The EXPERT Project: Advancing the State of the Art in Hybrid Translation Technologies

Constantin Orăsan[a], Alessandro Cattelan[b], Gloria Corpas Pastor[c], Josef van Genabith[d], Manuel Herranz[e], Juan José Arevalillo[f], Qun Liu[g], Khalil Sima'an[h] and Lucia Specia[i]

[a]University of Wolverhampton, UK, C.Orasan@wlv.ac.uk
[b]Translated, Italy, Alessandro@translated.net
[c]University of Malaga, Spain, GCorpas@uma.es
[d]Saarland University, Germany, Josef.Van_Genabith@dfki.de
[e]Pangeanic, Spain, M.Herranz@pangeanic.com
[f]Hermes, Spain, Juanjo.Arevalillo@hermestrans.com
[g]Dublin City University, Ireland, QLiu@computing.dcu.ie
[h]University of Amsterdam, Netherlands, K.Simaan@uva.nl
[i]Sheffied University, UK, L.Specia@sheffield.ac.uk

Abstract

This paper gives a brief overview of the EXPloiting Empirical appRoaches to Translation (EXPERT) project, an FP7 Marie Curie Initial Training Network, which is preparing the next generation of world-class researchers in the field of hybrid machine translation. The project is employing 15 Marie Curie fellows who are working on 15 individual, but interconnected, projects and is organising local and consortium wide training activities. The project has been running for three years and has already produced high-quality research. This paper presents the most important research achievements of the project.

1 Introduction

Machine translation is playing an increasingly important role in our multilingual society, but in many cases the technology is not mature enough to be able to produce high-quality translations completely automatically. Current research is addressing this problem by developing better translation methods and by improving the way human translators can use computers in the translation process. Despite its importance, the field is lacking enough world-class researchers to ensure its fast progress. This paper gives a brief overview of the EXPloiting Empirical appRoaches to Translation (EXPERT) project[1], an FP7 Marie Curie Initial Training Network which is focusing on these issues.

The purpose of the EXPERT project is two-fold. As a training network, the project is preparing 15 Marie Curie fellows to become future leaders in the field. This is achieved by employing 12 Early Stage Researchers (ESRs) and three Experienced Researchers (ERs) at one of the nine partners in the consortium, by organising dedicated training events and enabling intersectoral and transnational secondments. The researchers employed in the project work together with established researchers from the consortium to promote the research, development and use of hybrid language translation technologies. All the ESRs are registered on PhD programs at their hosting institutions and complete secondments at partner institutions in order to experience different sectors and develop transferable skills. The ERs are employed by the industrial partners and are developing commercial solutions based on some of the research carried out by the ESRs.

The project is delivered by a consortium coordinated by University of Wolverhampton, UK and which contains five other academic partners: University of Malaga, Spain; University of Sheffied, UK; Saarland University, Germany; Dublin City University, Ireland and University of

[1]http://expert-itn.eu

Proceedings of the 37th Conference Translating and the Computer, pages 18–23,
London, UK, November 26-27, 2015.

Amsterdam, Netherlands, as well as three industrial partners: Pangeanic, Spain; Translated, Italy and Hermes, Spain. In addition, the consortium benefits from the contribution of four associated partners: WordFast, France; Etrad, Argentina; Unbabel, Portugal and DFKI, Germany. The project started on the 1st Oct 2012 and has just completed the third year, with one more year left.

2 Description of the Research Carried Out in the Project

The researchers employed in the project are working on 15 individual, but related, projects which aim to improve the state of the art from five different directions: the user perspective, data collection and preparation, incorporation of language technology in translation memories, the human translator in the loop, and hybrid approaches to translation. This section gives an overview of the main achievements so far in each of these directions.

2.1 The User Perspective

The large number of tools available and the plethora of features that professional translators can access create challenges to professional translators when they try to integrate these tools in their translation workflow. This is largely due to the fact that in many cases the real needs of translators were not considered when designing these tools. To this end, a survey with professional translators was carried out in order to find out their views and requirements regarding various technologies, and their current work practices. Thanks to the help of the commercial partners in the project, the survey received 736 complete responses, from a total of over 1300 responses, which is more than in other similar surveys. A first analysis of the data is presented in (Zaretskaya et al., 2015) with more analyses underway.

