2nd Workshop on Natural Language Processing for Translation Memories (NLP4TM 2016)

28 May 2016

ABSTRACTS

Editors:

Constantin Orasan, Carla Parra, Eduard Barbu, Marcello Federico

Workshop Programme

9:00 – 10:30: Session 1: Invited talks

9:00 – 9:30
Marcello Federico, *Machine translation adaptation from translation memories in ModernMT*

9:30 – 10:00
Núria Bel, *Data fever in the 21st century. Where to mine Language Resources*

10:00 – 10:30
Samuel Läubli, *Data, not Systems: a Better Way to Conduct the Business of Translation*

10:30 – 11:00 Coffee break

11:00 – 12:00: Session 2: Research papers

11:00 – 11:20
A. Bellandi, G. Benotto, G. Di Segni, E. Giovannetti, *Investigating the Application and Evaluation of Distributional Semantics in the Translation of Humanistic Texts: a Case Study.*

11:20 – 11:40
Tapas Nayak, Santanu Pal, Sudip Kumar Naskar, Sivaji Bandyopadhyay, Josef van Genabith, *Beyond Translation Memories: Generating Translation Suggestions based on Parsing and POS Tagging*

11:40 – 12:00
Friedel Wolff, Laurette Pretorius, Loïc Dugast, Paul Buitelaar, *Methodological pitfalls in automated translation memory evaluation*

12:00 – 13:30: Session 3: Cleaning of translation memories shared task

12:00 – 12:20
Eduard Barbu, Carla Parra Escartín, Luisa Bentivogli, Matteo Negri, Marco Turchi, Marcello Federico, Luca Mastrostefano, Constantin Orasan, *1st Shared Task on Automatic Translation Memory Cleaning Preparation and Lessons Learned*

12:20 – 13:00
Presentations of the systems that took part in the shared task

13:00 – 13:30
Round table

Workshop Organizers

Constantin Orasan	University of Wolverhampton, UK
Carla Parra	Hermes, Spain
Eduard Barbu	Translated, Italy
Marcello Federico	FBK, Italy

Workshop Programme Committee

Juanjo Arevalillo	Hermes, Spain
Yves Champollion	WordFast, France
Gloria Corpas	University of Malaga, Spain
Maud Ehrmann	EPFL, Switzerland
Kevin Flanagan	Swansea University, UK
Corina Forascu	University “Al. I. Cuza”, Romania
Gabriela Gonzalez	eTrad, Argentina
Rohit Gupta	University of Wolverhampton, UK
Manuel Herranz	Pangeanic, Spain
Samuel Läubli	Autodesk, Switzerland
Liangyou Li	DCU, Ireland
Qun Liu	DCU, Ireland
Ruslan Mitkov	University of Wolverhampton, UK
Aleksandros Poulis	Lionbridge, Sweden
Gabor Proszeky	Morphologic, Hungary
Uwe Reinke	Flensburg University of Applied Sciences, Germany
Michel Simard	NRC, Canada
Mark Shuttleworth	UCL, UK
Masao Utiyama	NICT, Japan
Mihaela Vela	Saarland University, Germany
Andy Way	DCU, Ireland
Joern Wuebker	Lilt, United States
Marcos Zampieri	Saarland University and DFKI, Germany

Preface

Translation Memories (TM) are amongst the most used tools by professional translators, if not the most used. The underlying idea of TMs is that a translator should benefit as much as possible from previous translations by being able to retrieve how a similar sentence was translated before. Moreover, the usage of TMs aims at guaranteeing that new translations follow the client's specified style and terminology. Despite the fact that the core idea of these systems relies on comparing segments (typically of sentence length) from the document to be translated with segments from previous translations, most of the existing TM systems hardly use any language processing for this. Instead of addressing this issue, most of the work on translation memories focused on improving the user experience by allowing processing of a variety of document formats, intuitive user interfaces, etc.

