Názov: Rafinácia paralelných korpusov z webu
Autor: Bc. Jakub Kúdela
E-mailová adresa autora: jakub.kudela@gmail.com
Katedra: Katedra Softwarového Inženýrství
Vedúci práce: Doc. RNDr. Irena Holubová, Ph.D.
E-mailová adresa vedúceho: holubova@ksi.mff.cuni.cz
Konzultant práce: RNDr. Ondřej Bojar, Ph.D.
E-mailová adresa konzultanta: bojar@ufal.mff.cuni.cz

Abstrakt: Štatistický strojový preklad (SMT, statistical machine translation) je v súčasnosti jeden z najpopulárnejších prístupov ku strojovému prekladu. Tento prístup využíva štatistické modely, ktorých parametre sú získané z analýzy paralelných korpusov potrebných pre tréning. Existencia paralelného korpusu je najdôležitejšou prerekvizitou pre vytvorenie účinného SMT prekladača. Viaceré vlastnosti tohto korpusu, ako napríklad objem a kvalita, ovplyvňujú výsledky prekladu do značnej miery. Web môžeme považovať za neustále rastúci zdroj značného množstva paralelných dát, ktoré môžu byť rafinované a zahrnuté do trénovacieho procesu, čím môžu zdokonaliť výsledky SMT prekladača. Prvá časť práce sumarizuje niektoré z rozšírených metód pre získavanie paralelného korpusu z webu. Väčšina z metód hľadá páry paralelných webových stránok podľa podobnosti ich štruktúr. Veríme však, že existuje nezanedbatelné množstvo paralelných dát rozložených na webových stránkach, ktoré nezdieľajú podobnú štruktúru. V ďalšej časti predstavíme iný prístup ku identifikácii paralelného obsahu na webe. Tento prístup vôbec nezávisí na porovnávaní štruktúr webových stránok. Najskôr si predstavíme generickú metódu pre bilingválne zarovnanie dokumentov. Kľúčová časť našej metódy je postavená na kombinácii dvoch súčasne populárnych myšlienok, menovite bilingválneho rozšírenia modelu word2vec a lokálne-senzitívneho hašovania (locality-sensitive hashing). S metódou aplikovanou na úlohu získavania paralelných korpusov z webu, sme schopní efektívne identifikovať páry paralelných častí (paragrafov) nachádzajúcich sa na ľubovolných stránkach webovej domény bez ohľadu na štruktúru stránok. V poslednej časti naša práca opisuje experimenty vykonané s našou metódou. Prvý experiment využíva vopred zarovnané dáta a jeho výsledky sú vyhodnotené automaticky. Druhý experiment zahŕňa reálne dáta poskytované organizáciou Common Crawl Foundation a predstavuje riešenie pre úlohu rafinácie paralelných korpusov zo stoviek terabajtov veľkého množstva dát získaných z webu. Obidva experimenty ukazujú priaznivé výsledky, čo naznačuje, že navrhovaná metóda môže tvoriť nádejný základ pre nový spôsob získavania paralelných korpusov z webu. Veríme, že množstvo paralelných dát, ktoré je naša metóda schopná získať by mohlo zabezbečiť SMT prekladačom objemnejší tréning a tým pádom aj lepšie výsledky.

Kľúčové slová: rafinácia paralelných korpusov, bilingválne dokumentové zarovnanie, word2vec, lokálne-senzitívne hašovanie