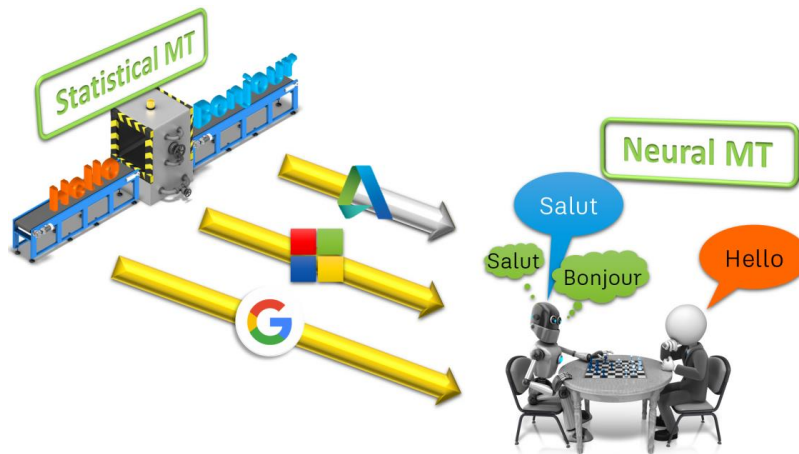
Neural Machine Translation timeline



Neural Machine Translation timeline



Goal



Assess quality of *Neural MT* versus *Autodesk MT*

Assumptions: MT systems

Assumptions: MT systems

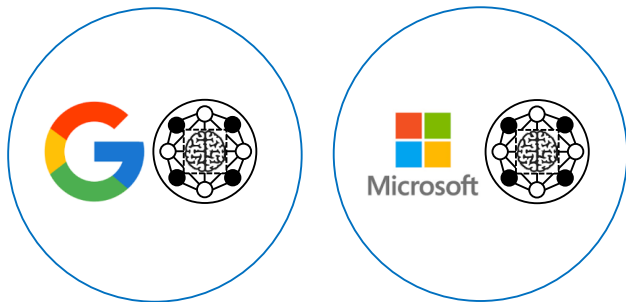


© 2017 Autodesk | Localization Solutions

9



Assumptions: MT systems

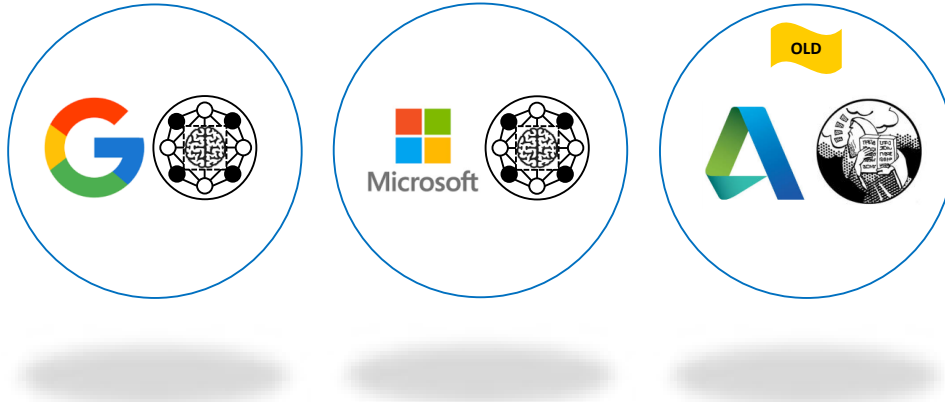


© 2017 Autodesk | Localization Solutions

10



Assumptions: MT systems

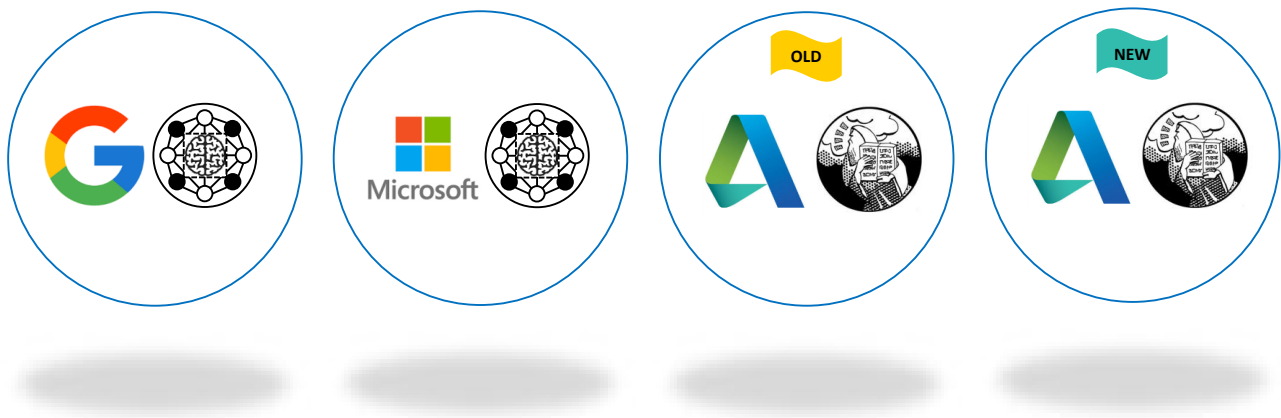


© 2017 Autodesk | Localization Solutions

11



Assumptions: MT systems

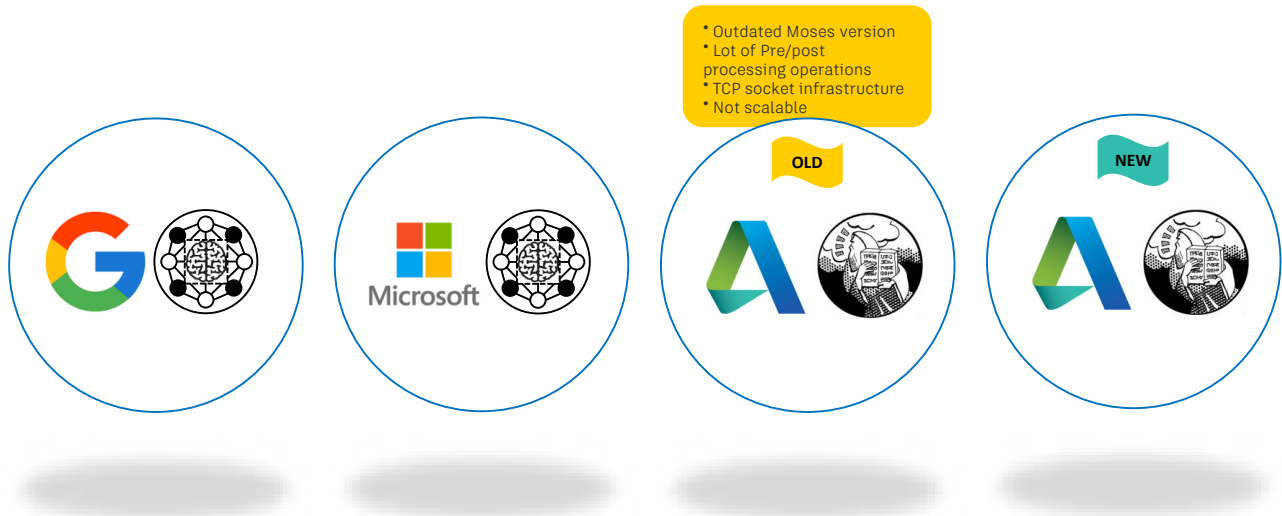


© 2017 Autodesk | Localization Solutions

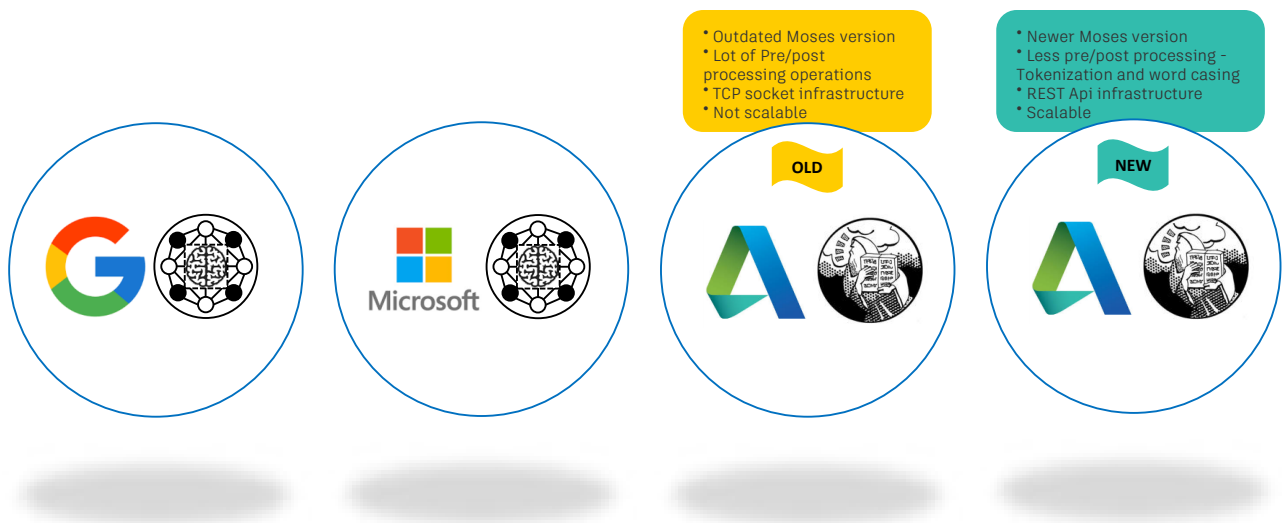
12



Assumptions: MT systems



Assumptions: MT systems



Assumptions: Products



Knowledge Network



Assumptions: Products

ADSK legacy product



Assumptions: Products

ADSK legacy product

AUTODESK MIX
Knowledge Network
Used to train ADSK MT
AUTODESK INFRAWORKS
Dynamo



Assumptions: Products

ADSK legacy product

AUTODESK MIX
Knowledge Network
Used to train ADSK MT
AUTODESK INFRAWORKS
Dynamo

ADSK new product or External product

Delcam
Apache OpenOffice™

Assumptions: ADSK legacy product

ADSK legacy product



- Human Translation for these products started from the **OLD ADSK MT** (translation is now post-editing)
- For some portions of *Infraworks* and *Dynamo* final Human Translation was then used to retrain the engines **ADSK MT, OLD** and **NEW**
- The nature of Autodesk content favors higher matches even on non-trained engines (i.e. Architecture, 3D and so on)
- For these products it looks like there isn't much difference whether an engine was retrained or not, therefore we will not make a distinction in the conclusions

© 2017 Autodesk | Localization Solutions

19

AUTODESK.

Assumptions: Products

- Cases which shouldn't give any advantage to **ADSK MTs**
- It was not easy to find content for which we haven't trained our engines. But looking at the results it is clear that we would benefit from more languages at least for the identified content.

For example we don't have such samples for **German** and **Simplified Chinese**.

ADSK new product or External product



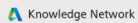





© 2017 Autodesk | Localization Solutions

20

AUTODESK.

Assumptions: Scope

PRODUCT	CATEGORY	Languages					
		German	French	Spanish	Japanese	Simplified Chinese	Portuguese Brazilian
 Dynamo	SW	45k	45k	45k	12k	45k	
	DOC	51k	51k	51k	12k	51k	
 AUTODESK INFRAWORKS	SW	45k	57k	56k	18k	17k	55k
	DOC	374k	437k	286k	89k	119k	427k
 Knowledge Network	DOC	166k	164k	151k	50k	43k	
 AUTODESK MIX	DOC	5k	6k	7k	2k	1.5k	6k
 Delcam	DOC		244k		57k		658k
 Apache OpenOffice	DOC		397k	282k	407k		

ADSK legacy product

ADSK new product or External product

Used to train ADSK MT

Approach

Approach



AUTOMATIC

- Automatic quality evaluation comparing machine's output and human translation

Approach



AUTOMATIC

- Automatic quality evaluation comparing machine's output and human translation

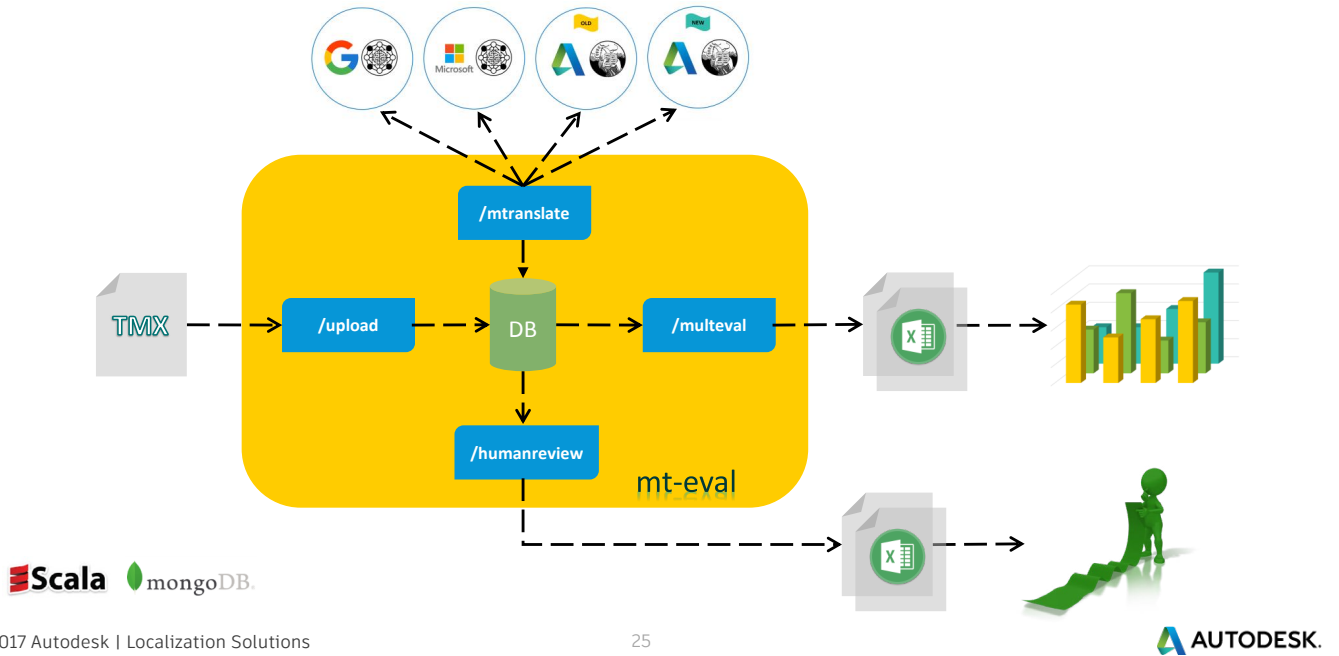


MANUAL

- Human review, involving internal native speakers and external reviewers

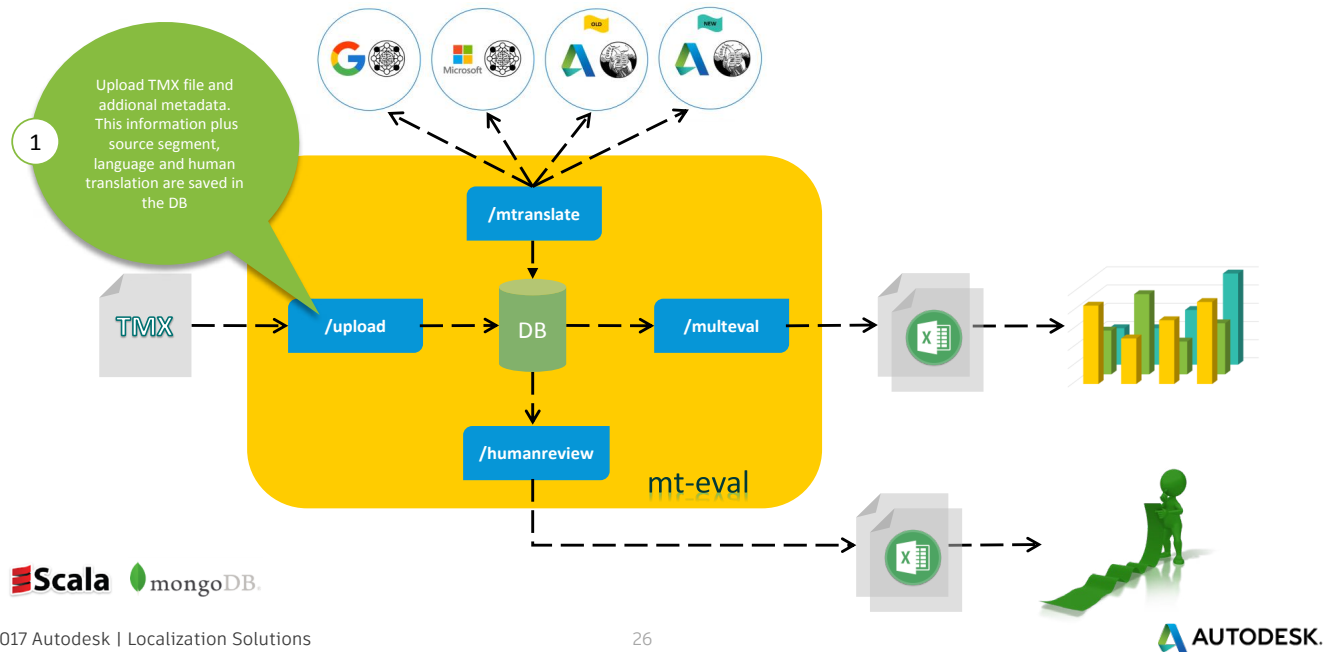
Automatic: mt-eval system

Ref. *
<https://git.autodesk.com/LocalizationServices/multeval>
<https://github.com/jhclark/multeval>

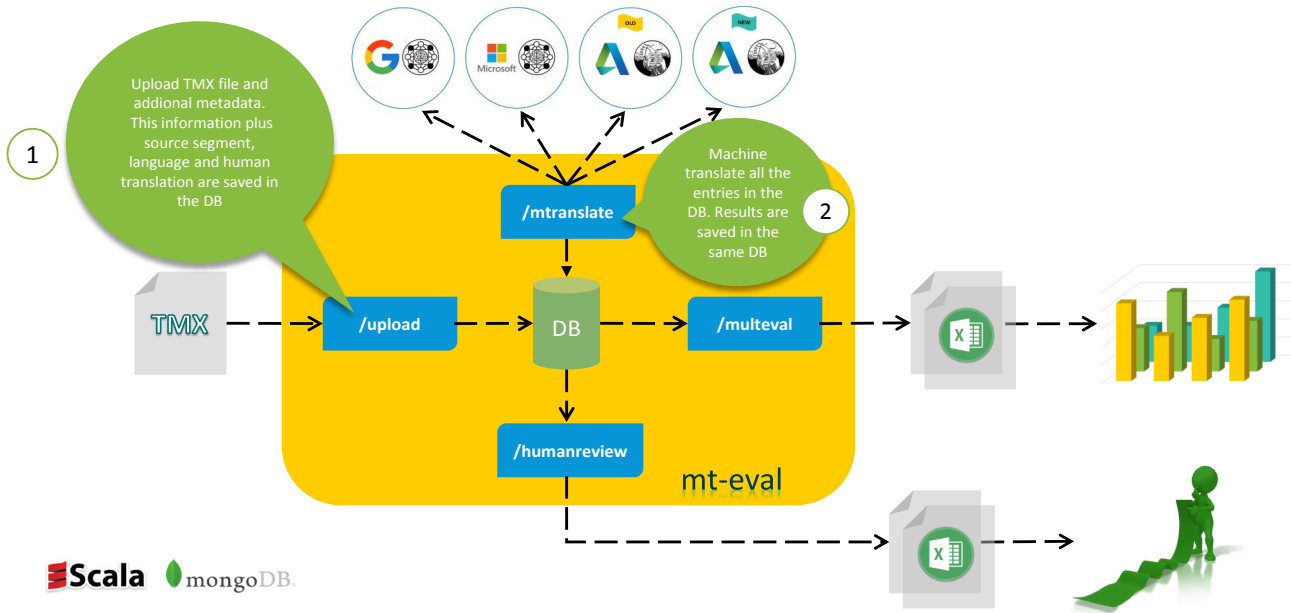


Automatic: mt-eval system

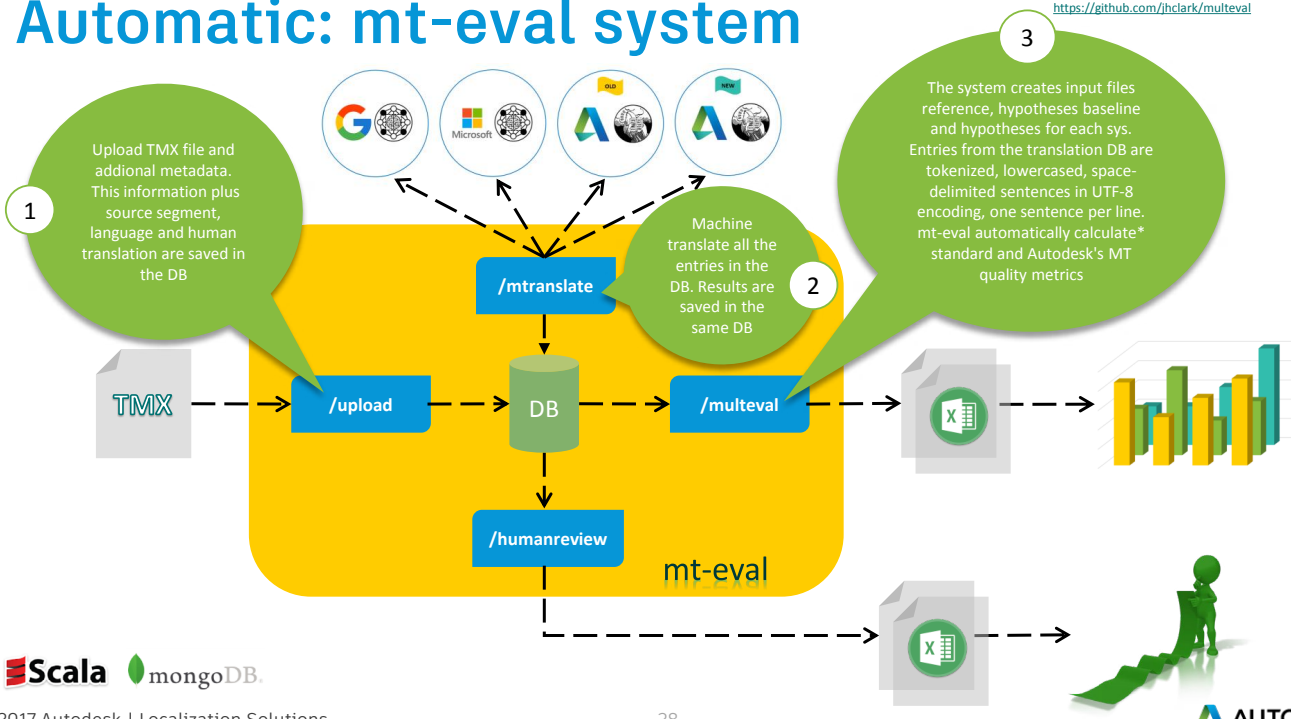
Ref. *
<https://git.autodesk.com/LocalizationServices/multeval>
<https://github.com/jhclark/multeval>



Automatic: mt-eval system

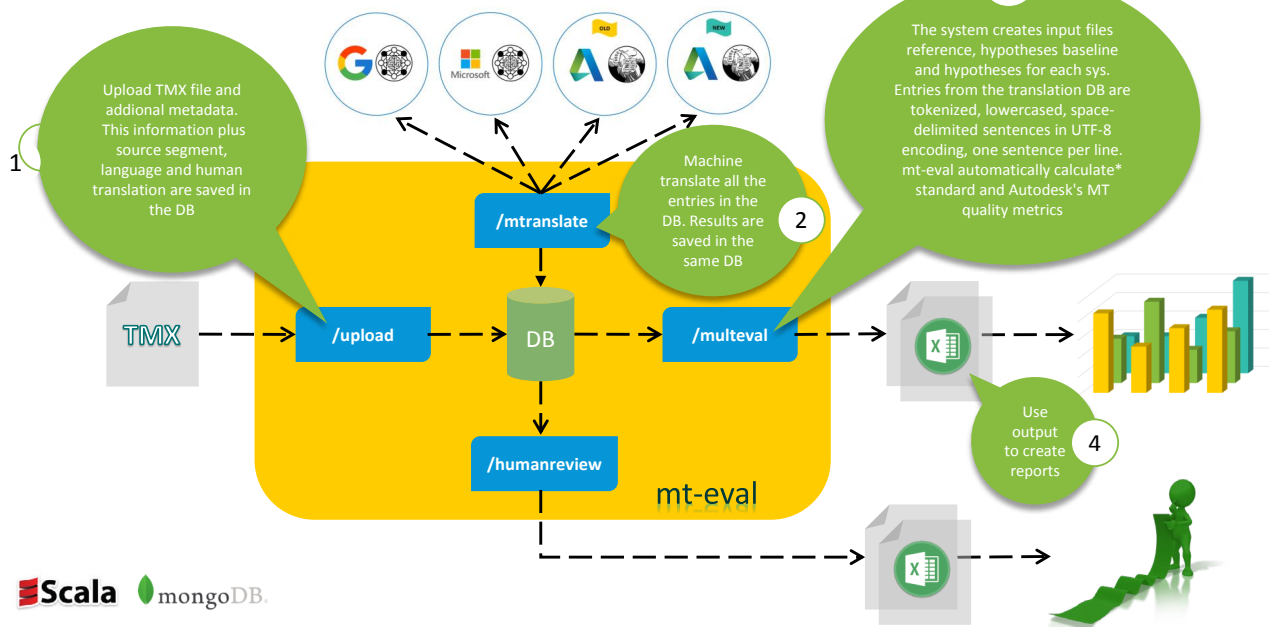


Automatic: mt-eval system



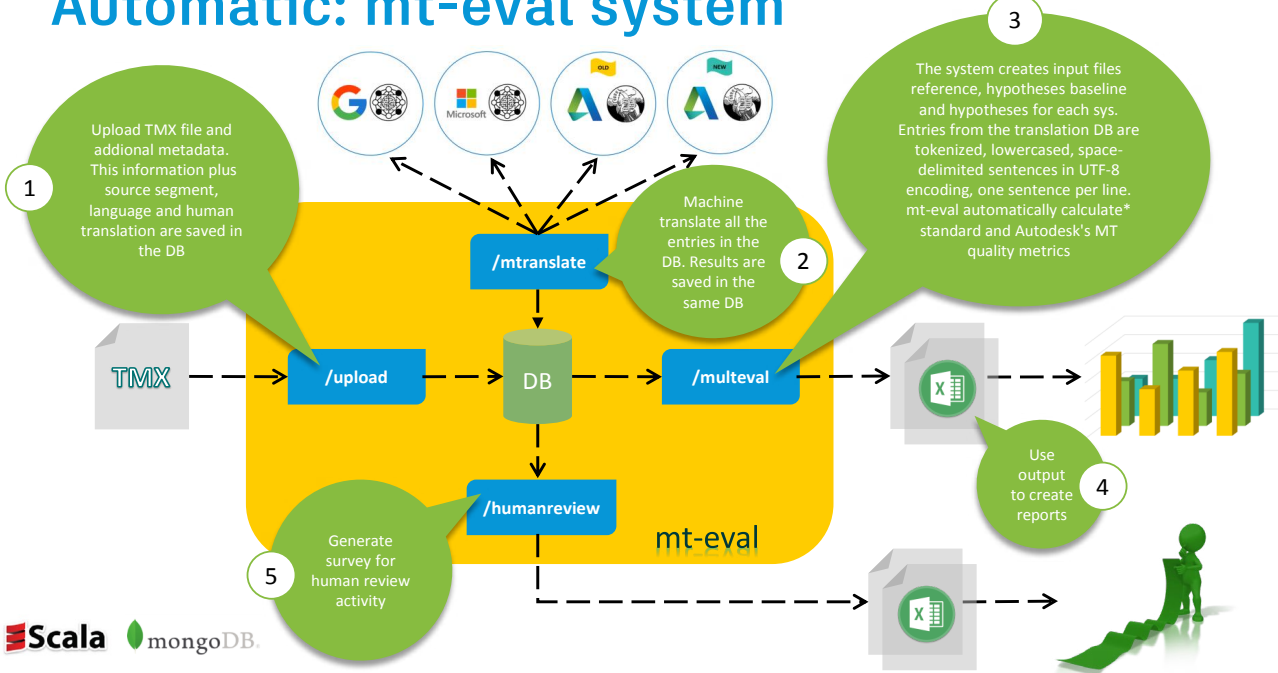
Automatic: mt-eval system

Ref. *
<https://git.autodesk.com/LocalizationServices/multeval>
<https://github.com/hclark/multeval>



Automatic: mt-eval system

Ref. *
<https://git.autodesk.com/LocalizationServices/multeval>
<https://github.com/hclark/multeval>



Automatic: MT quality metrics

COMMON

BLEU - Bilingual Evaluation Understudy

- Quality is considered to be the correspondence between a machine's output and that of a human. The closer a machine translation is to a professional human translation, the better it is (1)

METEOR - Metric for Evaluation of Translation with Explicit Ordering

- The metric evaluates translation hypotheses by aligning them to reference translations and calculating sentence-level similarity scores. It uses stemming and synonymy matching, along with the standard exact word matching. The metric was designed to fix some of the problems found in BLEU (2)

TER - Translation Error Rate

- A method to determine the amount of Post-Editing required for machine translation jobs. The automatic metric measures the number of actions required to edit a translated segment inline with one of the reference translations (3)

Length

- Machine's output length over professional human translation length as a percent. If it is 100%, machine and human translation output have the same length (4)



CFS - Character-based Levenshtein distance

- Levenshtein distance on character level

WFS - Word-based Fuzzy Score

- Levenshtein distance on word level

JFS - Joint Fuzzy Score

- It is a combination of the two above, taking the worse of the two scores for each segment and computing a joined score like this for the whole test set

All three below are based on the **Levenshtein** distance between the output and the reference translation, the higher the score the better.

Levenshtein distance between two words is the minimum number of single-character edits (i.e. insertions, deletions or substitutions) required to change one word into the other.

Ref.

(1) <https://en.wikipedia.org/wiki/BLEU>

(2) <http://www.cs.cmu.edu/~alavie/METEOR/>

(3) <https://kantanmtblog.com/2015/07/28/what-is-translation-error-rate-ter/>

(4) <https://git.autodesk.com/LocalizationServices/multeval>



Manual: Human review rating



Adequacy

How much of the meaning expressed in the source is also expressed in the target translation

- None:** Completely nonsense translation
- Little:** Sentence preserves some of the meaning of the source sentence but misses significant parts
- Most:** Sentence retains most of the meaning of the source sentence, but may have some grammar mistakes
- Everything:** Perfect translation: the meaning of the translation is completely consistent with the source, and the grammar is correct

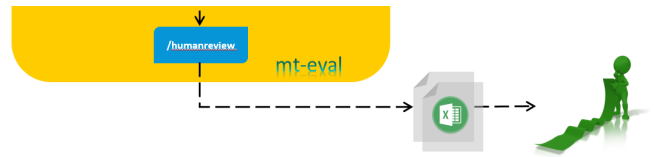
Fluency

Readability and naturalness of the translated text

- Incomprehensible:** The content is not fluent nor natural in the target language. The translated text is a word by word translation, therefore it is hard to read and understand.
- Disfluent:** The content reads like it was translated. Some sentence structures don't seem to be natural in the target language or are not idiomatic. It contains some literal translations.
- Good:** The content reads like it was originally written in the target language. It uses proper sentence structure and idiomatic expressions. But a few minor improvements might be necessary.
- Flawless:** The content reads like it was originally written in the target language. It uses proper sentence structure and idiomatic expressions.



Manual: Survey

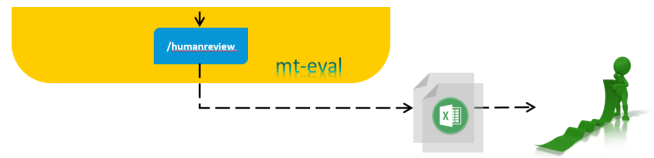


Accuracy Score	Fluency Score	Translation	Source
2	3	グループになるまで[Shift]キーを押してオブジェクトの選択が完了したらマウスクリックを使用します。	Group by pressing Shift until you have finished selecting the objects with mouse clicks.
4	4	条件より大きな値を表示します。	Shows values greater than the condition.
4	4	実習プログラムの一環としてオートデスクの教育機関限定ライセンスの対象ソフトウェアの使用が商用プロジェクトになる場合: オートデスクの教育機関限定ライセンスの使用条件により、その使用は教育と実習関連活動に限定されます。	If your use of Autodesk software subject to an Educational license as part of the apprentice program will be part of commercial projects: The terms and conditions for Autodesk Educational licenses restrict the use exclusively to teaching and exercising activities.
4	4	P&IDとの互換性	P&ID compatibility
1	1	長細いフィーチャが最適でない可能性が高いというこの方法でフライス加工します。	Long thin features probably are not best milled in this way.
4	4	{1}ベクトルの詳細{2}	{1}Vector Details{2}
4	4	{1}境界カーブ {2}の下で {3}境界 {4}をクリックします。	Under {1}Boundary curves{2}, click {3}Boundaries{4}.
4	4	Autodesk® Inventor Engineer-to-Order Server 2015, Autodesk® Inventor Engineer-to-Order Server 2014, Autodesk® Inventor Engineer-to-Order Server 2013	Autodesk® Inventor Engineer-to-Order Server 2015, Autodesk® Inventor Engineer-to-Order Server 2014, Autodesk® Inventor Engineer-to-Order Server 2013
4	4	改訂日	Revision Date
4	4	579H1	579H1
4	4	並び替え:	Sort by:
1	3	ジオメトリ, 平面	Geometry, Plane
2	3	Shiftキーを押しながらマウスをクリックしてオブジェクトの選択を完了します。	Group by pressing Shift until you have finished selecting the objects with mouse clicks.
3	4	{1}は、Revit ElementからElementCurveReferenceを抽出する必要があります。	{1} requires a ElementCurveReference extracted from a Revit Element!
4	4	オブジェクトを再度選択します。	Select the object again.
4	4	体験版ライセンスから有償ライセンスへの変換	Convert a Trial to a Paid License
2	3	電子メールが送信されましたが、提供されていません。	The email was sent but not delivered.

