Title: Towards Machine Translation Based on Monolingual Texts

Author: Ivana Kvapilíková

Institute: Institute of Formal and Applied Linguistics

Supervisor: doc. RNDr. Ondřej Bojar, Ph.D., Institute of Formal and Applied Linguistics

Abstract: The current state of the art in machine translation (MT) heavily relies on parallel data, i.e. texts that have been previously translated by humans. This type of resource is expensive and only available for several language pairs in limited domains. A new line of research has emerged to design models capable of learning to translate from monolingual texts which are significantly easier to obtain, e.g. by web-crawling. While it is impressive that such models achieve translation capabilities, the translation quality of the output they produce is still low for practical applications. This dissertation thesis strives to improve their performance. We explore the existing approaches of using monolingual resources to train translation models and propose a new technique to generate pseudo-parallel training data artificially without expensive human input. We automatically select similar sentences from monolingual corpora in different languages and we show that using them in the initial stages of MT training leads to a significant enhancement in translation quality. We also point out the limitations of existing MT models based on monolingual texts which often struggle with the translation of named entities and generally produce low-quality translations, especially in truly low-resource conditions where monolingual training data is limited and often suffers from a domain mismatch.

Keywords: machine translation, unsupervised learning, deep neural networks, low-resource languages, natural language processing