Topic Adaptation for Lecture Translation through Bilingual Latent Semantic Models

Nicholas Ruiz

University of Groningen
Free University of Bozen-Bolzano

Abstract

Language models (LMs) are used in Statistical Machine Translation (SMT) to improve the fluency of translation output by assigning high probabilities to sequences of words observed in training data. However, SMT systems are trained with large amounts of data that may differ in style and genre from a text to be translated. Language models can be adapted through various techniques, including topic modeling approaches, which describe documents as a mixture of topics.

Several bilingual topic modeling approaches have been recently constructed to adapt language models to reward translated word sequences that include words that better fit the topics that represent the translation text. Most topic modeling approaches use Latent Dirichlet Allocation, which makes a prior assumption about the distribution of topics within a document.

This work presents a simplified approach to bilingual topic modeling for language model adaptation by combining text in the source and target language into very short documents and performing Probabilistic Latent Semantic Analysis (PLSA) during model training. During inference, documents containing only the source language can be used to infer a full topic-word distribution on all words in the target language's vocabulary, from which we perform Minimum Discrimination Information (MDI) adaptation on a background language model (LM). We apply our approach on the English-French IWSLT 2010 TED Talk exercise, and report a 15% reduction in perplexity and relative BLEU and NIST improvements of 3% and 2.4%, respectively over a baseline only using a 5-gram background LM over the entire translation task. Our topic modeling approach is simpler to construct than its counterparts.

Keywords:
Machine Translation, Language Modeling, Topic Adaptation, Topic Modeling