Language Models for Machine Translation: Original vs. Translated Texts

Gennadi Lembersky Noam Ordan Shuly Wintner

Abstract

We investigate the differences between language models compiled from original target-language texts and those compiled from texts translated to the target language. Corroborating established observations of Translation Studies, we demonstrate that the latter are significantly better predictors of translated sentences than the former, and hence fit the reference set better. Furthermore, translated texts yield better language models for statistical machine translation than original texts.

Statistical machine translation (MT) uses large target language models (LMs) to improve the fluency of generated texts, and it is commonly assumed that for the construction of language models, “more data is better data” (Brants and Xu, 2009). Not all data, however, are created the same. In this work we distinguish between LMs compiled from texts originally written in the target language and LMs compiled from *translated* texts.

The motivation for our work stems from much research in Translation Studies that establishes the fact that original texts are significantly different from translated ones in various aspects (Gellerstam, 1986). Recently, corpus-based computational analysis corroborated this observation, and Kurokawa et al. (2009) apply it to statistical machine translation, showing that it is better to train a *translation* model for an English-to-French MT system on an English-translated-to-French parallel corpus and vice-versa. Our research question is whether a *language model* compiled from translated texts may improve the results of the translation further.

We test this hypothesis on three different translation tasks: Hebrew-to-English, German-to-English and French-to-English. First, for each language pair we build two English language models from two types of corpora: texts originally written in English, and human translations from the source language into English. We show that for each language pair, the latter language model better fits a set of reference translations in terms of perplexity. In other words, LMs can successfully distinguish between original and translated texts. Moreover, we demonstrate that the differences between the two LMs are not biased by content but rather reflect differences on abstract linguistic features.

Research in Translation Studies indicates that certain *translation universals* exist which cause translated texts from several languages to a single target language to resemble each other along various axes (Baker, 1993, 1995, 1996). To test this hypothesis, we compile additional English LMs, this time using texts translated to English from languages *other* than the source. Again, we use perplexity to assess the fit of these LMs to reference sets of source-language-translated-to-English sentences. We show that these LMs depend on the source language and differ from each other. They outperform the original-based LMs, but the LMs compiled from texts that were translated from the *source* language still fit the reference set best.

Finally, we train a phrase-based MT system (Koehn et al., 2003) for each language pair. We use parallel corpora comprising components translated from source and target languages. We use four LMs: one original, one translated from the source language, one translated from another language and one (baseline) oblivious to the source language. We show that the translated-from-source-language LMs provide a siginificant improvement in the quality of the translation output over all other LMs.

The main findings of this work, therefore, are that original and translated texts indeed exhibit significant, measurable differences, and LMs compiled from translated texts better fit translated references than LMs compiled from original texts (and, to a lesser extent, LMs compiled from texts translated from languages other than the source language). These differences yield significant improvement in the quality of MT systems that use LMs compiled from translated texts.

References

Mona Baker. Corpus linguistics and translation studies: Implications and applications. In Gill Francis Mona Baker and Elena Tognini-Bonelli, editors, *Text and technology: in honour of John Sinclair*, pages 233–252. John Benjamins, Amsterdam, 1993.

Mona Baker. Corpora in translation studies: An overview and some suggestions for future research. *Target*, 7(2):223–243, September 1995.

Mona Baker. Corpus-based translation studies: The challenges that lie ahead. In Gill Francis Mona Baker and Elena Tognini-Bonelli, editors, *Terminology, LSP and Translation. Studies in language engineering in honour of Juan C. Sager*, pages 175–186. John Benjamins, Amsterdam, 1996.

Thorsten Brants and Peng Xu. Distributed language models. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Tutorial Abstracts*, pages 3–4, Boulder, Colorado, May 2009. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/N/N09/N09-4002`.

Martin Gellerstam. Translationese in Swedish novels translated from English. In Lars Wollin and Hans Lindquist, editors, *Translation Studies in Scandinavia*, pages 88–95. CWK Gleerup, Lund, 1986.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54. Association for Computational Linguistics, 2003.

David Kurokawa, Cyril Goutte, and Pierre Isabelle. Automatic detection of translated text and its impact on machine translation. In *Proceedings of MT-Summit XII*, 2009.