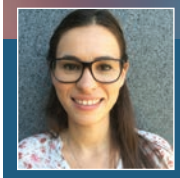
Crossing Sentence Boundaries in Statistical Machine Translation

Mascarell, Laura ; Rios, Annette ; Volk, Martin

Crossing sentence boundaries in statistical machine translation



Laura Mascarell is a PhD student in computational linguistics at the University of Zurich. Her PhD thesis studies textual dependencies across sentences, and their integration within MT systems.

Annette Rios is completing her postdoctoral at the Institute of Computational Linguistics at the University of Zurich, where she graduated in 2015.

Martin Volk is head of the Institute of Computational Linguistics at the University of Zurich. His research focuses on multilingual systems, in particular on MT.

Standard phrase-based statistical machine translation (SMT) systems translate one sentence at a time, completely ignoring discourse dependencies and the wider context of the document. As a consequence, words with multiple senses are often mistranslated when they are ambiguous in the local context. These translation errors decrease the quality of the translation, threatening the cohesion of the text. Research in discourse-aware SMT tackles document-level issues to improve the translation and to ensure that discourse features such as cohesion are maintained in the translation.

As an example, the English word *face* is most frequently translated into German as *Gesicht*. But when we deal with mountaineering, the word *face* may also refer to a specific side or part of a mountain — for example, "the north face of Mount Everest" or "the face has several cracks" — and must be translated into German as *Wand*. Therefore, the machine translation (MT) system needs to consider the context in order to determine the correct translation variant.

SMT has been dominated by phrase-based models for the last decade, and several freely available toolkits, such as Moses or Joshua, provide a fast way to obtain state-of-the-art translation systems. However, this approach comes with well-known limitations regarding the performance of these systems: phrase-based models need to make strong independence assumptions, since they translate each sentence independently and only consider local context during translation. This makes it hard to model dependencies across sentences, which can result in a loss of crucial information and sometimes a wrong translation. Research in discourse-aware MT is generally focused on specific problems related to document-level dependencies, such as lexical and grammatical cohesion.

For instance, the multiuse German word *Absatz* can be translated into English as *heel*, *paragraph* and *sale*. In

a phrase such as *hoher Absatz*, both a translation as *heel* or *sale* can be appropriate, depending on the context. Information about how this word has been translated in previous sentences, or about the general domain of the document, will help the MT system make the correct choice. Generally, words that have multiple possible translations are a challenge, especially if the correct translation in a specific context does not reflect the most frequent meaning of the word. A special case of this relation can be observed by words that are introduced as part of a nominal compound and maintain that meaning throughout the text, even if they appear on their own. Consider the following translation example of the German word *Typ*:

German: "Der ektomorphe Körpertyp neigt zur Schlankheit. Dieser Typ muss viel Krafttraining machen."

MT: "The ectomorph body type tends to be slender. This guy has to do a lot of strength training."

We observe that the translation of the second clause is grammatically correct, but it does not convey the meaning of the German sentence where *Typ* refers back to *Körpertyp* (body type) in the previous sentence, and therefore, the correct translation would be type, not guy. Discourse-aware SMT systems cross sentence boundaries, and the information that *Typ* in the second clause corefers back to the compound *Körpertyp*, helps to disambiguate *Typ* and translate it correctly. We assume that the head of a nominal compound — *typ* in *Körpertyp* — should have the same translation as a coreference and as part of the compound. Since the coreferring head noun in isolation may not produce a desired translation, we take advantage of the compound. Note that compounds are the result of joining multiple words, providing fewer translation variants than words consisting of a single root, and thus helping to reduce ambiguities when translating their parts. *Körpertyp* will be translated into *body type*, but not *body guy*.

In our experiments, we use the sentence-level translation system Moses. In order to enforce a correct translation across sentence boundaries, we employ two different methods: plugging in the correct translation to the system before translation or before post-editing. With the first method, we translate the document one sentence at a time. However, we cache, or store, the translation of the head of the compounds — for instance, type in body type — and encourage the translation system to use the corresponding cached translation for every coreference to a compound. To do so, we use the XML markup scheme integrated in Moses, which allows us to introduce the preferred translation, competing with the other translation candidates without changing the model. This approach improves the translation correctness of these coreferences from 80.1% to 86.7% when translating from German into French.

