Discovering Paraphrases from Movie Subtitles

Suggested as a **presentation** at the HELDIG Summit 2017

Mathias Creutz
Department of Modern Languages
Faculty of Arts, University of Helsinki, Finland
mathias.creutz@helsinki.fi

1. Abstract

Paraphrases are pairs of phrases in the same language that essentially "mean the same thing", such as *"Have a seat"* versus *"Sit down"* or *"It's what we do"* versus *"This is our job"*. The natural language processing community is interested in paraphrases for numerous reasons. Applications include, for instance, plagiarism detection, automatic text summarization, automatic evaluation of machine translation systems, as well as computer assisted language learning.

We are interested in computer assisted language learning, in particular. One goal is to lower the threshold for learners to use a new language actively, by providing tools that help people express themselves better, in more idiomatic and natural ways. Appropriate data is needed to train such a system, and there do exist publicly available paraphrase corpora in different languages (Dolan and Brockett, 2005; Ganitkevitch et al., 2013; Ganitkevitch and Callison-Burch, 2014; Pavlick et al., 2015). Unfortunately, the existing data collections mostly consist of fairly formal language, such as news text or transcripts of parliamentary proceedings.

A source of more colloquial, every-day language can be found in OpenSubtitles2016[1] (Lison and Tiedemann, 2016), which is a collection of translated movie and TV subtitles from www.opensubtitles.org. The collection contains subtitles in 65 languages. When subtitles exist for the same film in multiple languages, then the subtitles have been aligned sentence by sentence, such that translations of the same sentences are side by side. Such a sentence aligned corpus is typically utilized to train a machine translation system. We have used it to discover paraphrases, that is, "translations" of sentences within the same language.

Bannard and Callison-Burch (2005) introduced a "pivot" technique for finding paraphrases from parallel texts. A sentence in the target language is translated to another, so-called pivot language and then translated back. Assume, for example, that English *"Have a seat"* is aligned with Finnish *"Istu alas"* somewhere in the corpus. Additionally, *"Istu alas"* is aligned with *"Sit down"* somewhere else in the corpus. Thus, we can conclude that *"Have a seat"* and *"Sit down"* are paraphrases. However, because of noise in the sentence alignments, not all candidates obtained in this manner are valid paraphrases. Therefore, probabilities are involved to produce a ranking of the suggested paraphrase candidates.

We have implemented the pivot technique and experimented with different extensions to the original formula proposed by Bannard and Callison-Burch (2005), which was based on conditional probability. As target languages, six European languages from four different language families were used: German, English, Finnish, French, Russian, and Swedish. In turn, each language was the target language and the five other languages served as pivot languages. Millions of paraphrase candidates were discovered.

[1] Available as part of OPUS ("... the open parallel corpus"): opus.lingfil.uu.se

Analyzing the paraphrases suggested can be quite interesting. Depending on the ranking scheme, different types of paraphrases appear at the top of the list. One scheme favors frequently used phrases, such as *"Yes."* ↔ *"Yeah."*, *"Of course."* ↔ *"Sure."*, *"Hello."* ↔ *"Good morning."*, *"Are you okay?"* ↔ *"Are you all right?"*. Another scheme favors sentences with more information content, such as *"It was a last minute thing."* ↔ *"This wasn't planned."*, *"I have goose flesh."* ↔ *"The hair's standing up on my arms."*, *"It was a difficult and long delivery."* ↔ *"The delivery was difficult and long."*.

The less good suggestions are also worth studying, and in many cases it is hard to decide what is correct. Sometimes the two sentences could, in fact, refer to the same situation, although one would not consider them to mean the "same thing": *"All right, everybody out."* ↔ *"Everybody on the floor."*, *"Isn't it nice?"* ↔ *"That shit's good for you."*, *"It should have been open."* ↔ *"That was my bad."*, *"Just dinner."* ↔ *"Nothing else."*

In the near future, we intend to release the complete paraphrase lists for public use. We also intend to carry out manual annotations for part of the data, in order to have a subset of fully reliable paraphrases in addition to the automatically discovered ones.

2. Bibliographical References

Bannard, C. and Callison-Burch, C. (2005). Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 597–604, Ann Arbor, Michigan. Association for Computational Linguistics.

Dolan, B. and Brockett, C. (2005). Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*. Asia Federation of Natural Language Processing, January.

Ganitkevitch, J. and Callison-Burch, C. (2014). The multilingual paraphrase database. In *The 9th edition of the Language Resources and Evaluation Conference*, Reykjavik, Iceland, May. European Language Resources Association.

Ganitkevitch, J., Van Durme, B., and Callison-Burch, C. (2013). PPDB: The paraphrase database. In *Proceedings of NAACL-HLT*, pages 758–764, Atlanta, Georgia, June. Association for Computational Linguistics.

Lison, P. and Tiedemann, J. (2016). OpenSubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*.

Pavlick, E., Rastogi, P., Ganitkevitch, J., Van Durme, B., and Callison-Burch, C. (2015). PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers)*, pages 425–430, Beijing, China, July. Association for Computational Linguistics.