Mining Parallel Corpora from the Web

Charles University, Faculty of Mathematics and Physics

Aurhor: Mgr. Jakub Kúdela
Supervisor: Doc. RNDr. Irena Holubová, Ph.D.
Consultant: RNDr. Ondřej Bojar, Ph.D.

Overview

- We have proposed a new method for acquiring parallel corpora from the web.
- Unlike majority of others, our method does **not depend on structure comparison**.
- Using Hadoop, we have extracted Czech–English parallel corpus from the **149 TB** large dataset of web-crawled data.
- Our experiments show satisfactory results, implying that the method could become a **new promising baseline**.

Motivation

- Statistical machine translation (SMT) is one of the most popular approaches to machine translation today.
- Existence of a parallel corpus is the most important prerequisite for building an effective SMT system.
- The web can be considered as an ever-growing source of considerable amounts of parallel data.

Proposed Method

- Our generic method solves the task of **bilingual document alignment**.
- The method is supervised as it requires sentence-aligned training parallel corpus.
- It can effectively identify pairs of parallel paragraphs located anywhere on a web domain, regardless of its structure.
- The key step is based on recent ideas—the bilingual extension of **word2vec** and **locality-sensitive hashing**.
- Classification of candidate pairs is done using a **neural network** model based on a set of **custom features**.

CzEng Experiment

- Prealigned Czech–English parallel corpus unaligned to be aligned once again.
- Simulation with **147** web domains, each having app. **50,000** Czech and **50,000** English paragraphs, to be aligned.

Recall (%)	**63.28**
Precision (%)	**94.32**

Common Crawl Experiment

- Mining Czech–English parallel corpus from July 2015 Common Crawl dataset (**149 TB**, **1.84 billions** of web pages).
- Two MapReduce jobs to identify bilingual web domains and refine paragraphs from identified domains.
- We have identified **8,750** Czech–English web domains with **801,116** Czech and **5,931,091** English paragraphs.
- Our method has aligned **114,771** pairs of paragraphs from **2,178** web domains.
- The method's precision has been evaluated manually using random **500** aligned pairs.

Precision (%)	**94.60**

- To evaluate the recall, we have selected one web domain (**www.csa.cz**).

Recall (%)	**95.45**
Precision (%)	**97.67**