Latent Domain Models for Statistical Machine Translation .
C. Hoàng

A data-driven approach to model translation suffers from the data mismatch problem and demands domain adaptation techniques. Given parallel training data originating from a specific domain, training an MT system on the data would result in a rather suboptimal translation for other domains. But does suboptimality of translation happen only in such an extreme scenario of domain mismatch? This dissertation shows that training SMT systems on heterogeneous corpora (e.g. EuroParl, Common Crawl Corpus) may also result in suboptimal performance of statistical translation systems. Specifically, it is clear that a word/phrase could be translated in different ways when it comes to different domains. The translation statistics induced from word alignment models and phrase-based models, however, reflect translation preferences aggregated over diverse domains in heterogeneous corpora. In this sense, they can be considered as coarse and domain-confused statistics.

The first contribution of this dissertation is showing that domain-confused statistics may harm performance of both word alignment and phrase-based models. Another important contribution of this dissertation is to provide a principled way to address the problem. We focus on learning the translation statistics with respect to each of diverse domains (i.e. domain-focused translation statistics). With our method of domain induction for translation, we present a comprehensive study of domain adaptation for statistical machine translation, including four specific case studies Data Selection, Phrase-Based Translation, Word Alignment and Rewarding Domain Invariance in translation.

Finally, we briefly describe Scorpio, the ILLC-UvA Adaptation System submitted to an adaptation task at WMT 2016, which participated with the language pair of English-Dutch. This system consolidates the ideas in the thesis on latent variable models for adaptation. Results validate the effective adaptation performance in a competitive setting.