The post-editing approach is similar: we perform the caching step, but instead of plugging a specific translation into the translation system, we automatically edit the MT output, replacing all coreferences with their cached translation. There are advantages and disadvantages with both the plugging and the post-editing method. During the translation process, several components, or models, are combined to provide the best translation. Each of these models has a different function related to translation, reordering of the words, and fluency of the output. Post-editing is a straightforward approach to get the desired translation, but it is not included upfront in the translation process. As a consequence, the other models integrated in the translation system cannot contribute to verify whether the new translation is affecting the word order or the fluency of the output. With the use of the XML markup scheme provided by Moses in the plugging approach, the translation output can benefit from the other models. However, this is not optimal since the cached translation competes with the other translation candidates without proper probability scores.

An optimal solution that takes advantage of other models can be implemented using the document-level translation system Docent developed at the University of Uppsala by Christian Hardmeier, which offers more flexibility at modeling discourse dependencies. At every step of the translation process, Docent produces a complete translation of the entire document, and it accepts a new document translation when the combination of all model scores is higher than the score of the previous translation. To integrate our solution into Docent, we implemented a new model that gives higher scores when the translation of the head of a given compound and its coreferences are the same.

In a more general approach, we seek to improve the consistency in the translation of all ambiguous words, not just parts of compounds. Generally, if a word in a given source language has different translation variants in the target language, we can often infer the intended meaning by specific words in the context that we call trigger words. These words can, but need not, be in the same sentence, and be extracted from both the source and the target side. For example, on the one hand, the German words in the source: *ektomorph*, *Körpertyp*, *Muskelmasse*; and the English words in the target *ectomorph*, *metabolism*, *bodybuilding* trigger the translation of English *type* for German *Typ*. On the other hand, German *jung, friendly* or *band* trigger the translation guy. Thus, we define trigger words as words in the surrounding sentences that trigger a specific translation for a given ambiguous word.

In order to disambiguate a word using such trigger words, we need to find them first. For this purpose, we look at changes in the translation distribution when a specific trigger word candidate appears in the context of a given ambiguous word. For instance, when *ektomorph* appears in the context of *Typ*, the translation as *type* has the highest probability in the translation distribution, whereas if *ektomorph* is not in the context, *guy* has the highest translation probability. Using this method, we automatically extract the trigger words of all ambiguous words from a large parallel corpus. During translation, we check whether the detected trigger words appear in the surrounding sentences, and can thus conclude which translation is the most likely. We then plug the correct translation of the term into the translation system or do a post-editing step in exactly the same way as described above to obtain the desired translation.

Another area of research in discourse for MT is pronouns, most prominently personal pronouns like English *he, she, it* and German *er, sie, es* — these words are especially hard to translate, since the form of a pronoun is in many languages determined by gender and number of its antecedent — the noun it stands for. Therefore, in order to pick the correct translation, the system must know the gender and number of the word that a given pronoun refers to. For instance, a standard phrase-based MT system has problems with the translation of the following snippet from an article about South Korea's president Park Geun-hye from Spanish to English:

Spanish: "Para muchos surcoreanos, la elección de Park como candidata es segura. Si gana, será la consecuencia de su seriedad y tenacidad, no de su herencia política."

Human translation: "To many South Koreans, the election is now Park's to lose. If she wins, it will be the result of her seriousness and tenacity, not her political heritage."

MT without coreference resolution: "For many South Koreans, Park's election as a candidate is safe. If it does, it will be the result of her seriousness and tenacity, not his political legacy."

MT with coreference resolution: "For many South Koreans, Park's election as a candidate is safe. If she wins, it will be the result of her seriousness and tenacity, not her political legacy."

First of all, Spanish only uses subject pronouns for emphasis, but otherwise omits them, as is the case in the second sentence *Si gana* — if [she] wins. The MT system is smart enough to insert a pronoun, but since it does not have any information about the actual subject, it inserts the more frequent pronoun *it* instead of *she*. Furthermore, the Spanish possessive pronoun *su* is unspecified for gender, whereas English uses distinctive forms depending on the gender and number of the possessor. Since the MT system is ignorant of the actual possessor, it mistranslates *su* as *his* instead of *her*.

In our research, we use a coreference resolution system on the Spanish source text and annotate possessive pronouns with their respective morphological features. Additionally, we insert placeholders for the omitted subject pronouns that indicate gender and number of the subject. This can be done prior to the translation with English as the target language, since the gender distinction is only relevant for humans, and it is safe to assume that their grammatical gender does not change when translated. For languages that use grammatical gender for all nouns, the issue is more complicated, since for a correct translation we need to know not only the antecedent of a given pronoun, but also the gender of the translation of that antecedent, which can be different from the gender which the noun had in the source language.

As we have shown, translation systems benefit from discourse dependencies to improve translation choices. However, MT has been focused mostly on sentence-level translation for the last decade, and we should move on to MT systems that take into account discourse knowledge to outperform the quality of the translation. [M]