Internal ~250 segments | External ~ 2500 segments

*OLD ADSK not rated

Manual: Survey

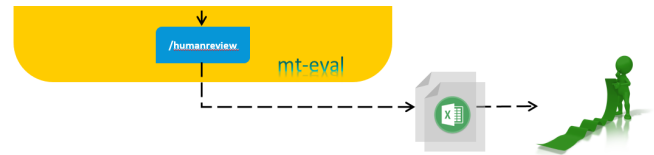


Accuracy Score	Fluency Score	Translation	Source	Product	Category	Type
2	3	グループになるまで[Shift]キーを押してオブジェクトの選択が完了したらマウスクリックを使用します。	Group by pressing Shift until you have finished selecting the objects with mouse clicks.	OPENOFFICE	DOC	ADSK
4	4	条件より大きな値を表示します。	Shows values greater than the condition.	OPENOFFICE	DOC	GOOGLE
4	4	実習プログラムの一環としてオートデスクの教育機関限定ライセンスの対象ソフトウェアの使用が商用プロジェクトになる場合: オートデスクの教育機関限定ライセンスの使用条件により、その使用は教育と実習関連活動に限定されます。	If your use of Autodesk software subject to an Educational license as part of the apprentice program will be part of commercial projects: The terms and conditions for Autodesk Educational licenses restrict the use exclusively to teaching and exercising activities.	AKN	DOC	HT
4	4	P&IDとの互換性	P&ID compatibility	ADSKNT	DOC	GOOGLE
1	1	長細いフィーチャが最適でない可能性が高いというこの方法でフライス加工します。	Long thin features probably are not best milled in this way.	DELCAM	DOC	ADSK
4	4	{1}ベクトルの詳細{2}	{1}Vector Details{2}	DYNAMO	DOC	ADSK
4	4	{1}境界カーブ {2}の下で {3}境界 {4}をクリックします。	Under {1}Boundary curves{2}, click {3}Boundaries{4}.	DELCAM	DOC	MICROSOFT
4	4	Autodesk® Inventor Engineer-to-Order Server 2015, Autodesk® Inventor Engineer-to-Order Server 2014, Autodesk® Inventor Engineer-to-Order Server 2013	Autodesk® Inventor Engineer-to-Order Server 2015, Autodesk® Inventor Engineer-to-Order Server 2014, Autodesk® Inventor Engineer-to-Order Server 2013	AKN	DOC	HT
4	4	改訂日	Revision Date	DYNAMO	SW	GOOGLE
4	4	579H1	579H1	AKN	DOC	ADSK
4	4	並び替え:	Sort by:	DELCAM	DOC	GOOGLE
1	3	ジオメトリ, 平面	Geometry, Plane	DYNAMO	DOC	GOOGLE
2	3	Shiftキーを押しながらマウスをクリックしてオブジェクトの選択を完了します。	Group by pressing Shift until you have finished selecting the objects with mouse clicks.	OPENOFFICE	DOC	GOOGLE
3	4	{1}は、Revit ElementからElementCurveReferenceを抽出する必要があります。	{1} requires a ElementCurveReference extracted from a Revit Element!	DYNAMO	SW	GOOGLE
4	4	オブジェクトを再度選択します。	Select the object again.	DELCAM	DOC	GOOGLE
4	4	体験版ライセンスから有償ライセンスへの変換	Convert a Trial to a Paid License	AKN	DOC	HT
2	3	電子メールが送信されましたが、提供されていません。	The email was sent but not delivered.	ADSKNT	DOC	HT

Internal ~250 segments | External ~ 2500 segments

*OLD ADSK not rated

Manual: Survey



Accuracy Score	Fluency Score	Translation	Source	Product	Category	Type
2	3	グループになるまで[Shift]キーを押してオブジェクトの選択が完了したらマウスクリックを使用します。	Group by pressing Shift until you have finished selecting the objects with mouse clicks.	OPENOFFICE	DOC	ADSK
4	4	条件より大きな値を表示します。	Shows values greater than the condition.	OPENOFFICE	DOC	GOOGLE
4	4	実習プログラムの一環としてオートデスクの教育機関限定ライセンスの対象ソフトウェアの使用が商用プロジェクトになる場合: オートデスクの教育機関限定ライセンスの使用条件により、その使用は教育と実習関連活動に限定されます。	If your use of Autodesk software subject to an Educational license as part of the apprentice program will be part of commercial projects: The terms and conditions for Autodesk Educational licenses restrict the use exclusively to teaching and exercising activities.	AKN	DOC	HT
4	4	P&IDとの互換性	P&ID compatibility	ADSKNT	DOC	GOOGLE
1	1	長いフィーチャが最適でない可能性が高いというこの方法でフェイス加工します。	Long thin features probably are not best milled in this way.	DELCAM	DOC	ADSK
4	4	{1}ベクトルの詳細{2}	{1}Vector Details{2}	DYNAMO	DOC	ADSK
4	4	{1}境界カーブ {2}の下で {3}境界 {4}をクリックします。	Under {1}Boundary curves{2}, click {3}Boundaries{4}.	DELCAM	DOC	MICROSOFT
4	4	Autodesk® Inventor Engineer-to-Order Server 2015, Autodesk® Inventor Engineer-to-Order Server 2014, Autodesk® Inventor Engineer-to-Order Server 2013	Autodesk® Inventor Engineer-to-Order Server 2015, Autodesk® Inventor Engineer-to-Order Server 2014, Autodesk® Inventor Engineer-to-Order Server 2013	AKN	DOC	HT
4	4	改訂日	Revision Date	DYNAMO	SW	GOOGLE
4	4	579H1	579H1	AKN	DOC	ADSK
4	4	並び替え:	Sort by:	DELCAM	DOC	GOOGLE
1	3	ジオメトリ。平面	Geometry.Plane	DYNAMO	DOC	GOOGLE
2	3	Shiftキーを押しながらマウスをクリックしてオブジェクトの選択を完了します。	Group by pressing Shift until you have finished selecting the objects with mouse clicks.	OPENOFFICE	DOC	GOOGLE
3	4	{1}は、Revit ElementからElementCurveReferenceを抽出する必要があります。	{1} requires a ElementCurveReference extracted from a Revit Element!	DYNAMO	SW	GOOGLE
4	4	オブジェクトを再度選択します。	Select the object again.	DELCAM	DOC	GOOGLE
4	4	体験版ライセンスから有償ライセンスへの変換	Convert a Trial to a Paid License	AKN	DOC	HT
2	3	電子メールが送信されましたが、提供されていません。	The email was sent but not delivered.	ADSKNT	DOC	HT

Internal ~250 segments | External ~ 2500 segments

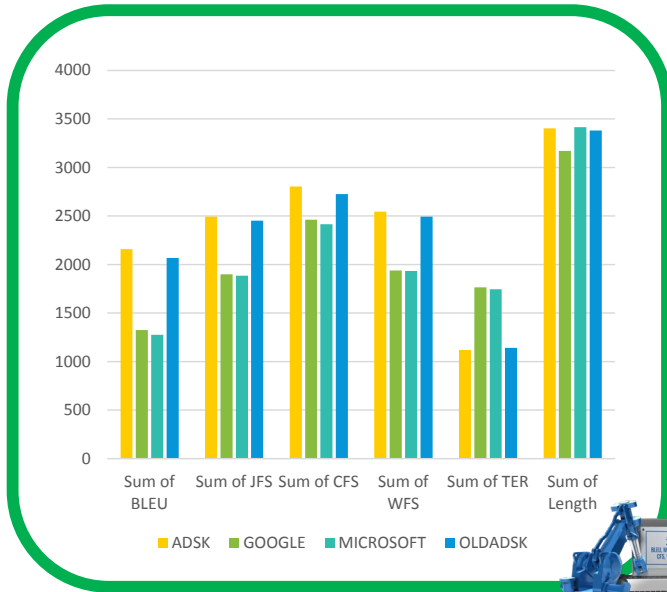
*OLD ADSK not rated

Results: Automatic



Results: Automatic

ADSK legacy product



© 2017 Autodesk | Localization Solutions

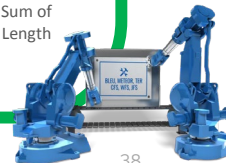
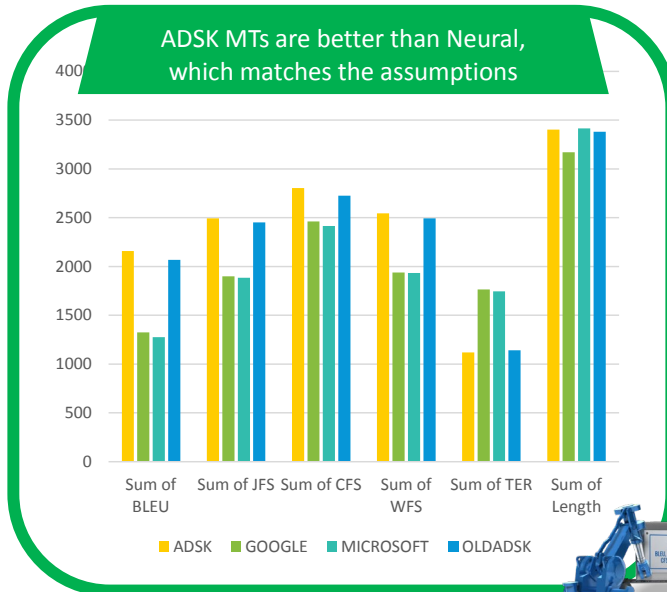
37

* METEOR only for FR and DE – not in the graph



Results: Automatic

ADSK legacy product



© 2017 Autodesk | Localization Solutions

38

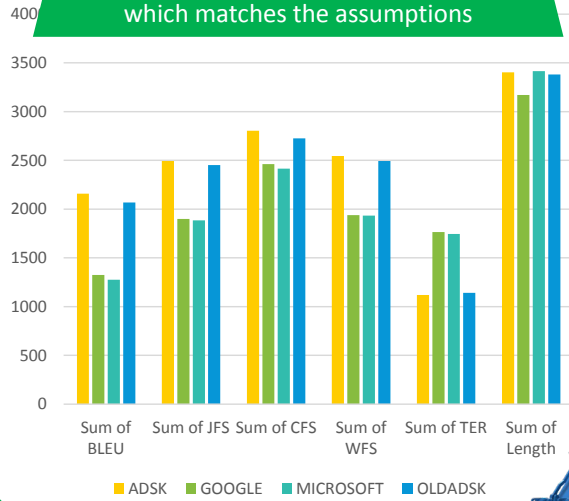
* METEOR only for FR and DE – not in the graph



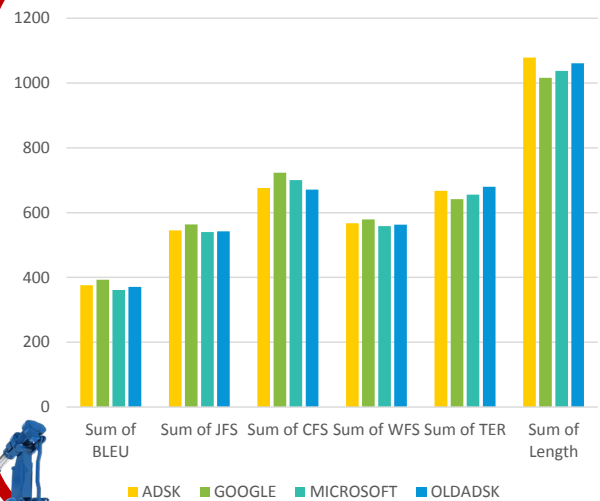
Results: Automatic

ADSK legacy product

ADSK MTs are better than Neural, which matches the assumptions



ADSK new product or External product



© 2017 Autodesk | Localization Solutions

39

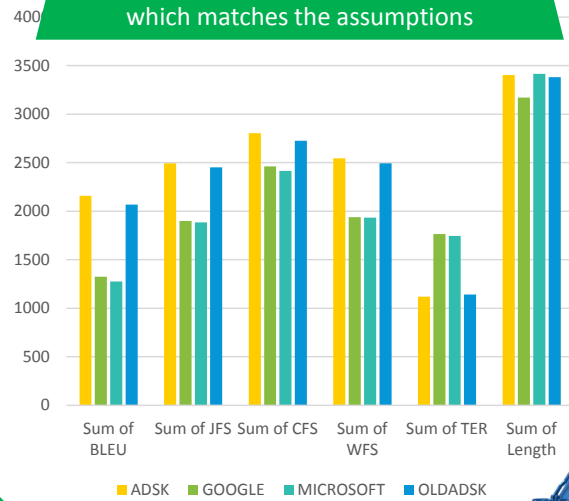
* METEOR only for FR and DE – not in the graph



Results: Automatic

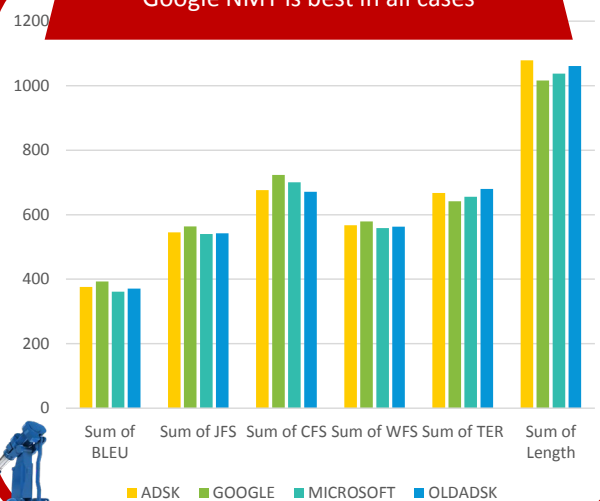
ADSK legacy product

ADSK MTs are better than Neural, which matches the assumptions



ADSK new product or External product

Google NMT is best in all cases



© 2017 Autodesk | Localization Solutions

40

* METEOR only for FR and DE – not in the graph

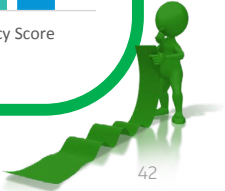
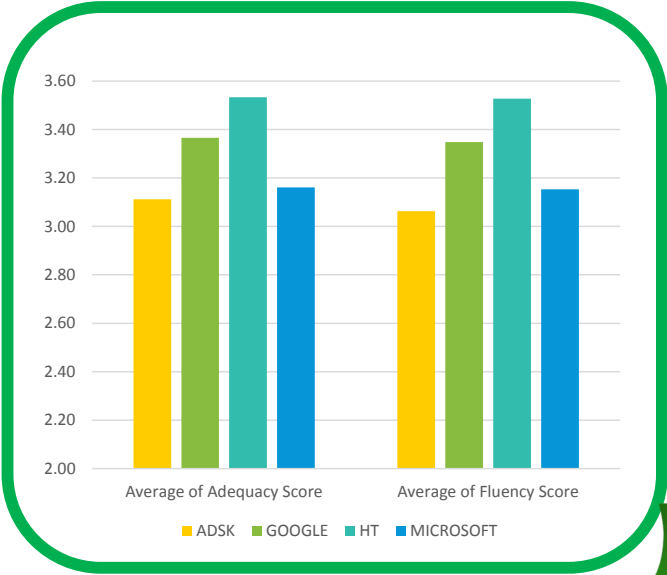


Results: Manual



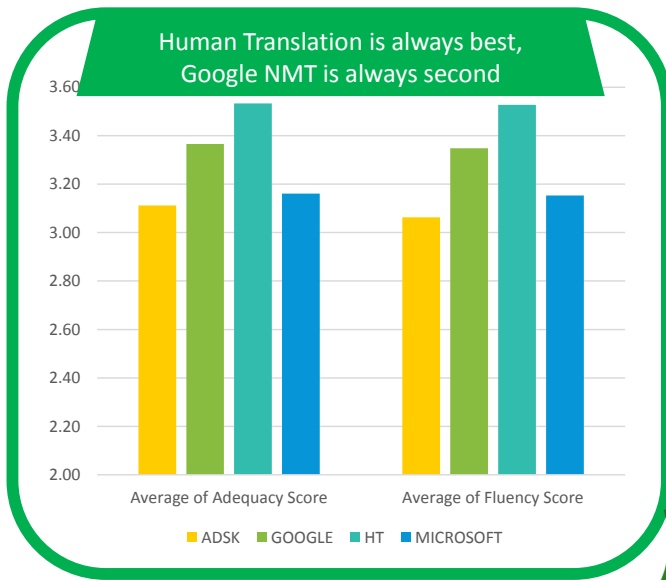
Results: Manual

ADSK legacy product



Results: Manual

ADSK legacy product



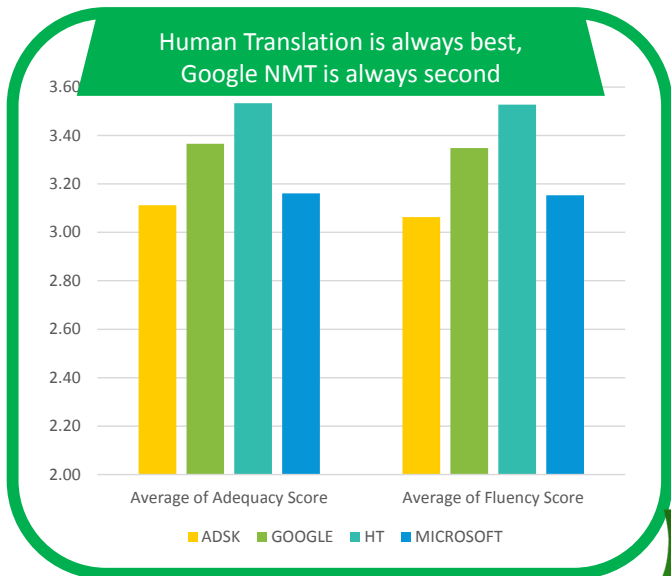
© 2017 Autodesk | Localization Solutions

43



Results: Manual

ADSK legacy product

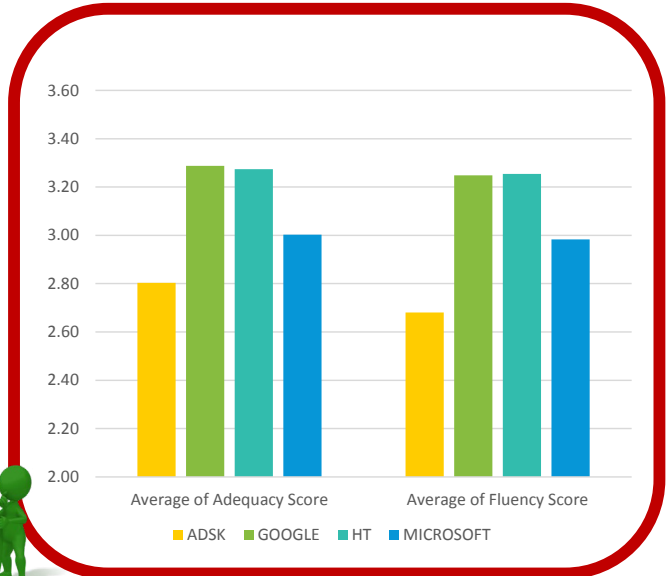


© 2017 Autodesk | Localization Solutions

44

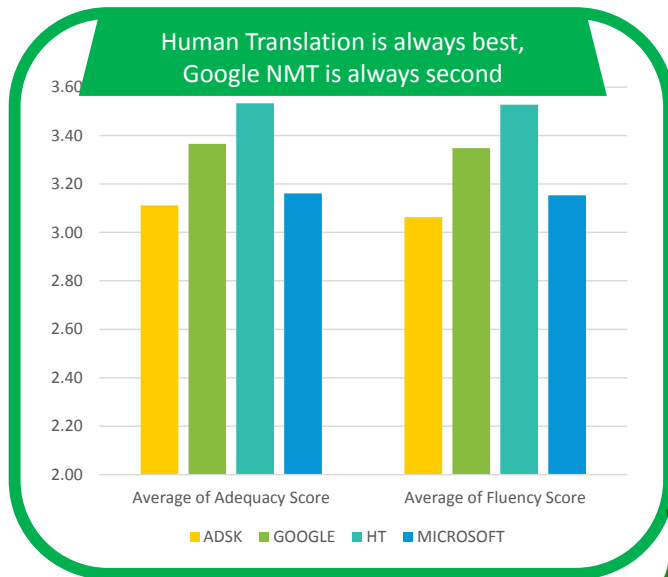


ADSK new product or External product

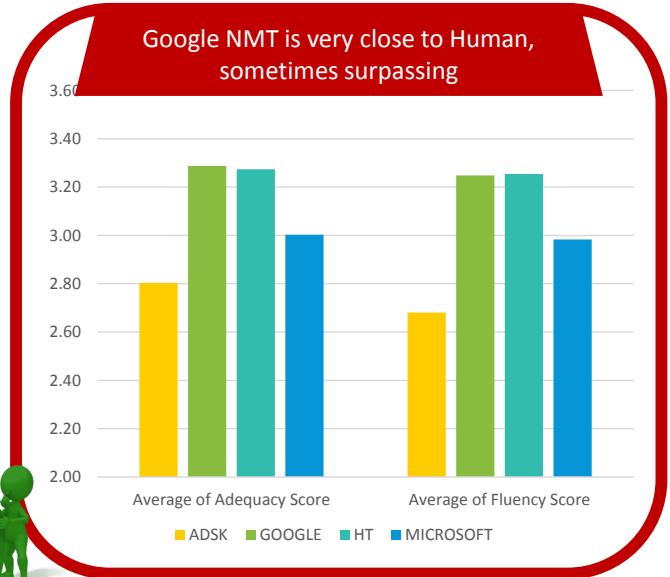


Results: Manual

ADSK legacy product



ADSK new product or External product



© 2017 Autodesk | Localization Solutions

45

AUTODESK.

Conclusions

- Commercial Neural MT are **viable**
- Moses Engines** are still useful on legacy products
- Next Steps:
 - Explore **Open source solutions** (i.e. OpenNMT)
 - Use the best **MT system** that matches current context (i.e. product, language, content type, etc.)



© 2017 Autodesk | Localization Solutions

46

AUTODESK.



