RANLP 2015

Natural Language Processing for Translation Memories (NLP4TM)

Proceedings of the Workshop

September 11, 2015
Hissar, Bulgaria

The Workshop on
Natural Language Processing
for Translation Memories (NLP4TM)
associated with THE INTERNATIONAL CONFERENCE
RECENT ADVANCES IN
NATURAL LANGUAGE PROCESSING 2015

PROCEEDINGS

Hissar, Bulgaria
11 September 2015

Introduction

Translation Memories (TM) are amongst the most used tools by professional translators. 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. 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, and so on.

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 the matching of subsentential chunks, the editing of distance operations between syntactic trees, and the incorporation of semantic and discourse information in the matching process. Terminologies, glossaries and ontologies are also very useful for translation memories, 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 which can be integrated in the translation process. 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 4 long contributions and 3 short papers dealing with the aforementioned issues.

Vít Baisa, Aleš Horák and Marek Medved' discuss in *Increasing Coverage of Translation Memories with Linguistically Motivated Segment Combination Methods* how it is possible to extend existing translation memories using linguistically motivated segments combining approaches concentrated on preserving high translational quality.

Linked to the topic of enhancing existing translation memories, in the paper *Spotting false translation segments in translation memories* Eduard Barbu presents a method for identifying false translations in translation memories thought as a classification task.

In *Improving translation memory fuzzy matching by paraphrasing*, Konstantinos Chatzitheodorou explores the use of paraphrasing in retrieving better segments from translation memories. The method relies on NooJ and performs consistently better than the state of the art on EN-IT language pair.

CATaLog: New Approaches to TM and Post Editing Interfaces presents a new CAT tool by Tapas Nayek, Sudip Kumar Naskar, Santanu Pal, Marcos Zampieri, Mihaela Vela and Josef van Genabith. The aim of the tool is to improve both the performance and the productivity of post-editing.

Carla Parra Escartín aims at bridging the gap between academic research on Translation Memories (TMs) and the actual needs and wishes of translators in *Creation of new TM segments: Fulfilling translators' wishes*. She presents a pilot study where the requests of translators are being implemented in the translation workflow.

Katerina Timonera and Ruslan Mitkov explore the use of clause splitting in better retrieval of segments. Their paper *Improving Translation Memory Matching through Clause Splitting* show that their method leads to a statistically significant increase in the number of retrieved matches when both the input segments and the segments in the TM are first processed with a clause splitter.

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).

Constantin Orăsan and Rohit Gupta

Organizers:

Constantin Orăsan, University of Wolverhampton, UK

Rohit Gupta, University of Wolverhampton, UK

Program Committee:

Manuel Arcedillo, Hermes, Spain
Juanjo Arevalillo, Hermes, Spain
Eduard Barbu, Translated, Italy
Yves Champollion, WordFast, France
Gloria Corpas, University of Malaga, Spain
Maud Ehrmann, EPFL, Switzerland
Kevin Flanagan, Swansea University, UK
Gabriela Gonzalez, eTrad, Argentina
Manuel Herranz, Pangeanic, Spain
Qun Liu, DCU, Ireland
Ruslan Mitkov, University of Wolverhampton, UK
Gabor Proszeky, Morphologic, Hungary
Uwe Reinke, Cologne University of Applied Sciences, Germany
Michel Simard, NRC, Canada
Mark Shuttleworth, UCL, UK
Masao Utiyama, NICT, Japan
Andy Way, DCU, Ireland
Marcos Zampieri, Saarland University and DFKI, Germany
Ventsislav Zhechev, Autodesk

Invited Speaker:

Marcello Federico, Fondazione Bruno Kessler, Italy

Table of Contents

Conference Program

9:00–9:10 *Welcome*
Constantin Orasan

9:10–10:10 *Automatically tidying up and extending translation memories (invited talk)*
Marcello Federico, FBK, Italy

10:15–11:00 *Creation of new TM segments: Fulfilling translators' wishes*
Carla Parra Escartín

11:30–12:15 *Spotting false translation segments in translation memories*
Eduard Barbu

12:15–13:00 *Improving Translation Memory Matching through Clause Splitting*
Katerina Raisa Timonera and Ruslan Mitkov

14:30–15:00 *Improving translation memory fuzzy matching by paraphrasing*
Konstantinos Chatzitheodoroou

15:00–15:30 *Increasing Coverage of Translation Memories with Linguistically Motivated Segment Combination Methods*
Vít Baisa, Ales Horak and Marek Medveď

15:30–16:00 *CATaLog: New Approaches to TM and Post Editing Interfaces*
Tapas Nayek, Sudip Kumar Naskar, Santanu Pal, Marcos Zampieri, Mihaela Vela and Josef van Genabith

16:15–17:00 ***Round table and final discussion***