SCATE – Smart Computer-Aided Translation Environment – Year 3 (/4)

Peer-reviewed author version

Vandeghinste, Vincent; Vanallemeersch, Ton; Augustinus, Liesbeth; Van Eynde, Frank; Pelemans, Joris; Verwimp, Lyan; Wambacq, Patrick; Heymans, Geert; Moens, Marie-Francine; van de Lek-Ciudin, Iulianna; Steurs, Frieda; Rigouts, Terryn Ayla; Lefever, Els; Tezcan, Arda; Macken, Lieve; COPPERS, Sven; VAN DEN BERGH, Jan; LUYTEN, Kris & CONINX, Karin (2017) SCATE – Smart Computer-Aided Translation Environment – Year 3 (/4). In: Proceedings of the 20th annual conference of the European Association for Machine Translation EAMT 2017,p. 19-19.

Handle: http://hdl.handle.net/1942/24898

SCATE – Smart Computer-Aided Translation Environment – Year 3 (/4)

Vincent Vandeghinste
Tom Vanallemeersch
Liesbeth Augustinus
Frank Van Eynde
Joris Pelemans
Lyan Verwimp
Patrick Wambacq
Geert Heyman
Marie-Francine Moens
Iulianna van der Lek-Ciudin
Frieda Steurs
University of Leuven
first.lastname@kuleuven.be

Ayla Rigouts Terryn
Els Lefever
Arda Tezcan
Lieve Macken
Ghent University
first.lastname@ugent.be

Sven Coppers
Jan Van den Bergh
Kris Luyten
Karin Coninx
UHasselt – tUL – EDM
first.lastname@uhasselt.be

Abstract

We aim to improve translators' efficiency through improvements in the technology. Funded by Flemish Government IWT-SBO, project No. 130041. http://www.ccl.kuleuven.be/scate

1 Tree-based MT and TM

We have aligned parse trees based on semantic predicates and roles, and building a tree-to-tree decoder for syntax-based SMT. We create parallel node-aligned treebanks and make them available online. We investigate different fuzzy matching metrics and how to integrate them with MT.

2 Detecting grammatical errors in SMT

As grammatical errors are the most frequent error types in MT output, we develop a methodology that detects grammatical errors in SMT output by using monolingual morpho-syntactic word representations in combination with surface and syntactic context windows.

3 Term Extraction from Comparable Corpora

Framing the induction of translations as a classification problem, we learn from a seed dictionary what word pairs are translations. We combine word and character-level features and induce features on character-level from training data. For evaluation we developed an annotation scheme with detailed guidelines, resulting in high inter-annotator agreement. In addition to monolingual annotations, we are also working on a bilingual gold standard, where terms are linked with their translations.

4 Post-Editing via ASR

We are investigating domain adaptation by boosting language model probabilities of domain-specific terminology. The terminology is inferred from the already corrected material, either directly by keeping a word cache or indirectly, by using word and/or topic similarity. In addition, the language model is enriched with character-level information which enables modeling out-of-vocabulary words, which are very common in new domains.

5 Intelligible Translator Interfaces

A thorough redesign of translator interfaces has been established, integrating the different types of MT and TM, term corpora and consistency checks in such a way translators can minimize focus shifts and optimize usage of these tools. We included support for multiple translators working on different pieces of the same text and personalized workflows as part of the online translator interface.

6 Integration

We have built a demo system which combines the different research aspects into one demo, and are working with translators to collect feedback on the interface.