Parra Escartín (2015) carried out another study with professional translators in an attempt to find out "missing functionalities" of translation memories that could potentially improve their productivity. An interesting feature suggested was to generate segments on fly from fragments of previously translated segments. An implementation based on pattern matching showed that even such a simple approach can be potentially useful.

Another way to address the needs of translators is to design flexible interfaces. Lewis et al. (2014) propose a framework in which new components of a user interface can be consistently tested, compared and optimised based on user feedback. HandyCAT is an implementation of the proposed framework.

The output of machine translation systems is usually evaluated using standard metrics such as BLEU (Papineni et al., 2002). However, these metrics are not necessarily that useful to translation companies. To this end, research is currently going on to develop a method that can predict the post-editing effort required by a given sentence (Béchara, 2015; Parra Escartín and Arcedillo, 2015a; Parra Escartín and Arcedillo, 2015b; Parra Escartín and Arcedillo, 2015c).

2.2 Data Collection and Preparation

Given that the focus of the EXPERT project is on data-driven translation technologies, a significant amount of work is dedicated to collecting and preparing of relevant data. Costa et al. (2014) shows how it is possible to compile comparable corpora from the Internet using distributional similarity measures. This method is currently being integrated in a web-based application capable of semi-automatically compiling multilingual comparable and parallel corpora (Costa et al., 2015a).

Resources like MyMemory[2] contain large number of bi-segments that can be used in translation memories, but not all the bi-segments are true translations. For this reason,

[2]https://mymemory.translated.net/

Barbu (2015) proposed a method based on machine learning for cleaning existing translation memories.

2.3 Incorporation of Language Technology in Translation Memories

Translation memories are among the most successfully used tools by professional translators. However, most of these tools rely on little language processing when they match and retrieve segments. Research carried out in the EXPERT project shows that even incorporation of simple language processing such as paraphrasing can help translators (Gupta and Orăsan, 2014). Rather than expanding the segments stored in a translation memory with all the possible paraphrases, the proposed method incorporates paraphrases in the edit distance algorithm. An experiment with human translators shows that by using paraphrasing it is possible to reduce the number of keystrokes required to produce a correct translation by 33%, whilst the time reduces by 10% (Gupta et al., 2015). Integration of this technology in a real-world environment is currently being explored.

An alternative way of improving the retrieval from translation memories is by integrating relevant ontologies and terminology databases. However, it is not unusual that these resources are not available for all the domains. To this end, Tan and Pal (2014) proposed several methods for terminology extraction and ontology induction with the aim of integrating them in translation memories and statistical machine translation.

2.4 The Human Translator in the Loop

Post-editing is one of the most promising ways of integrating the output of machine translation methods in the workflows used by translation companies. Quality estimation methods are used to decide whether a sentence should be translated from scratch or it is good enough to be given to a post-editor. Most of the existing methods focus on estimating the quality of sentences, but in some cases it is necessary to estimate the quality of the translation of a whole document. The work carried out by Scarton and Specia (2014) in the EXPERT project focuses on document level quality estimation.

Automatic post-editing provides an additional way to simplifying the work of professional translators. Pal (2015) shows how it is possible to apply Hierarchical Phrase Based Statistical Machine Translation to the task of monolingual Statistical Automatic Post-editing. Evaluation using standard MT metrics shows that automatically post-edited texts are better than the raw translations. In addition, an experiment with four professional translators reveals that the post-editing effort is also reduced.

Logacheva and Specia (2015) investigate ways to collect and generate negative human feedback in various forms, including post-editing, and learn how to improve machine translation systems from this feedback, for example, by building word-level quality estimation models to mimic user feedback and introducing the predictions in SMT decoders.

2.5 Hybrid Approaches to Translation

All the existing methods in MT have strengths and weaknesses and one of the most common ways to improve their performance is to combine them. Li et al. (2014) proposed a method for incorporating translation memories and linguistic knowledge in SMT, showing that for English-Chinese and English-French the proposed methods lead to better translations.

