What's in a Domain? Towards Fine-Grained Adaptation for Machine Translation.

M.E. van der Wees

Summary

Machine translation (MT) uses software to translate texts written in one language (for example German) to another language (for example English). Modern-day MT systems are built using large amounts of example translations between the two languages of interest, so-called *parallel corpora*. If parallel corpora are sufficiently large and of high quality, created by professional translators who adhere to language standardizations, an MT system can produce good translations.

However, sizable high-quality parallel corpora only exist for a limited number of translation tasks, or *domains*, such as parliamentary proceedings. The picture looks less bright when training data is scarce for a domain of interest, for example if one wishes to translate medical texts. In such cases, the available training data differs from the translation task in both writing style and vocabulary, a mismatch that can cause large drops in translation quality. In recent years, this problem has been addressed by *domain adaptation*, in which an MT system is adapted to the domain of interest and translation quality is often improved.

Unfortunately, the notion of a domain is not uniformly defined. Typically, domain means ‘different data set,’ and is thus a hard-labeled concept that is often directly used to optimize an MT system. This definition ignores three important facts: first, documents or sentences *within* a single domain may vary at many levels, such as *topics*, *genres*, and *register*, which may be useful information to adapt an MT system. Second, some domains or genres may require different strategies than others to improve translation quality due to their inherent differences. Third, available domain labels may not provide the most useful information for effective adaptation.

To shed light on the concept of a domain and its impact on MT, the core question in this thesis is: *“What’s in a domain?”* Guided by this question, we distinguish various aspects that together make up a domain, i.e., topic, genre, register, dialogue acts, speakers, and speaker gender. We study to what extent MT output differs among these aspects, and how we can use them to perform fine-grained adaptation for MT. We are particularly interested in *informal* and *conversational* genres, which lack standardization and are notorious for poor MT output. In addition, we aim to develop methods that do not, or at most partially, rely on manual domain or genre information.

By studying what’s in a domain and showing how we can use different aspects of language to improve MT, we take in this thesis a step forward towards fine-grained adaptation for machine translation.