Vícejazyčné vyhledávání informací v oblasti medicíny

Shadi Saleh

V posledních letech došlo k exponenciálnímu růstu objemu digitálního obsahu dostupného na internetu, což koreluje s rostoucím počtem neanglicky mluvících uživatelů internetu v důsledku celosvětového rozšíření internetu. To zvyšuje význam zpřístupnění informací pro ty, kteří si chtějí vyhledat informace bez omezení na jazyky, jimž rozumějí, zejména pro ty, kteří chtějí využít internet k vyhledání lékařského obsahu týkajícího se jejich zdravotního stavu. Získávání informací napříč jazyky (Cross-Lingual Information Retrieval, CLIR) překonává jazykové bariéry tím, že vyhledává dokumenty napsané v jiném jazyce, než je jazyk dotazu.

Tato disertační práce se zabývá úlohou CLIR v lékařské doméně v sedmi evropských jazycích. Vyvinuli jsme systémy pro statistický strojový překlad (Statistical Machine Translation, SMT) adaptované na oblast lékařství, které jsou doladěny pro metodu překladu dotazů (Query Translation, QT) a pro metodu překladu dokumentů (Document Translation, DT).

Na empirických datech dokládáme, že oproti tomu, co se předpokládalo od konce 90. let., v úloze CLIR v lékařské doméně metoda DT nepředčí metodu QT pro žádný ze zkoumaných jazyků, a to ani při použití SMT, ani při použití neuronového strojového překladu (Neural Machine Translation, NMT). Navrhujeme též reranker založený na strojovém učení, který mění pořadí překladových hypotéz systému SMT, aby dosáhl lepších výsledků v úloze CLIR. Systém jsme nejprve navrhli pro české, francouzské a německé systémy CLIR a poté přizpůsobili pro španělštinu, maďarštinu, švédštinu a polštinu. Dle našich zjištění nejlepší překlad dotazů vybraný výchozím SMT překladačem nemusí být nutně nejlepším překladem pro účely CLIR. Náš reranker produkuje překlady, které jsou optimalizovány pro CLIR, a statisticky významně zlepšují úspěšnost CLIR. Představujeme též novou metodu obohacení přeložených dotazů o nová slova, která napomáhá odstranění problémů s vágností přeložených dotazů. Každému uvažovanému slovu je přiřazeno skóre užitečnosti pomocí lineárního regresního modelu a prahového skóre, které je nastaveno tak, aby se zabránilo přidávání irelevantních slov, která by mohla poškodit původní dotaz. Naše metoda zlepšuje jak úspěšnost systémů CLIR ve všech zkoumaných jazycích, tak i jednojazyčné získávání informací (IR) v angličtině.

Pro porovnání SMT a NMT v kontextu CLIR jsme natrénovali NMT model zaměřený na úlohu CLIR, který překládá lékařské dotazy. Tento NMT model statisticky významně předčil přístup založený na technologii SMT, a to ve všech zkoumaných jazycích.

V průběhu našeho výzkumu jsme vyvinuli rozšířený datový soubor pro CLIR v lékařské doméně, který je založen na existujících datových souborech z vyhodnocování CLEF eHealth IR z let 2013–2015. Tento datový soubor je přístupný veřejnosti on-line.