The term second generation translation memories has been around for more than ten years and it promises translation memory software that integrates linguistic processing in order to improve the translation process. This linguistic processing can involve matching of subsentential chunks, edit distance operations between syntactic trees, incorporation of semantic and discourse information in the matching process. Terminologies, glossaries and ontologies are also very useful for translation memories, by facilitating the task of the translator and ensuring a consistent translation. The field of Natural Language Processing (NLP) has proposed numerous methods for terminology extraction and ontology extraction. The building of translation memories from corpora is another field where methods from NLP can contribute to improving the translation process.

We are happy we could include in the workshop programme four contributions dealing with the aforementioned issues. In addition, the programme of the workshop in complemented by the presentations of three well-known researchers.

The first edition of this workshop organised at RANLP 2015 confirmed the fact that there is interest in the research community for the topics proposed. In addition, it highlighted the need for automatic methods for cleaning translation memories. For this reason, the second edition of the NLP4TM workshop also organises a shared task on cleaning translation memories in an attempt to make the creation of resources for translation memories easier.

The Organising Committee would like to thank the Programme Committee, who responded with very fast but also substantial reviews for the workshop programme. This workshop would not have been possible without the support received from the EXPERT project (FP7/2007-2013 under REA grant agreement no. 317471, http://expert-itn.eu).

Session 1: Invited talks

Saturday 28 May, 9:00 – 10:30

Machine translation adaptation from translation memories in ModernMT

Marcello Federico, FBK Trento, Italy

Adapting machine translation systems to specific customers or domains usually requires long time and extensive effort in training and optimizing the system. ModernMT sets out to do away with this by developing a new MT technology that seamlessly integrate translation memories into the MT system and train it on the fly without any disruption of service nor any user intervention. ModernMT is a new MT technology funded by the European Union that will overcome the typical limitations of traditional MT systems. From the software engineer perspective, ModernMT will provide an easy to install, fast to train, and simple to scale platform, capable to simultaneously serve tens of thousands of translators working simultaneously. Users will be able to integrate ModernMT in their favorite CAT tool, such as MateCat and Trados Studio, via a plugin that will merge translation memory and machine translation functions. In particular, by uploading and connecting their private translation memory to ModernMT, and by updating it during their work, they will not only receive better matches, but also more appropriate machine translation suggestions, that will significantly enhance their user experience and productivity. Real-time training and adaptation of machine translation from multiple translation memories, however, requires very efficient processing, e.g. for text cleaning, tokenisation, tag management, word alignment, adaptation, etc. In my talk I will present the overall ModernMT architecture, discuss its development roadmap and report preliminary results.

Data fever in the 21st century. Where to mine Language Resources.

Núria Bel, Universitat Pompeu Fabra, Spain

Language Resources, especially parallel corpora and bilingual glossaries, are raw materials for a number of application tasks and are considered a critical supply for Natural Language Processing-based applications such as Machine Translation. More and more efforts are being made with the aim of finding and exploiting deposits of multilingual data. The web is considered the most obvious mine: special crawlers are devised to find multilingual webs from where to extract parallel corpora. But other sources are also being explored such as open data. Open data is a quite promising source of data, particularly in the case of public administration data, although some issues concerning access and formats must be taken into consideration. Moreover, Linked Open Data has also proved to be useful for producing multilingual glossaries.

In this presentation, I will review different initiatives to mine Multilingual Language Resources which might be of interest for producing domain-specific Translation Memories for different language pairs. Besides the creation of new Translation Memories, these resources may also be used for enriching, curating or quality assuring already existing ones.

Data, not Systems: a Better Way to Conduct the Business of Translation

Samuel Läubli, Autodesk Development S.à.r.l., Switzerland

Over the past two decades, Autodesk has been acquiring large volumes of professional translations through localizing software products into more than 20 languages. Besides classical translation memory leveraging, we use this data for providing natural language processing services such as full text search, terminology harvesting, or domain-specific statistical machine translation.

While these services have been shown to positively impact translation quality and/or throughput (e.g., Plitt & Masselot, 2010), the fact that they are normally accessed via purpose-built systems creates inefficiencies for providers and consumers alike. From a corporate perspective, the effort needed for their maintenance and support would be better spent on improving the services as such. Translators, on the other hand, lose time in familiarizing themselves with client-specific systems as well as switching between them and their usual translation workbench. Recent research suggests that bundling these components into mixed-initiative interfaces (Horvitz, 1999) makes translation much more efficient and rewarding (Green et al., 2015).

