NAIST® Research Highlights

Nara Institute of Science and Technology | **Computational Linguistics Laboratory** www.naist.jp/en

Machine translation

Better bilingual dictionaries

By combining two computational methods, NAIST researchers have enhanced the accuracy of machine-readable bilingual dictionaries

Machine-readable bilingual dictionaries aid computer translation, but existing frameworks are plagued with inaccuracies. To remedy this situation, researchers at Japan's NAIST have devised a method for generating more accurate bilingual dictionaries for machine translation; this method combines two existing computational techniques — topic modelling and word alignment[1].

"Bilingual dictionaries that focus on specific domains, for example medicine or tourism, are useful for many applications," explains Kevin Duh, who worked with NAIST colleagues Yuji Matsumoto and Xiaodong Liu on the project. "Existing topic models automatically group words into specific domains, so we included this as an integral component of our method. We are the first to use topic models for this purpose."

A topic model is a mathematical framework that predicts the proportion of words related to different topics in a particular text. It then generates groups of words that are likely to correspond to individual topic areas. The team applied topic models to English and Japanese 'real-world' documents (or 'corpora') written on the same subject, taken from the online encyclopaedia Wikipedia. This gave them lists of English and Japanese words that correspond to individual topics. Although these lists have the advantage of being easier to deal with than entire original texts, topic modelling did not help the researchers correctly identify which English words translated into which Japanese words in each list.

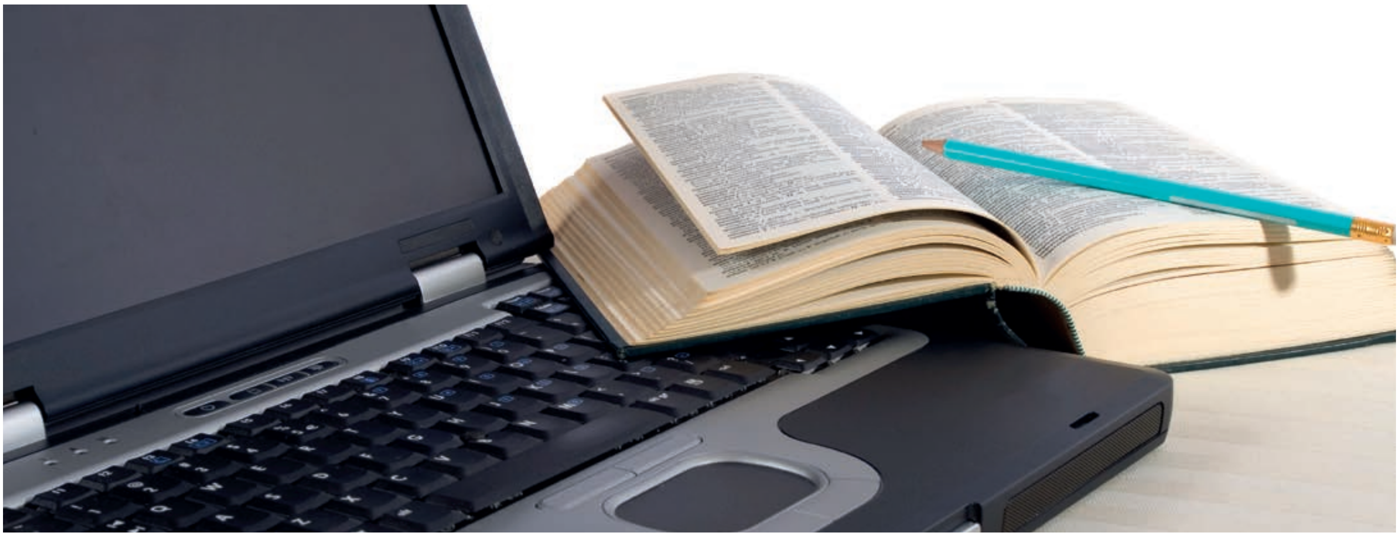
"Let's say each topic list contains about 500 English words and 500 Japanese words," says Duh. "There are then 500 × 500 translation possibilities in each list. Our insight came when we realized that this problem is similar to the word alignment problem in the field of statistical machine translation."

Word alignment is a computational technique used for linking words that are close translations of one another within texts. Matsumoto's team carried out word alignment after running topic models, enhancing the chances of accurate bilingual translation. In fact, they discovered that bilingual dictionary extraction became more accurate when they incorporated more subject-specific documents from multiple languages. Additional language data meant that each topic model became more and more specific, allowing word alignment to 'zone in' on more precise translations relevant to the original texts.

"Including multiple word phrases, such as compound nouns, is an important next step because they are so common in languages," explains Duh. "This is a surprisingly difficult task, because it dramatically increases both the computation time and the number of translation candidates. However, it will enhance the usefulness of our method."

Reference

1. Liu, X., Duh, K. & Matsumoto, Y. Multilingual topic models for bilingual dictionary extraction. *ACM Transactions on Asian Language Information Processing* **9**, 39 2014.



© marcelmooij/iStock/Thinkstock

Multilingual Wikipedia entries on the same subject were used to create more accurate bilingual dictionaries by combining existing topic model and word alignment techniques.

More information about the group's research can be found at Yuji Matsumoto's webpage: http://www.naist.jp/en/about_naist/offices/administration_bureau/yuji_matsumoto/index.html