Neural Machine Translation with Translation Selection Ambiguity

2310086 Kuon Takata

In recent years, the performance of Machine Translation (MT) has improved dramatically due to methods based on neural networks. However, regarding the translation of words with multiple meanings (polysemous words), challenges remain because polysemous words need to be translated into different words depending on the meaning. Word Sense Disambiguation (WSD) plays an important role in translating polysemous words. WSD is the task of identifying the meaning (sense) of a polysemous word appearing in a certain context, and it becomes possible to translate a polysemous word into an appropriate word in the target language by using this as a preprocessing step for translation. On the other hand, since current Neural Machine Translation (NMT) is a method that trains End-to-End models, incorporating WSD as preprocessing has been disregarded. To improve translation performance between Korean and Vietnamese, one of the previous studies identifies the senses of polysemous words using WSD, represents them with different tokens for each sense, and provides sentences with identified word senses as input. However, sense-level tokenization increases the vocabulary size handled by the translation model and increases training parameters, leading to the problem of requiring more training data.

The goal of this study is to explore a method to improve machine translation performance by applying WSD only to translation-oriented ambiguous words and assigning unique tokens for each sense. A translation-oriented ambiguous word is defined as a polysemous word where each sense is translated into a different word in the target language. For example, the English word "bank" has two senses, "financial institution" and "land along a river," which are translated into Japanese as "*ginko*" and "*dote*," respectively; thus, it is a translation-oriented ambiguous word. On the other hand, the word "wall" has two senses, "partition" and "social barrier," but both are translated as "*kabe*" in Japanese, so it is not a translation-oriented ambiguous word. By limiting this sense-level tokenization to translation-oriented ambiguous words, we aim to improve machine translation performance for polysemous words while suppressing the increase in vocabulary size. Note that the languages handled in this study are Japanese as the source language and English as the target language.

First, translation-oriented ambiguous words are automatically identified. A Japanese-English parallel corpus is prepared, and then the word alignment tool GIZA++ is applied for it to identify the correspondence between Japanese words and English words. Next, by aggregating word correspondences across the entire corpus, each of a Japanese word is associated with

a list of its multiple corresponding English words. Finally, post-processing such as removing words with low and extremely high frequency is performed. Through this procedure, a lexicon of translation-oriented ambiguous words is constructed. This lexicon consists of pairs of translation-oriented ambiguous words (Japanese words) and their corresponding multiple English words.

Next, WSD is performed on translation-oriented ambiguous words in the source language sentences of the parallel corpus. The original words are subdivided by sense IDs identified by WSD (e.g., word_senseID) and treated as individual tokens for each sense. After this processing is performed, the MT model is trained using the parallel corpus as training data. The Transformer is employed as the translation model. It is trained from scratch with randomly initialized parameters.

This study proposes multiple methods based on the definition of senses and the WSD method. The first method performs WSD using the Iwanami Japanese Dictionary as the definition of senses. This method is denoted as "JDIC." An existing WSD tool, KyWSD, is used for WSD. This method has the advantage that WSD can be applied to all translation-oriented ambiguous words, but it also has the disadvantage that dictionary senses are not necessarily associated with translated words; that is, different senses in the dictionary are not necessarily translated into different English words. The second method performs WSD using the list of translated words in the lexicon of translation-oriented ambiguous words as the definition of senses. This method is denoted as "TR." However, the WSD model used in TR needs to be trained independently. A dataset to train the WSD model is constructed by labeling Japanese words with their corresponding translations (senses) based on word alignment results from GIZA++. Three WSD models are proposed: (1) a model that uses the pre-trained language model BERT to obtain embeddings of target words and learns multiple Fully Connected Layers (FCLs) to classify the sense of each word using these embeddings as features (called TR-pre), (2) a model that fine-tunes BERT individually for each word (called TR-ft-indi), and (3) a model that learns a BERT model shared among all words and multiple FCLs to classify the sense of each word (called TR-ft-shared). For the translation experiments, we adopted the first and third methods in terms of computational cost and performance. As extensions of the third model, we also train a model that classifies the sense as "unknown" when it does not correspond to any translation in the lexicon of translation-oriented ambiguous words (called TR-ft-shared-unk), and a model that classifies the sense as "uncertain" when the WSD model cannot predict the sense with high confidence (called TR-ft-shared-uncer).

Several experiments are conducted to evaluate the effectiveness of the proposed methods. One million sentence pairs randomly extracted from the

Japanese-English parallel corpus JParaCrawl is used as training data, and another 4,000 sentences is used as test data. As evaluation metrics, in addition to BLEU, a standard metric for MT, a unique metric called "translation selection accuracy" is used. This metric approximately estimates the ratio at which translation-oriented ambiguous words were translated into correct words. The method that trains the translation model without WSD (called vanilla) achieved a BLEU of 27.9 and a translation selection accuracy of 0.523, while the method that performs WSD on all words corresponding to previous research (called JDIC-all) achieved a BLEU of 27.2 and a translation selection accuracy of 0.523. In contrast, among the proposed methods that perform WSD only on translation-oriented ambiguous words, TR-ft-shared achieved the best performance, with a BLEU of 28.1 and a translation selection accuracy of 0.534. JDIC, which uses a dictionary as the definition of senses, did not outperform the baselines (vanilla and JDIC-all). This may be because the dictionary senses are not directly associated with the ambiguity of translation selection. Other proposed methods improved BLEU or translation selection accuracy compared to the baselines, but there were few cases where both were improved, thus their effectiveness was limited compared to TR-ft-shared. Furthermore, when measuring the translation performance of the proposed method under the ideal condition where tokens were subdivided by correct senses, that is, under the situation that the WSD accuracy was 100%, the BLEU became 28.9 and the translation selection accuracy became 0.689. Compared to the vanilla and JDIC-all baselines, the translation selection accuracy improved significantly in particular. This supports the validity of the approach of the proposed method, which performs WSD limited to translation-oriented ambiguous words, subdivides tokens for each sense, and then trains the translation model.