Autodesk and the Autodesk logo are registered trademarks or trademarks of Autodesk, Inc., and/or its subsidiaries and/or affiliates in the USA and/or other countries. All other brand names, product names, or trademarks belong to their respective holders. Autodesk reserves the right to alter product offerings, and specifications and pricing at any time without notice, and is not responsible for typographical or graphical errors that may appear in this document. © 2017 Autodesk. All rights reserved.

Result: Breakdown

Approach	Results
ADSK legacy product	AUTOMATIC <ul style="list-style-type: none"> *NEW and OLD ADSK MTs are clearly better than Neural - which matches the assumptions *NEW and OLD ADSK MTs tend to have very similar results, except for <i>German</i> *Between Neural MTs, only <i>Japanese</i> shows better results with Microsoft than Google
	MANUAL <ul style="list-style-type: none"> *Human Translation is always best except one case only for fluency for <i>Portuguese</i> where Google Neural is a little bit better *Google Neural is always second *Hard to say whether ADSK or Microsoft are best, it varies between languages but globally they are quite a bit lower than the others and close together
ADSK new product or External product [Breakdown]	AUTOMATIC <ul style="list-style-type: none"> *Google Neural tends to be best in all cases except <i>Japanese</i> *For <i>Japanese</i> Microsoft Neural is the best *Neural is better than ADSK MT, NEW and OLD
	MANUAL <ul style="list-style-type: none"> *Google Neural is very close to Human, sometimes surpassing *Microsoft and ADSK are often close alternating third position <p><i>For OPENOFFICE we had to ignore Human Translation scores</i></p>

Breakdown: ADSK legacy product (1/2)

Language	Approach	Ranking	Notes
German	AUTOMATIC	1.NEW ADSK 2.OLD ADSK 3.Google Neural / Microsoft Neural	*NEW ADSK is the best and quite a bit better than the OLD ADSK *Google Neural and Microsoft Neural have very similar results, which are quite a bit lower than ADSK
	MANUAL	1.Human Translation 2.Google Neural 3.Microsoft Neural 4.NEW ADSK	*Human is best *Second Google Neural, not too much lower *Third is Microsoft Neural *Worst is NEW ADSK * <i>Adequacy</i> and <i>Fluency</i> same pattern for all
Spanish	AUTOMATIC	1.NEW ADSK / OLD ADSK 2.Google Neural 3.Microsoft Neural	*NEW and OLD ADSK are the best and very close *Google Neural is better than Microsoft Neural, but quite a bit lower than ADSK
	MANUAL	1.Human Translation 2.Google Neural 3.NEW ADSK / Microsoft Neural	*Human is best *Second Google Neural, then NEW ADSK then Microsoft Neural > these three are very close * <i>Adequacy</i> and <i>Fluency</i> same pattern for all
French	AUTOMATIC	1.NEW ADSK / OLD ADSK 2.Google Neural 3.Microsoft Neural	*NEW and OLD ADSK are the best and very close *Google Neural is better than Microsoft Neural, but quite a bit lower than ADSK
	MANUAL	1.Human Translation 2.Google Neural 3.NEW ADSK / Microsoft Neural	*Human is best *Second Google Neural *Then ADSK and then MS > these two are very close * <i>Adequacy</i> and <i>Fluency</i> same pattern for all

Breakdown: ADSK legacy product (2/2)

Language	Approach	Ranking	Notes
Portuguese	AUTOMATIC	1.NEW ADSK / OLD ADSK 2.Google Neural 3.Microsoft Neural	*NEW and OLD ADSK are the best and very close *Google Neural is better than Microsoft Neural, but quite a bit lower than ADSK
	MANUAL	<i>Adequacy</i> 1.Human Translation 2.Google Neural 3.NEW ADSK / Microsoft Neural <i>Fluency</i> 1.Google Neural 2.Human Translation 3.NEW ADSK / Microsoft Neural	* <i>Adequacy</i> • Human is best, Goggle Neural second quite a bit lower * <i>Fluency</i> • Google Neural is best, Human is close *NEW ADSK and Microsoft Neural are quite a bit lower and close for both Adequacy and FL
Japanese	AUTOMATIC	1.NEW ADSK / OLD ADSK 2.Google Neural 3.Microsoft Neural	*NEW and OLD ADSK MT are the best and very close *Microsoft Neural is better than Google Neural, but lower than ADSK *One score, CFS > all results are incredibly close
	MANUAL	1.Human Translation 2.Google Neural 3.Microsoft Neural 4.NEW ADSK	*Human is best *Second Google Neural, not too much lower *Third is Microsoft Neural *Worst is NEW ADSK * <i>Adequacy</i> and <i>Fluency</i> same pattern for all
Simplified Chinese	AUTOMATIC	1.NEW ADSK / OLD ADSK 2.Google Neural 3.Microsoft Neural	*NEW and OLD ADSK MT are the best and very close *Google Neural is quite a bit better than Microsoft Neural, but quite a bit lower than NEW ADSK
	MANUAL	<i>Adequacy</i> 1.Human Translation 2.Google Neural 3.NEW ADSK 4.Microsoft Neural <i>Fluency</i> 1.Human Translation 2.Google Neural 3.Microsoft Neural 4.NEW ADSK	*Human is best - both Adequacy and FI *Google Neural is second best - both Adequacy and FI * <i>Adequacy</i> • NEW ADSK is better than Microsoft Neural * <i>Fluency</i> • Microsoft Neural is slightly better than ADSK MT

ADSK legacy product: Trained VS Not-Trained

Approach	TRAINED: DYNAMO (SW), INFRAWORKS (SW/DOC) [Breakdown]	NOT-TRAINED: DYNAMO (DOC), AKN (DOC), ADSK MIX (DOC) [Breakdown]
MANUAL	<ul style="list-style-type: none"> • Human Translation is always best • Google Neural is second in most of the languages • NEW ADSK is close to or a little bit better than Google Neural in French, Spanish and Portuguese • Microsoft Neural is worst in most of the languages except Japanese and German Fluency 	<ul style="list-style-type: none"> • Human Translation is always best except Portuguese where Google Neural is best • Google Neural is second and close to Human Translation in most of the languages • Microsoft Neural is third in most of the languages except Spanish • NEW ADSK is worst not far away from Microsoft Neural
AUTOMATIC	<ul style="list-style-type: none"> • NEW ADSK is always best • OLD ADSK is always second except Japanese • Google Neural and Microsoft Neural are close in most of the languages except <ul style="list-style-type: none"> • Simplified Chinese where Google is clearly better than Microsoft Neural • Japanese where Microsoft Neural is clearly better than Google Neural 	<ul style="list-style-type: none"> • OLD ADSK is always best • NEW ADSK is always second except CFS in Japanese and Simplified Chinese • Google Neural is third • Microsoft Neural is fourth, very close to Google Neural in most of the languages

Breakdown: ADSK new product or External product (1/2)

Product	Language	Approach	Ranking	Notes
DELCAM	French	AUTOMATIC	1. Google Neural / Microsoft Neural 2. NEW ADSK 3. OLD ADSK	*Google Neural and Microsoft Neural are the best and very close *NEW ADSK is a bit lower than Neural, and quite a bit better than OLD ADSK
		MANUAL	1. Human Translation 2. Google Neural 3. Microsoft Neural 4. NEW ADSK	*Human is best *Google is second not too far from Human *Microsoft Neural is third quite a bit lower *NEW ADSK last not too far from Microsoft Neural
	Japanese	AUTOMATIC	1. Microsoft Neural 2. Google Neural 3. NEW ADSK / OLD ADSK	*Microsoft Neural is the best and quite a bit better than Google Neural *NEW and OLD ADSK are lower and very close
		MANUAL	1. Google Neural 2. Human Translation / Microsoft Neural 3. NEW ADSK	*Google Neural is best *Followed by Human and Microsoft Neural being very close together *NEW ADSK last a bit lower
	Portuguese	AUTOMATIC	1. Google Neural 2. Microsoft Neural 3. NEW ADSK / OLD ADSK	*Google Neural is the best *Google and MS Neural are the best and close *NEW and OLD ADSK are lower and very close
		MANUAL	<u>Adequacy</u> 1. Human Translation 2. Google Neural 3. Microsoft Neural 4. NEW ADSK <u>Fluency</u> 1. Google Neural 2. Human Translation 3. Microsoft Neural 4. NEW ADSK	* <u>Adequacy</u> • Human is best, Google Neural second but very close * <u>Fluency</u> • Opposite, Google Neural best with Human very close *Third is Microsoft Neural followed closely by NEW ADSK

Breakdown: ADSK new product or External product (2/2)

Product	Language	Approach	Ranking	Notes
OPENOFFICE	French	AUTOMATIC	1.Google Neural 2.Microsoft Neural 3.NEW ADSK 4.OLD ADSK	*Google Neural is the best *Google Neural and Microsoft Neural are the best and close *NEW and OLD ADSK are lower and close
		MANUAL	1.Google Neural 2.Microsoft Neural 3.NEW ADSK	*Google Neural is best *Microsoft Neural is second *NEW ADSK last not too far
	Japanese	AUTOMATIC	1.Microsoft Neural (except BLEU) 2.OLD ADSK 3.Google Neural 4.NEW ADSK	*Microsoft Neural is the best except for BLEU where OLD ADSK wins *OLD ADSK is generally higher than Google Neural
		MANUAL	1.Google Neural / Microsoft Neural 2.NEW ADSK	*Google Neural and Microsoft Neural are best very close * <u>Adequacy</u> <ul style="list-style-type: none"> Microsoft Neural a little better, opposite for <u>Fluency</u> *NEW ADSK is quite a bit lower
	Spanish	AUTOMATIC	1.Google Neural 2.OLD ADSK / NEW ADSK / Microsoft Neural	*Google Neural is the best *The rest is lower and quite similar results
		MANUAL	1.Google Neural 2.NEW ADSK 3.Microsoft Neural	*Google Neural is best *NEW ADSK is second *Microsoft Neural is last *All very close

A Reception Study of Machine Translated Subtitles for MOOCs

Ke Hu, Sharon O'Brien, Dorothy Kenny
ADAPT Centre, SALIS, Dublin City University



Overview

www.adaptcentre.ie

- § Context
 - § Why MOOCs?
 - § Why MT?
 - § Why reception?
- § Research question and hypothesis
 - § Methodologies
 - § Reception model
 - § Sub-hypotheses
- § Pilot study
- § What's next?

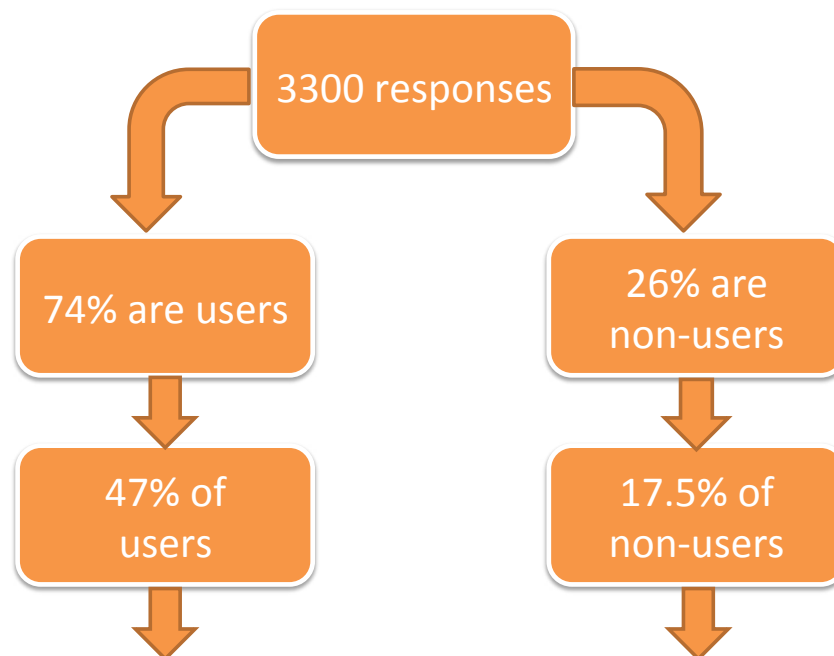
2008	“MOOC” coined by Dave Cormier (2008, online) Massive Open Online Courses E.g.: Coursera, Udacity, edX...
2013	Coursera has over 30 university partners, 2.8 million registered students, 1.4 million course enrolments every month (Cusumano, 2013)
In China	
Early 2013	Ø Chinese universities started to join MOOCs Ø 4 universities joined edX, 6 universities joined Coursera
2014	Ø 2 universities joined FutureLearn Ø Over 50 MOOCs offered by Chinese universities on international platforms (Yuan & Liu, 2014)
Now	Around 20 Chinese MOOC platforms (unclear)



Developed by Tsinghua University, largest Chinese MOOC platform, offers 504 MOOCs to 1,290,000 registered students from 126 countries (Ma, 2015)

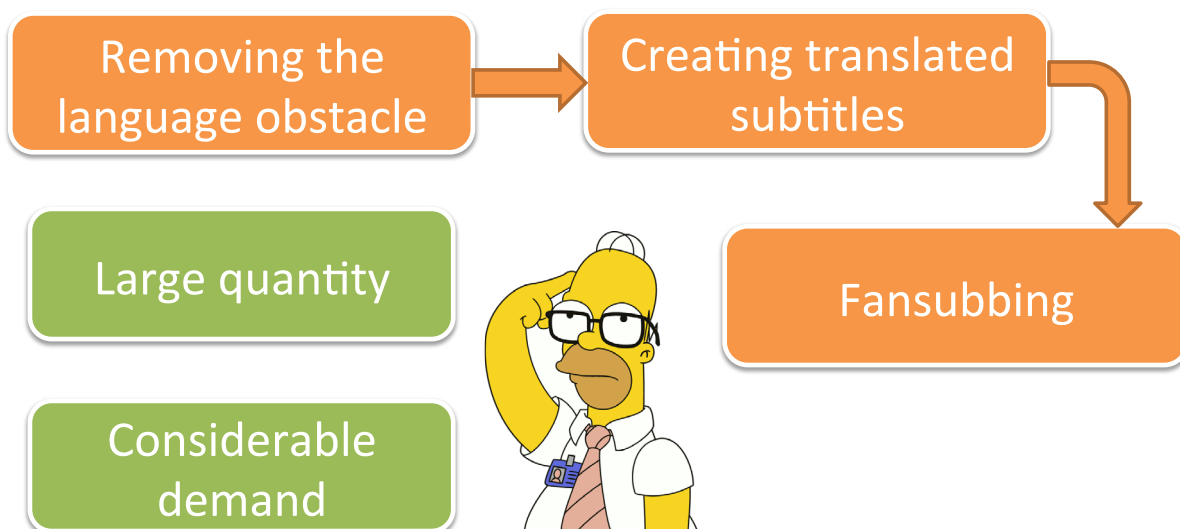


Survey by MOOC学院 (mooc.guokr.com) in 2014:



Language was a barrier to learning via MOOCs





Machine Translation!!



Why reception?

“It is not the software but the human side of the implementation cycle that will block progress in seeing that delivered systems are used effectively.”

-- Peter G. W. Keen (1991:1249)

Questions:

What are the needs of MT users?

What can affect user experience of MT?

How well do end users receive MT content?

...



Main research question:

Is there a difference in reception between participants who are offered raw MT subtitles and those who are offered full PEMT subtitles?

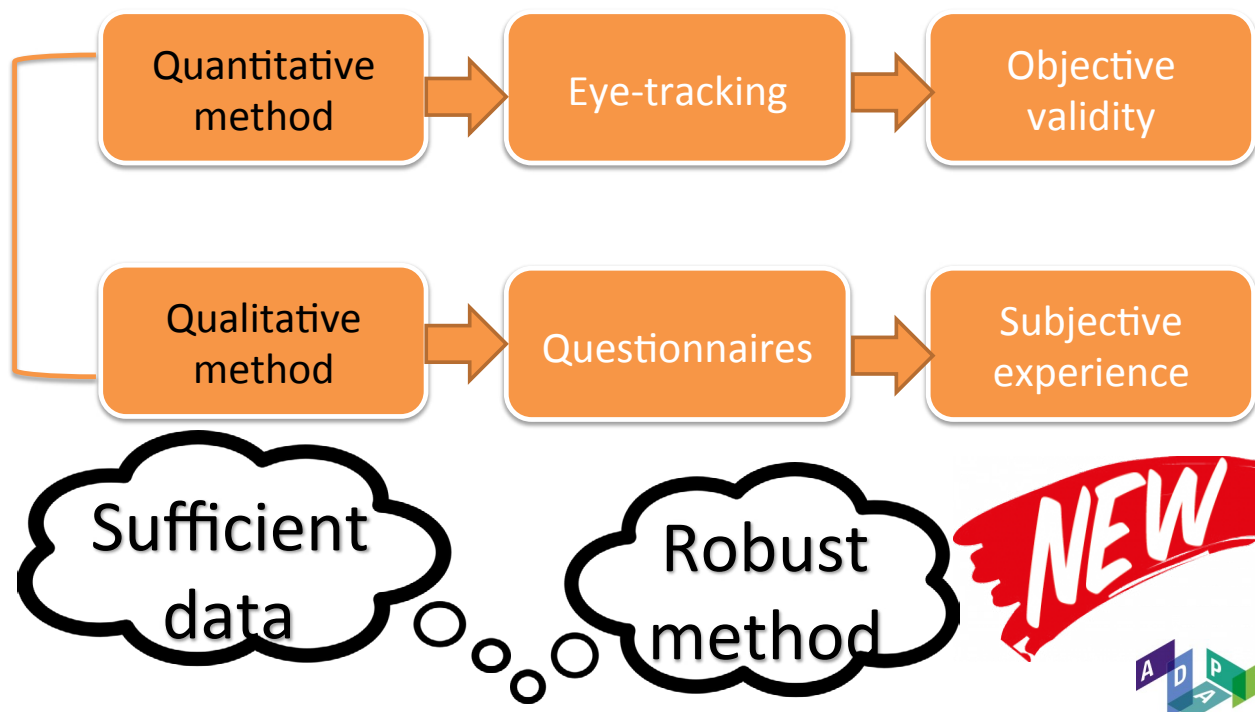
Main hypothesis:

Participants who are offered full PEMT subtitles will score higher on our reception metrics compared with those who are offered raw MT subtitles.



Methodologies

A mixed-methods approach



Element	Related to	Reflected in	Measured by
Response	Perceptual decoding	Attentional processes	Eye-tracking
Reaction	Psycho-cognitive issue	Processing effort and comprehension	Eye-tracking and comprehension testing
Repercussion	Attitudinal issues and sociocultural dimensions	Attitudes and beliefs	Background survey and attitude questions

Based on Gambier's model (Gambier, 2009)



Sub-hypotheses

Response:

Hypothesis 1: Fewer subtitles are skipped when participants are watching full PEMT subtitles. (measured by visit count)

Hypothesis 2: Relatively more attention is allocated to the image area when full PEMT subtitles are displayed than when raw MT subtitles are displayed. (measured by fixation count and visit duration)

Reaction:

Hypothesis 3: Comprehension score is higher with full PEMT subtitles. (measured by comprehension testing)

Hypothesis 4: Mean fixation duration is shorter when full PEMT subtitles are displayed.

Repercussion:

Hypothesis 5: Attitudes toward machine translation are better among participants shown full PEMT subtitles. (measured by attitude questions)



- Ø DCU, May 2017
- Ø Video: “What is physical activity?” (6”59”) under the MOOC “Sit Less, Get Active” on Coursera.
- Ø MT system: Google Translate (EN-ZH)
- Ø Two versions of subtitles (Number: 114 & 115)



Participants

- Ø MOOCs: university students, 18-25 years old
- Ø China: 50.94 out of 100, English Proficiency Index 2016 by EF



Ø Ideal participants: Chinese undergraduates with low English level

Ø Four Chinese participants (two groups)

Gender	1 female, 3 male
Age	22-33
Education	2 PhD students, 1 Post-doc, 1 final-year undergraduate
English level	1 intermediate 3 upper intermediate



Step 1: Pre-recruitment questionnaire & Online English test (Cambridge English Language Assessment)

Step 2: Watching MOOC video with eye-tracker (SMI REDn Scientific)

Step 3: Post-task questionnaire: comprehension testing (multiple choice) and attitude survey (five-point Likert scale)



Results

- 😓 All hypotheses were NOT supported by the results.
 - Tiny sample
- 😓 A few questions could be answered by common sense.
 - Questionnaire needs to be modified
- 😓 Vagaries of participants' memories and concentration.
 - Irresistible force



Main experiment in China!

Larger sample: over 30 Participants

One more group: human translated text added!

Statistical methods: ANOVA and t-test



Ke Hu: ke.hu2@mail.dcu.ie

Sharon O'Brien: sharon.obrien@dcu.ie

Dorothy Kenny: dorothy.kenny@dcu.ie



Recent Developments



Joss Moorkens & Yota Georgakopoulou

MT Summit XVI



Table of contents

- The TraMOOC Project
- NMT systems for TraMOOC
- Comparative Evaluation of Neural MT and Phrase-Based SMT
- Crowdsourced evaluations (explicit & implicit)
- Task-based evaluations

Joss Moorkens, Sheila Castilho, Federico Gaspari, Andy Way (DCU/ADAPT) – Ireland
Yota Georgakopoulou, Maria Gialama (Deluxe Media) – Greece/United Kingdom
Rico Sennrich, Antonio Valerio Miceli Barone (University of Edinburgh) - United Kingdom
Valia Kordoni, Markus Egg, Maja Popović (Humboldt University of Berlin) - Germany
Vilemini Sosoni (Ionian University, Corfu) - Greece
Iris Hendrickx (Radboud University Nijmegen) – The Netherlands
Menno van Zaanen (Tilburg University) – The Netherlands



9月2017年

Joss Moorkens & Yota Georgakopoulou



- **Reliable Machine Translation (MT) for Massive Open Online Courses (MOOCs)**
- The main expected outcome is a **high-quality semi-automated machine translation service** for educational text data on a MOOC platform
- Open educational platform for MT and a replicable process for creating such a service



9月2017年

Joss Moorkens & Yota Georgakopoulou



- Create domain-specific SMT NMT engines – 3 iterations
- Crowdsourced evaluation of MT quality
- Explicit and implicit evaluation stages
- Task-based evaluations
- Free and premium platform due 2018



9月2017年

Joss Moorkens & Yota Georgakopoulou



- Make existing monolingual educational material available to speakers of other languages
 - multi-genre and heterogeneous textual course material
 - Subtitles – video lectures
 - assignments
 - tutorial text
 - social web text posted on MOOC blogs and fora (questions/answers/comments)
- Reusing existing linguistic infrastructure and MT resources extending existing models
- Test on a MOOC platform and on the VideoLectures.Net digital video lecture library

- Users who want access to open online education that is not constrained by language barriers.
- MOOC providers, who wish to offer high-quality, integrated multilingual educational services.
- Machine Translation developers, who need a platform for promoting, testing and comparing their solutions.
- Language Technology Engineers, who want access to accurate and wide-coverage linguistic infrastructure, even for less widely spoken languages.

