Statistical Machine Translation has received significant attention from the academic community over the past decade. This research has led to significant improvements in machine translation quality. As a result, it is widely adopted in the industry (Google, Microsoft, Twitter, Facebook, ...etc.) as well as the government (http://nist.gov). The biggest factor in this improvement has been the availability of ever increasing sources of training data as digital multilingual communication and information dissemination become ubiquitous. Relatively little research has been done on training data analysis and selection, despite training data being the main contributor of machine translation quality.

In this work, we first examine fundamental properties of translated and authored text. We introduce a new linguistically motivated feature (Part of Speech Tag Minimal Translation Units) that outperforms prior work in sentence level translation direction detection. Next, we develop a cross-domain data matrix that enables comparison between different features in the translation direction detection task. We extend our previously introduced feature for translation direction detection to use statistically trained brown clusters instead of part of speech tags. This new feature outperforms all prior work in all cross-domain data matrix combinations.

Data selection in machine translation is performed in different scenarios with different objectives including: reducing training resource consumption, domain adaptation, improving quality or reducing deployment size. We develop an efficient (computational complexity and memory consumption is linear in training data size) framework for training data selection and compression called Vocabulary Saturation
Filter (VSF). In our experiments we show the machine translation system trained on data selected using VSF is comparable to prior data selection methods with quadratic computational complexity. However, VSF is sensitive to data order. Therefore, we experiment with different orderings of the data and compare the results.

Finally, inspired by the features introduced earlier for detecting translation direction, we propose a generalization of VSF to use various features leveraging joint models of source and target using Minimal Translation Units (MTU) in addition to adding source side context using brown clusters. This enables leveraging word alignment and semantic information in the data selection process. Furthermore, we calculate the empirical probability distribution function for these data sets before and after data selection and compare different data selection methods based on the KL-divergence between the two probability distribution functions. In an attempt to preserve the original data distribution while reducing the data size, we also tuning the VSF threshold parameter for each n-gram according to its relative frequency and demonstrate its effectiveness in preserving original data distribution after data selection. Using these techniques, we are able to achieve higher translation accuracy (measured by BLEU) with less training data compared to prior work.

Journal Publications:

Conference Proceedings: