早稲田大学大学院情報生産システム研究科

博 士 論 文 概 要

論 文 題 目

A Study of Efficient Bidirectional Latent Variable Models in Machine Translation

申 請 者
Hao WANG
情報生産システム工学専攻
用例翻訳・言語処理 研究

2018 年 10 月

Machine translation between distant languages, like English-Japanese or Chinese-Japanese, is problematic, because of the existence of translation divergences (i.e., cross-linguistic differences). These differences may lie in the way words correspond across languages (*phrasal,* alignment), in the way words are ordered in sentences (*structural*, syntax) and in the way words are decomposed (*lexical, segmentation*). In machine translation, a number of approaches have attempted to account for divergences by introducing latent variables into the translation models. Such latent variables, i.e., hidden parameters, capture the underlying classes or structures. Previous research has shown that latent variables benefit the learning of the translation. However, most of these latent variable models are asymmetric (i.e., mono-directional or monolingual) models. They are only able to partially capture the translation divergences. This results in lower translation scores when translating distant language pairs. Apart from that, these models ignore the fact that cross-linguistic equivalents might exist in vocabulary, syntax or phrasal alignments. It is natural to exploit bidirectional models that might take advantage of these equivalents.

This dissertation proposes a systematic solution to the translation divergence problem, called bidirectional latent variable framework. This framework consists of several bidirectional latent variable models. Each model deals with the corresponding lexical, phrasal or structural divergences sequentially. Thus, this leads to more efficient learning of distant language translation. The efficiency of these bidirectional models are investigated through three machine translation tasks: phrasal alignment, syntax-based reordering and word segmentation. Experimental results shows that bidirectional latent variable models effectively reduce conflicts caused by asymmetric models in these tasks, which yield (a) state-of-the-art performance, while leading to (b) shorter training times and/or (c) smaller model sizes.

The dissertation is organized as follows:

Chapter 1 [**Introduction**] introduces the existing problems in current MT approaches, basic notions of latent variable and explains the motivations of this work.

Chapter 2 [**Background**] provides the necessary concepts in statistical machine translation (SMT) and neural machine translation (NMT).

Chapter 3 [**Exploiting Bidirectional Latent Variable Models in Phrase-based SMT: Phrasal Alignment**] investigates a fundamental latent variable in phrase-based SMT models, i.e., phrasal alignment. Conventional methods extract phrasal alignments (many-to-many) relying on asymmetric word alignment models, i.e., IBM models (1~5) [Brown, 1993]. This requires a

bi-directional training procedure and the additional process of symmetrization, resulting in long training times and large translation tables.

This dissertation proposes a novel hybrid method for symmetric phrasal alignment. It rests on the effective approximation method for IBM model 2 [Riley and Gildea, 2012] (more accurate, faster) and a beam-search variation of the hierarchical sub-sentential alignment (HSSA) method [Lardilleux et al., 2012] (symmetric). This enables the proposed method to obtain better estimation of initial parameters, to train faster, and to deliver more accurate phrasal alignments.

Compared to other techniques (GIZA++, the state-of-the-art implementation for IBM models), this novel method (a) keeps the translation accuracy in various language pairs. But it is much more efficient because (b) it requires only 4% of the training time of GIZA++ and (c) it outputs much smaller translation tables (50% in average). Compared to fast_align [Dyer et al., 2013], another fast and efficient word aligner, it (a) significantly surpasses it (+1.37 BLEU points, p-value < 0.01) in English-Japanese within (b) the same training time while (c) reducing by half the size of the translation tables.

Chapter 4 [**Exploiting Bidirectional Latent Variable Models in Syntax-based SMT: Syntactic Representation**] investigates the bidirectional latent variable of shared syntactic structures across languages. Phrase-based SMT suffers from the long-distance reordering problem, especially for distant language pairs, typical of English-Japanese. Recent works on syntax-based SMT have shown that the use of syntactic structures improves long-distance reorderings. Syntax-based SMT methods involves either making use of syntactic parsers to obtain parse trees, or employing large rule tables such as hierarchical SMT [Chiang, 2007].

The bracketing transduction grammar formalism (BTG) [Wu, 1997] constitutes a simple bilingual parsing model to answer this problem. BTG allows different word orders licensed by the same syntactic structure. Recent research has widely studied BTG-based preordering in phrase-based SMT, however there are less attention on BTG-based decoding. This dissertation proposes a novel latent BTG-based decoding method based on the TD-BTG method [Nakagawa, 2015]. Although TD-BTG is the state-of-the-art method, it is sensitive to initial word alignments. Aiming at more accurate reordering models, this dissertation first improves the training algorithm in the TD-BTG method by adopting ensemble/mini-batch learning techniques with a forest-margin-based parameter updating strategy. On the top of that, a bottom-up BTG-based decoder is built. It constitutes a log-linear model where incorporating the TD-BTG-based reordering model with other models.

As for efficiency in translation, when used for preordering, the improved method (a) leads to statistically significant gains in English–Japanese and Japanese–English translation accuracy (+0.5 and +0.8 BLEU point respectively, p-value < 0.01). (b) It is four times as fast in training compared to the state-of-the-art method (c) without difference in model sizes. The use of the BTG-based reordering model (a) leads to a statistically significant improvement (+0.43 BLEU point, p-value < 0.05) in English-Japanese translation accuracy for our BTG-based decoder over the standard Moses phrase-based decoder (b) with similar decoding speed while (c) reducing the size of the lexical reordering models (20% of state-of-the-art model size).

Chapter 5 [**Exploiting Bidirectional Latent Variable Models in Seq2seq NMT: Sub-word Segmentation**] investigates the bidirectional latent variable of vocabulary for word segmentation. Word segmentation is used for the computation of the vocabulary of East Asian languages without word separators (Chinese or Japanese). Conventional supervised word segmenters massively produce rare words which MT systems cannot translate in different quantity for different segmenters. This results in lower and different translation scores in SMT or NMT.

To tackle the rare word problem in East Asian languages and reduce sensitivity to segmentation, this dissertation proposes a novel bilingual unsupervised sub-word segmentation method based on the principle of minimum description length (MDL). This method learns a single (F) finite-size vocabulary, where sub-words are common to the two languages, each with (M) a minimal frequency. In addition, this enables to share word embedding layer between the encoder and the decoder in Seq2seq NMT.

When compared with state-of-the-art word segmenters, Juman for Japanese and Stanford Segmenter for Chinese, in both SMT and NMT experiments in Japanese-Chinese and Chinese-Japanese, the proposed method (a) leads to statistically significant improvements in translation accuracy (+1.0 BLEU, p-value < 0.01) (b) Times cannot be compared as Juman and Stanford Segmenter are frozen models, but (c) the vocabulary size is largely reduced (20 times smaller) as well as the size of the word embedding layer (50% smaller than in a monolingual setting). When comparing with two other sub-word models for NMT sentence piece model [Wu et al. 2016] and byte pair encoding [Sennrich et al. 2016], the proposed method (a) leads to comparable or above translation accuracy in both monolingual and bilingual cases while (b) being 3 to 8 times as fast, (c) for the same vocabulary sizes.

Chapter 6 [**Contributions and Conclusions**] summarizes the dissertation and mentions future directions. This dissertation investigated to reduce

translation divergences using bidirectional latent variable models. The proposed methods not only yield (a) state-of-the-art or above performance but also (b) have shorter training times and/or (c) smaller model sizes. This is a significant contribution to the field of machine translation which pushes the limits of MT. For all these reasons, this is worth a PhD degree.