Translation into morphological rich languages poses challenges to current methods in statistical machine translation. For this problem, Daiber and Sima'an (2015) propose a method which consists of two steps: first the source string is enriched with target morphological features and then fed into a translation model which takes care of reordering and lexical choice that

matches the provided morphological features. The resulting system performs better than a baseline phrase-based system.

The quality of SMT systems depends very much on the data they are trained. Cuong and Sima'an (2014b) propose a new statistical approach which works in two steps: first it exploits the in-domain data to identify least relevant instances, which it considers as pseudo-out-domain corpus, and secondly, it trains a novel full latent domain translation model aiming at measuring the degree of relevance for each instance in the mix-domain corpus using the statistical contrast between in-domain and pseudo-out-domain data. Continuing this line of research, Cuong and Sima'an (2014a) present a new method for domain adaptation for phrase-based models based on estimating latent domain variable statistics over phrase pairs from large heterogenuous parallel corpora, whilst Cuong and Sima'an (2015) proposes a new latent domain approach to word alignments and shows the advantages over the domain agnostic methods.

3 Training Activities

In order to successfully prepare the researchers for their future career, the EXPERT project also organises local and consortium-wide training events. The local training events focus on skills specific to the research carried out at that site, whereas the consortium-wide training events are delivered for the whole consortium and focus on skills and knowledge relevant to all the fellows employed in the project. The consortium has already organised three training events, with a final one planned for the fourth year. Slides from all the training events are available on the project's website.

The first training event delivered scientific and technical training which covered the fundamental themes of the EXPERT project. It was organised once most of the researchers were appointed and ensured that all of them acquired the necessary background to complete their projects.

The second training event focused on complementary skills and prepared researchers for the planning and exploitation of their research outcomes and improving their career prospects for jobs in industry and academia.

A scientific and technological workshop gave the opportunity to all the researchers employed in the project to present their work so far and to interact with other researchers, both employed on the project and attending the workshop. This third training event was organised as a mini conference where all the EXPERT fellows had to prepare and present a paper. A volume containing all the papers was produced (Costa et al., 2015b).

The final training event will be a business showcase for the tools developed by the ERs in the project. It will give them the opportunity to disseminate the outcomes of EXPERT to potential end-users: translators and the general public.

4 Conclusions

This paper has presented a brief overview of the EXPERT project, a Marie Curie Initial Training Network which focuses on hybrid translation technologies. The project has been running for three years and has already produced significant high quality research and trained excellent researchers.

Acknowledgements

The research presented in this paper is supported by the People Programme (Marie Curie Actions) of the European Union's Framework Programme (FP7/2007-2013) under REA grant agreement no 317471.

References

Eduard Barbu. 2015. Spotting false translation segments in translation memories. In *Proceedings of the International Workshop on Natural Language Processing for Translation Memories (NLP4TM)*, pages 9 – 16, Hissar, Bulgaria.

Hanna Béchara. 2015. The Role of Semantic Textual Similarity in Machine Translation Evaluation. Technical report, University of Wolverhampton, Wolverhampton, UK.

Hernani Costa, Gloria Corpas Pastor, and Miriam Sighiri. 2014. iCompileCorpora: A Web-based Application to Semi-automatically Compile Multilingual Comparable Corpora. In *Translating and the Computer 36 - AsLing*, pages 51 – 55, London, UK.

Hernani Costa, Gloria Corpas Pastor, Ruslan Mitkov, and Miriam Sighiri. 2015a. Towards a Web-based Tool to Semi-automatically Compile, Manage and Explore Comparable and Parallel Corpora. In *Proceedings of AIETI7 Conference: New Horizons in Translation and Interpreting Studies*, Malaga, Spain.

Hernani Costa, Anna Zaretskaya, Gloria Corpas Pastor, Lucia Specia, and Miriam Seghiri, editors. 2015b. *Proceedings of the EXPERT Scientific and Technological Workshop*. Malaga, Spain.