- 10 partners from 6 European countries
 - Humboldt University (Coordinator)
 - Dublin City University
 - University of Edinburgh
 - Ionian University
 - Radboud University
 - Tilburg University
 - Deluxe Media Europe LTD
 - Knowledge 4 All Foundation LTD
 - EASN Technology Innovation Services
 - (Iversity) HPI



9月2017年

Joss Moorkens & Yota Georgakopoulou



Which MT paradigm?

- Project had originally planned to compare Syntax-Based and Phrase-Based SMT
- Comparative Evaluation of Neural MT (Nematus) and Phrase-Based SMT (Moses)
- English to German, Greek, Portuguese, and Russian
- MT engines trained on open and educational data



9月2017年

Joss Moorkens & Yota Georgakopoulou



Chinese→English				German→English			
#	Ave %	Ave z	System	#	Ave %	Ave z	System
1	73.2	0.209	SogouKnowing-nmt	1	78.2	0.213	online-B
	73.8	0.208	uedin-nmt		76.6	0.169	online-A
	72.3	0.184	xmunmt		76.6	0.165	KIT
4	69.9	0.113	online-B	76.6	0.162	uedin-nmt	
	70.4	0.109	online-A	75.8	0.131	RWTH-nmt-ensem	
	69.8	0.079	NRC	74.5	0.098	SYSTRAN	
7	67.9	0.023	jhu-nmt	7	72.9	0.029	LIUM-NMT
	66.9	-0.016	afri-mitll-opennmt	8	70.2	-0.058	TALP-UPC
	67.1	-0.026	CASICT-cons		69.8	-0.072	online-G
	65.4	-0.058	ROCMT		68.6	-0.103	C-3MA
11	64.3	-0.107	Oregon-State-Uni-S	11	64.1	-0.260	online-F
12	61.7	-0.209	PROMT-SMT	English→German			
	61.2	-0.265	NMT-Ave-Multi-Cs	#	Ave %	Ave z	System
	60.0	-0.276	UU-HNMT	1	72.9	0.257	LMU-nmt-reranked
	59.6	-0.279	online-F	2	70.2	0.158	online-B
59.3	-0.305	online-G	69.8		0.139	uedin-nmt	
					68.9	0.092	SYSTRAN
					66.9	0.035	LMU-nmt-single
					66.7	0.022	KIT
					66.4	0.015	xmu
					66.6	0.006	LIUM-NMT
				66.0	-0.003	RWTH-nmt-ensem	
English→Chinese							
#	Ave %	Ave z	System				
1	73.2	0.208	SogouKnowing-nmt				
	72.5	0.178	uedin-nmt				
	72.0	0.165	xmunmt				
4	69.8	0.065	online-B				



- Main strength of NMT is grammatical improvements, but possible degradation in lexical transfer (Neubig, Morishita, Nakamura 2015)
- Output conditioned on full source text and target history
- Some problems:
 - Networks have fixed vocabulary → poor translation of rare/unknown words
 - Models are trained on parallel data; how do we use monolingual data?
 - Recent solutions:
 - Subword models allow translation of rare/unknown words (Sennrich, Birch, Haddow 2016a)
 - Train on back-translated monolingual data (Sennrich, Birch, Haddow 2016b)



- 4 datasets (250 segments) from EN MOOC data translated into German, Greek, Portuguese, and Russian using TraMOOC engine prototype 2
- PB-SMT/NMT mixed, random task order
- 2-4 professional translators in Deluxe Media
- **Detailed results presented by Sheila Castilho in Research Track and in proceedings of MT Summit XVI**



9月2017年

Joss Moorkens & Yota Georgakopoulou

11

- PBSMT
 - Moses, MGIZA is used to train word alignments, and KenLM is used for language model training and scoring (Huck and Birch 2015)
- NMT Tools Used:
 - Nematus: <https://github.com/rsennrich/nematus>
 - Amun: <https://github.com/amunmt/amunmt> (for deploying the models)
- Domain adaptation:
 - Models initially trained on all available data, then continually trained on in-domain data, which effectively adapts the system to the domain NMT



9月2017年

Joss Moorkens & Yota Georgakopoulou

12

- For all 4 language pairs:

FLUENCY
1. No fluency
2. Little fluency
3. Near native
4. Native

	EN-DE		EN-EL		EN-PT		EN-RU	
% scores assigned 3-4 fluency value (SMT, NMT)	54.2	67.6	65	75	73.8	79.5	60.2	75.1
% scores assigned 1-2 fluency value (SMT, NMT)	45.8	32.4	35	25	26.2	20.5	39.8	24.9



9月2017年

Joss Moorkens & Yota Georgakopoulou



- For all 4 language pairs:

ADEQUACY
1. None of it
2. Little of it
3. Most of it
4. All of it

	EN-DE		EN-EL		EN-PT		EN-RU	
% scores assigned 3-4 adequacy value (SMT, NMT)	73.5	66.4	89	89	94.7	97.1	72.8	77.5
% scores assigned 1-2 adequacy value (SMT, NMT)	26.5	33.6	11	11	5.3	2.9	27.2	22.5



9月2017年

Joss Moorkens & Yota Georgakopoulou



Words per second (all PEs)	SMT	NMT
German	0.21	0.22
Greek	0.22	0.24
Portuguese	0.29	0.30
Russian	0.14	0.14

Previous work by Moorkens & O'Brien (2015) found an average speed of 0.39 WPS for EN-DE professional PE.

SMT, NMT	German		Greek		Portuguese		Russian	
POST-EDITED SENTENCES (CHANGED)	940	813	928	863	874	844	930	848
UNCHANGED SMT, NMT	60	187	72	137	126	156	70	152



9月2017年

Joss Moorkens & Yota Georgakopoulou



- In this study, using these language pairs, in this domain...
- Fluency is improved, word order errors are fewer using NMT
- Fewer segments require editing using NMT
- NMT produces fewer morphological errors
- No clear improvement for omission or mistranslation using NMT
- NMT for production: no great improvement in post-editing throughput
 - “Errors are more difficult to spot”
- Based on the pace of improvement of NMT however, TraMOOC moved to NMT exclusively



9月2017年

Joss Moorkens & Yota Georgakopoulou



Using the Crowdflower platform for all 11 language pairs:

- *Clear instructions available during the entire translation procedure.*
- *Test Questions to validate the accuracy of the participants' input.*
- *Post-editing question should be displayed first, hiding the rest of the questions to avoid influencing the contributors' judgment.*
- *Fluency for ST and TT, adequacy and error mark-up for TT*
- *Multiple error mark-up supported.*

For QA and language coverage, 5-10% expert evaluation by DME

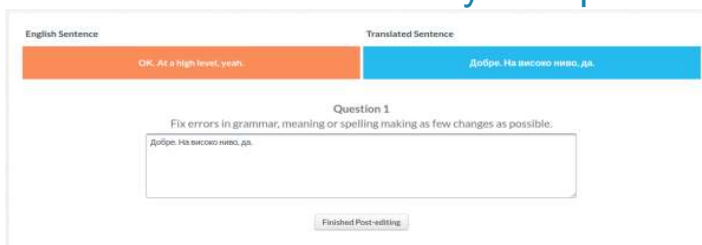


9月2017年

Joss Moorkens & Yota Georgakopoulou

17

- Post-editing (expert and crowd): “Make changes in the translation if there are errors in grammar, meaning or spelling”
 - Basic rules regarding spelling apply. If there are any typos or slight grammatical/syntactic mistakes in the original, please fix them in the translation
 - Do not implement corrections for stylistic reasons only
 - No need to restructure sentences only to improve the natural flow of the text

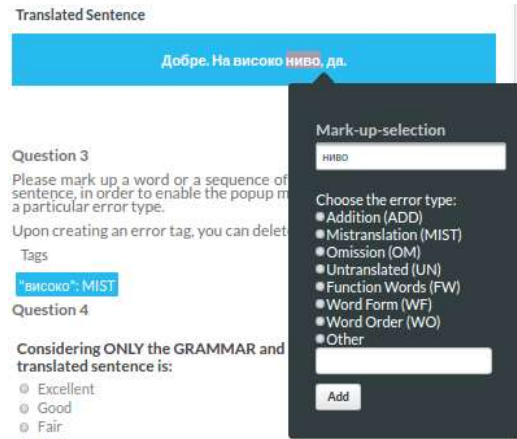



9月2017年

Joss Moorkens & Yota Georgakopoulou

18

- Change the mark-up error type list (for expert group) so as to map onto DQF-MQM typology: **Addition, Mistranslation, Omission, Untranslated, Function Words, Word Form, and Word Order**



9月2017年

Joss Moorkens & Yota Georgakopoulou



- Unforeseen delays:
 - Crowdsourcing contracts
 - Change of MOOC partner
- Crowd behaviour issues →

Crowd behaviour issue	Solution(s)
Malicious behaviour	Constant monitoring, manual and automated
Use of Google Translate	Source language text is an image. Manual check with Google Translate feature in Chrome.
BR performing EU-PT tasks	Target specific countries
No change, yet low score on quality	Popup alerts
Poor coverage/ low contributor flow	Increase HIT payment; expand geographical reach & channel; decrease contributor level; decrease text question difficulty



9月2017年

Joss Moorkens & Yota Georgakopoulou



Malicious behaviour	Solutions
Blank translations	Change tactics for test questions, binary evaluation answers, distributed randomly
Random symbols	Increase the minimum time per page
Repetitive answers	Increase contributors' level
Other language characters	Constant manual and script-based (automated) monitoring: Python scripts for blanks, Latin characters in non-Latin languages, etc.
Multiple malicious accounts	Customised alerts scripts (blanks, length, time per page, etc.); flag malicious contributors; ban specific channels



9月2017年

Joss Moorkens & Yota Georgakopoulou



Underway: Crowdsourced implicit evaluations

Implicit evaluation: Annotation of entities, topics and terms in the source and target texts

- Generate a thesaurus of tag-sets that allows for the **implicit evaluation** of MT output through the comparison of the source and target tag-sets

Activities:

1. Entity annotation via Wikification
2. Topic & sentiment annotation

Course Title: Business and Negotiation

Comment

This is such a poor transcript. I've completed a few MOOC courses in the past and they all had transcripts that helped students quickly review materials or search keywords. This doesn't look like a transcript, at least not a useful one, and doesn't look like subtitles either. I can't understand what it is and how could anyone use it. I had to go over and over some lines a few times in order to understand what they are saying. The subject was boring and the text was annoying. Misleading thread title :)

1) Is the topic of this comment mainly about:

The course
 The translation
 Something else

2) Does the post express a sentiment towards this topic that is:

Positive (happy, excited, enthusiastic, complimentary)
 Neutral/I don't know (unrelated comments, mixed feelings)
 Negative (hate, anger, sadness, frustration)

Positive Neutral Negative



9月2017年

Joss Moorkens & Yota Georgakopoulou



- openHPI - European MOOC platform plus TraMOOC API
 - Launched by the Hasso Plattner Institute (HPI) for Digital Engineering in Potsdam, Germany
- Users will be able to switch between the original course language and automatically translated content
- Users will be able to request translation for specific forum contributions
- Feedback via surveys on the translation content and the integration of the translation tools into the openHPI platform



9月2017年

Joss Moorkens & Yota Georgakopoulou



ありがとうございました



ありがとう



9月2017年

Joss Moorkens & Yota Georgakopoulou