In this talk, I will detail our transition from providing translators with the data rather than the software we think they need.

Session 2: Research papers

Saturday 28 May, 11:00 – 12:00

Investigating the Application and Evaluation of Distributional Semantics in the Translation of Humanistic Texts: a Case Study

A. Bellandi, G. Benotto, G. Di Segni, E. Giovannetti

Digital Humanities are persisting ascending and the need for translating humanistic texts using Computer Assisted Translation (CAT) tools demands for a specific investigation both of the available technologies and of the evaluation techniques. Indeed, humanistic texts can present deep differences from texts that are usually translated with CAT tools, due to complex interpretative issues, the request of heavy rephrasing, and the addition of explicative parts in order to make the translation fully comprehensible to readers and, also, stylistically pleasant to read. In addition, these texts are often written in peculiar languages for which no linguistic analysis tool can be available. We faced this situation in the context of the project for the translation of the Babylonian Talmud from Ancient Hebrew and Aramaic into Italian. In this paper we describe a work in progress on the application of distributional semantics to the informing of the Translation Memory, and on the evaluation issues arising from its assessment.

Beyond Translation Memories: Generating Translation Suggestions based on Parsing and POS Tagging

Tapas Nayak, Santanu Pal, Sudip Kumar Naskar, Sivaji Bandyopadhyay,
Josef van Genabith

This paper explores how translations of unmatched parts of an input sentence can be discovered and inserted into Translation Memory (TM) suggestions generated by a Computer Aided Translation (CAT) tool using a parse tree and part of speech (POS) tags to form a new translation which is more suitable for post-editing. CATaLog (Nayek et al., 2015) is a CAT tool based on TM and a modified Translation Error Rate (TER) (Snover et al., 2006) metric. Unmatched parts of the sentence to be translated can often be found in some other TM suggestions or in sentences which are not part of TM suggestions. Therefore, we can find the translations of those unmatched parts within the TM database itself. If we can merge the translations of the unmatched parts into one single sentence in a meaningful way, then post-editing effort will be reduced. Inserting the translations for the unmatched parts into TM suggestions may lead to loss of fluency in the generated target sentence. To avoid that, we use parsing and POS tagging together with a back off POS n-gram model to generate new translation suggestions.

Methodological pitfalls in automated translation memory evaluation

Friedel Wolff, Laurette Pretorius, Loïc Dugast, Paul Buitelaar

A translation memory system attempts to retrieve useful suggestions from previous translations to assist a translator in a new translation task. While assisting the translator with a specific segment, some similarity metric is usually employed to select the best matches from previously translated segments to present to a translator. Automated methods for evaluating a translation memory system usually use reference translations and also use some similarity metric. Such evaluation methods might be expected to assist in choosing between competing systems. No single evaluation method has gained widespread use; additionally the similarity metric used in each of these methods are not standardised either. This paper investigates the choice of fuzzy threshold during evaluation, and the consequences of different choices of similarity metric in such an evaluation method. Important considerations for automated evaluation of translation memory systems are presented.

Session 3: Shared task on cleaning of translation memories

Saturday 28 May, 12:00 – 13:30

1st Shared Task on Automatic Translation Memory Cleaning Preparation and Lessons Learned

Eduard Barbu, Carla Parra Escartín, Luisa Bentivogli, Matteo Negri, Marco Turchi, Marcello Federico, Luca Mastrostefano, Constantin Orasan

This paper summarizes the work done to prepare the first shared task on automatic translation memory cleaning. This shared task aims at finding automatic ways of cleaning TMs that, for some reason, have not been properly curated and include wrong translations. Participants in this task are required to take pairs of source and target segments from TMs and decide whether they are right translations. For this first task three language pairs have been prepared: English – Spanish, English – Italian, and English – German. In this paper, we report on how the shared task was prepared and explain the process of data selection and data annotation, the building of the training and test sets and the implemented baselines for automatic classifiers comparison.