Hoang Cuong and Khalil Sima'an. 2014a. Latent Domain Phrase-based Models for Adaptation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 566 – 576, Doha, Qatar.

Hoang Cuong and Khalil Sima'an. 2014b. Latent Domain Translation Models in Mix-of-Domains Haystack. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1928 – 1939, Dublin, Ireland.

Hoang Cuong and Khalil Sima'an. 2015. Latent Domain Word Alignment for Heterogeneous Corpora. In *Proceedings of The 2015 Annual Conference of the North American Chapter of the ACL*, pages 398 – 408, Denver, Colorado.

Joachim Daiber and Khalil Sima'an. 2015. Machine Translation with Source-Predicted Target Morphology. In *Proceedings of MT Summit XV*, Miami, Florida.

Rohit Gupta and Constantin Orăsan. 2014. Incorporating Paraphrasing in Translation Memory Matching and Retrieval. In *Proceedings of the Seventeenth Annual Conference of the European Association for Machine Translation (EAMT2014)*, pages 3 – 10, Dubrovnik, Croatia.

Rohit Gupta, Constantin Orăsan, Marcos Zampieri, Mihaela Vela, and Josef Van Genabith. 2015. Can Translation Memories afford not to use paraphrasing? In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation (EAMT)*, pages 35 – 42, Antalya, Turkey.

David Lewis, Qun Liu, Leroy Finn, Chris Hokamp, Felix Sasaki, and David Filip. 2014. Open, web-based internationalization and localization tools. *Translation Spaces*, 3:99 – 132.

Liangyou Li, Andy Way, and Qun Liu. 2014. A Discriminative Framework of Integrating Translation Memory Features into SMT. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas*, volume 1, pages 249–260, Vancouver, Canada.

Varvara Logacheva and Lucia Specia. 2015. The role of artificially generated negative data for quality estimation of machine translation. In *18th Annual Conference of the European Association for Machine Translation*, pages 51 – 58, Antalya, Turkey.

Santanu Pal. 2015. Statistical Automatic Post Editing. In *Proceedings of the EXPERT Scientific and Technological Workshop*, pages 13 – 22, Malaga, Spain.

Kishore Papineni, Salim Roukos, Todd Ward, and Wj Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics Annual Meeting (ACL)*, pages 311 – 318, Philadelphia, Pennsylvania.

Carla Parra Escartín and Manuel Arcedillo. 2015a. A fuzzier approach to machine translation evaluation: A pilot study on post-editing productivity and automated metrics in commercial settings. In *Proceedings of the ACL 2015 Fourth Workshop on Hybrid Approaches to Translation (HyTra)*, pages 40–45, Beijing, China.

Carla Parra Escartín and Manuel Arcedillo. 2015b. Living on the edge: productivity gain thresholds in machine translation evaluation metrics. In *Proceedings of The Fourth Workshop on Post-editing Technology and Practice*, Miami, Florida.

Carla Parra Escartín and Manuel Arcedillo. 2015c. Machine translation evaluation made fuzzier: A study on post-editing productivity and evaluation metrics in commercial settings. In *Proceedings of MT Summit XV*, Miami, Florida.

Carla Parra Escartín. 2015. Creation of new TM segments : Fulfilling translators' wishes. In *Proceedings of the International Workshop on Natural Language Processing for Translation Memories (NLP4TM)*, pages 1 – 8, Hissar, Bulgaria.

Carolina Scarton and Lucia Specia. 2014. Document-level translation quality estimation: exploring discourse and pseudo-references. In *Proceedings of the Seventeenth Annual Conference of the European Association for Machine Translation (EAMT2014)*, pages 101 – 108, Dubrovnik, Croatia.

Liling Tan and Santanu Pal. 2014. Manawi: Using multi-word expressions and named entities to improve machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 201 – 206, Baltimore, Maryland, USA.

Anna Zaretskaya, Gloria Corpas Pastor, and Miriam Sighiri. 2015. Translators' requirements for translation technologies: Results of a user survey. In *Proceedings of AIETI7 Conference: New Horizons in Translation and Interpreting Studies*, Malaga, Spain.