Smart Computer Aided Translation Environment – SCATE

IWT – Agentschap voor Innovatie door Wetenschap en Technologie
Strategic basic research
Project Nr. 130041
http://www.ccl.kuleuven.be/scate

University of Leuven (CCL - ESAT/PSI - LIIR – Fac. Arts Antwerp), Belgium
University of Ghent (LT3), Belgium
Hasselt University (tUL - iMinds, Expertise Centre for Digital Media), Belgium

Project duration: March 2014 – February 2018

Summary

We aim at improving the translators' efficiency through five different **scientific objectives**.
Concerning **improvements in translation technology**, we are investigating syntax-based fuzzy matching in which we estimate similarity based on syntactic edit distance or similar measures. We are working on syntax-based MT using synchronous tree substitution grammars induced from parallel node-aligned treebanks, and are building a decoder to use these grammars in translation.
Concerning **improvements in evaluation of computer-aided translation**, we have developed a taxonomy of typical MT errors and are constructing a manually annotated corpus of 3000 segments of Google Translate MT errors. Post-editing behaviour of translators is being monitored.
Concerning **improvements in automated terminology extraction from comparable corpora**, we have developed C-BiLDA, a multilingual topic model. It does not assume linked documents to have identical topic distributions. On the task of cross-lingual document categorization, we trained it on a comparable corpus of Wikipedia documents, and inferred cross-lingual document representations on a dataset for document categorization. The document representations and category labels are fed to an SVM classifier: we train on the source language and predict the labels for the target language documents. C-BiLDA outperforms the state-of-the-art in multilingual topic modeling.
Concerning **improvements in speech recognition accuracy,** we clustered words by their translations in multiple languages. If words share a translation in many languages, they are considered synonyms. By adding context and by filtering out those that do not belong to the same part of speech, we find meaningful word clusters to incorporate into a language model. We found no improvements, and attribute this in part to errors made by the MT system and to the incorporation technique (hard clustered class-based n-grams). We will take context into account during evaluation and/or further improve the word clusters by using the translations as features in vector space modeling techniques.
Concerning **improvements in work flows and personalised user interfaces**, we *reviewed existing translation systems*, and created an inventory of the various features and configuration options of the systems. Six Flemish companies are *interviewed* regarding their practices and their vision for future CAT tools. A *worldwide survey* has been conducted with more than 135 responses. Detailed analyses of translators' practices have been conducted by observing more than 7 translators by conducting a *contextual inquiry*.
In the upcoming period, the results of the different studies will be analysed in order to obtain a model of how CAT tools can support workflows for specific translators. This model will be used as a base for the personalised visualisations as part of interfaces for translation work. In contrast with traditional engineering approaches, this model will also be usable by translators as part of the configuration of their personal CAT tool.