Prihláška do súťaže TP Cup

Názov tímu: aDictIT
Téma projektu: Štatistický preklad textu
Vedúci projektu: Ing. Dušan Zeleník
e-mail: adictit@googlegroups.com
web: http://labss2.fiit.stuba.sk/TeamProject/2011/team04is-si/

Motivácia

Na svete existuje viac než 6900 rôznych jazykov. To vytvára jazykovú bariéru medzi miliónmi ľudí na svete, ktorí neovládajú rovnaký jazyk a tým pádom sa nemôžu dohovoriť. Tiež sa znižuje dostupnosť informácií, čo má za následok nižšiu informovanosť a vzdelanosť ľudí po celom svete. Svetový technologický a vedecký pokrok by mohol byť oveľa rýchlejší, ak by mali všetci dostupné literárne zdroje v materinskom jazyku. Situáciu najlepšie vystihuje biblické podobenstvo o babylonskej veži, kde jazyková bariéra zmarila stavbu veže siahajúcej do nebies.

História strojového prekladu siaha do obdobia po druhej svetovej vojne, no stále nedosahuje uspokojivé výsledky. Najznámejším webovým prekladačom textu je Google Translate. Jeho preklad dosahuje pomerne dobrú úspešnosť. Prístup k službe prekladača cez programátorské rozhranie Google nedávno spoplatnil. Celosvetová akademická obec by si však zaslúžila a určite aj využila bezplatný nástroj na preklad veľkého množstva textu. Nehovoriac o podnikateľskej sfére, kde zníženie nákladov často znamená konkurenčnú výhodu. Nami navrhovaný prístup k prekladu textu pritom v porovnaní s doposiaľ prezentovaným prístupom konkurencie prináša presnejšie výsledky efektívnejším spôsobom.

Naše riešenie

Cieľom nášho projektu je vytvoriť webovú službu, ktorá bude prekladať súvislý text, pre rôzne dvojice jazykov. Toto chceme dosiahnuť použitím jednoduchého prekladového slovníka a webu ako zdroju rozsiahlych textových korpusov v rôznych jazykoch. Náš nápad spočíva v tom, že vetu najskôr preložíme slovo po slove. Vygenerujeme všetky možné kombinácie prekladov vety, čo zahŕňa aj vygenerovanie všetkých slovných tvarov. Pre vygenerované vety potom analyzujeme pravdepodobnosti výskytov postupností slov, ktoré získame z rozsiahlych textových korpusov. Následne identifikujeme, ktorá veta vo vzorke jazyka je výsledným prekladom. Štatistický preklad podporujeme heuristikou, vo forme empiricky získaných pravidiel užitočných pri preklade a mechanizmami sociálnej kolaborácie pri spresňovaní prekladu podobne ako je to úspešné v konkurenčných riešeniach. Do projektu chceme zahrnúť aj spresnenie a personalizovanie prekladu zisťovaním kontextu. Toto chceme uskutočniť identifikovaním kľúčových slov, aby sme zistili doménu o ktorej text pojednáva a tým určili presnejší preklad. Výsledný produkt budeme porovnávať s webovými prekladačmi Google Translate a Bing Translator, aby sme demonštrovali dosiahnutú kvalitu.

Pri práci na projekte plánujeme využiť súčasné technológie (PHP, MySQL, Javascript, J2EE) . Pre vyhľadávanie v rozsiahlych textových korpusoch použijeme technológiu Elastic Search, ktorá v

súčasnosti dosahuje bezkonkurenčnú efektivitu v danej doméne a pre podporenie optimálneho výkonu využijeme distribuovanú architektúru Hadoop, ktorá je pre nami navrhované riešenie vhodná.

Výhody nášho riešenia

- bezplatná webová služba pre vývojárov
- webové používateľské rozhranie pre bežných používateľov
- možnosť využitia danej metódy prekladu pre veľkú množinu dvojíc jazykov
- spresnenie prekladu využitím kontextu
- efektívnosť dosiahnutá distribuovanou architektúrou

Náš tím

Tím aDictIT tvorí sedem mladých nadšencov pre moderné informačné technológie, ktorí sú závislí na neustálom získavaní nových znalostí a skúseností. Všetci majú skvelý prehľad v oblasti webových technológií. Teoretické vedomosti získané pri štúdiu si väčšina z nás stihla overiť aj v profesijnom živote. Našou ambíciou je nielen splniť úlohu, ale aj reálne prispieť našim projektom akademickej obci a spoločnosti všeobecne.

Bc. Róbert Horváth má odborné skúsenosti s využitím informačných technológií pre zisťovanie výkladu slova na základe kontextu, ktoré obhájil vo svojom bakalárskom projekte.

Bc. Peter Jurčík má rozsiahle znalosti webových technológií a webdizajnu, ktoré využíva aj vo svojej profesii.

Bc. Peter Macko sa okrem najmodernejších webových technológií venuje aj tvorbe počítačovej grafiky.

Bc. Vladimír Ruman je skúsený v oblasti webových technológií a certifikovaným odborníkom na sieťové technológie.

Bc. Peter Sládeček okrem skúseností z vývoja vlastného CMS získal aj cenné skúsenosti v oblasti testovania rozsiahlych informačných systémov zo svojej profesie.

Bc. Maroš Ubreži získal rozsiahle znalosti relačných databázových systémov počas štúdia a práci na svojej bakalárskej práci.

Bc. Matúš Vacula ako vývojár softvéru pre verifikáciu faktúr získal cenné skúsenosti so spracovaním veľkého množstva dát.