Title: Mining Parallel Corpora from the Web
Author: Bc. Jakub Kúdela
Author's e-mail address: jakub.kudela@gmail.com
Department: Department of Software Engineering
Thesis supervisor: Doc. RNDr. Irena Holubová, Ph.D.
Supervisor's e-mail address: holubova@ksi.mff.cuni.cz
Thesis consultant: RNDr. Ondřej Bojar, Ph.D.
Consultant's e-mail adress: bojar@ufal.mff.cuni.cz

Abstract: Statistical machine translation (SMT) is one of the most popular approaches to machine translation today. It uses statistical models whose parameters are derived from the analysis of a parallel corpus required for the training. The existence of a parallel corpus is the most important prerequisite for building an effective SMT system. Various properties of the corpus, such as its volume and quality, highly affect the results of the translation. The web can be considered as an ever-growing source of considerable amounts of parallel data to be mined and included in the training process, thus increasing the effectiveness of SMT systems. The first part of this thesis summarizes some of the popular methods for acquiring parallel corpora from the web. Most of these methods search for pairs of parallel web pages by looking for the similarity of their structures. However, we believe there still exists a non-negligible amount of parallel data spread across the web pages not sharing similar structure. In the next part, we propose a different approach to identifying parallel content on the web, not dependent on the page structure comparison at all. We begin by introducing a generic method for bilingual document alignment. The key step of our method is based on the combination of two recently popular ideas, namely the bilingual extension of the word2vec model and the locality-sensitive hashing. With the method applied to the task of mining parallel corpora from the web, we are able to effectively identify pairs of parallel segments (i.e. paragraphs) located anywhere on the pages of a web domain, regardless of their structure. The final part of our work describes the experiments conducted with our method. One experiment uses pre-aligned data, and its results are evaluated automatically. The other experiment involves real-world data provided by the Common Crawl Foundation and presents a solution to the task of mining parallel corpora from a hundreds of terabytes large set of web-crawled data. Both experiments show satisfactory results, implying that our proposed method is a promising baseline for acquiring parallel corpora from the web. We believe the amount of parallel data obtainable with our method might enable SMT systems to get trained better and eventually achieve superior translation results.

Keywords: mining parallel corpora, bilingual document alignment, word2vec, locality-sensitive hashing