

論文の内容の要旨

Abstract

論文題目 Translation and Description Methods for Multilingual Text Understanding
 (多言語テキスト理解のための翻訳および語義説明手法)

氏 名 石渡 祥之佑

The development of Web technologies has been rapidly accelerating human communication and sharing of knowledge. The massive amount of text on the new communication platforms or knowledge sources often consists of documents in several domains and multiple languages. In order to obtain fresh and diverse information from multilingual and diverse text source such as Twitter, Wikipedia, or arXiv, we need to cope with language barrier while also paying attention to domain differences.

Let us move on to the general topic of natural language processing. Machine translation, as one of the most important applications of natural language processing, has been playing an important role in overcoming the language barrier. The development of machine learning techniques and huge annotated corpora have kept improving the performance of machine translation, and make it more and more widely used. In the face of the increasing use of multilingual platforms and knowledge sources, can machine translation help us understand the real data in various domains? Can machine translation be applied to languages pairs whose vocabulary and grammar are significantly different (such as English vs. Japanese)? More generally, can machine directly help humans understand text written in unfamiliar domains/languages? These are the central topics of this thesis.

To answer the above questions, we propose **an instant domain adaptation method**, **an accurate translation method for English-to-Japanese translation**, and **a description generator for unknown phrases**.

Instant domain adaptation for Statistical Machine Translation:

To translate text in various domains, the most basic method is domain adaptation.

Most studies on domain adaptation require supervised in-domain resources such as parallel corpora or in-domain dictionaries. The necessity of supervised data has made such methods difficult to adapt to practical machine translation systems. In this thesis, we thus propose a method that adapts translation models without in-domain parallel corpora. Our method improves out-of-domain translation from Japanese to English by 0.5-1.5 BLEU score.

Accurate translation method for English-to-Japanese translation:

English-to-Japanese translation is more difficult than other language pairs such as English-to-German or English-to-French translations. This is mainly because (1) Japanese sentence has much more words in a sentence compared to English, and (2) Japanese is a free-word-order language. To cope with these problems, we propose a chunk-based decoder for neural machine translation. Our method improves English-to-Japanese translation by 0.93 BLEU score and achieved a state-of-the-art performance on WAT'16 translation task.

Description generator for unknown phrases:

Even if a text is translated perfectly, or written in our familiar languages, it is still common for humans to become stuck on unfamiliar words and phrases. To help humans understand unknown phrases which are not included in hand-crafted dictionaries, we undertake a task of describing a given phrase in natural language based on its contexts. In contrast to the existing methods, our model appropriately takes important clues from contexts and achieves state-of-the-art performance in four description generation datasets.

To Help humans understand real multilingual text is a challenging task because (1) the target domain is unknown, (2) the source language may extremely differ from the user's language, and (3) the users may be unfamiliar with the words/phrases in the text. Our proposed methods tackle these problems by (1) instant domain adaptation, (2) accurate English-to-Japanese translation, and automatic description generation. We expect that this thesis will provide a promising future direction for research into multilingual text understanding.