Slovenská technická univerzita
Fakulta informatiky a informačných technológií
Ilkovičova 3, 812 19 Bratislava

Priebežná správa

Tímový projekt II

Tím č. 4 – aDictIT

Bc. Róbert Horváth
Bc. Peter Jurčík
Bc. Peter Macko
Bc. Vladimír Ruman
Bc. Peter Sládeček
Bc Maroš Ubreži
Bc. Matúš Vacula

Vedúci: Ing. Dušan Zeleník
Akademicky rok: 2010/2011

Definícia cieľov a očakávaných dopadov/prínosov

Našimi cieľmi sú:

- vytvorenie prekladača
- použitie štatistík a jednostranného korpusu
- poskytnutie rozhrania pre externé aplikácie
- rýchly prístup k prekladom
- možnosť personalizácie prekladu

Prínosy, ktoré ponúka naše riešenie:

- rozhranie prekladača pre externé aplikácie
- jednoduchá rozšíriteľnosť prekladača o nové jazyky
- poskytnutie prispôsobeného prekladu na konkrétneho používateľa

Opis problémovej oblasti

Preklad textu z jedného jazyka do iného je netriviálny problém, ktorý sa ľudia snažia automatizovať už niekoľko desaťročí. Prvé pokusy o automatizáciu procesu prekladu sa datujú už od obdobia druhej svetovej vojny. Aj napriek tomu, že sa táto problematika skúma už dlhší čas, dnešné prekladače nie sú ani zďaleka dokonalé. Ich preklady sú často nepresné a z dlhších viet často vytvoria nezmyselné skupiny slov.

Väčšina z týchto prekladačov pritom funguje na princípe paralelného korpusu, ktorého získanie je však príliš zložité a preto väčšina menších prekladačov tohto typu nedosahuje pri prekladoch dostatočne dobré výsledky.

Veľké množstvo projektov na našej fakulte, ale aj vo svete, využíva možnosti automatických prekladačov ako Google Translate. Ten však v nedávnom čase spoplatnil svoje rozhranie pre vývojárov a tak nepríjemne ovplyvnil veľa nádejných projektov.

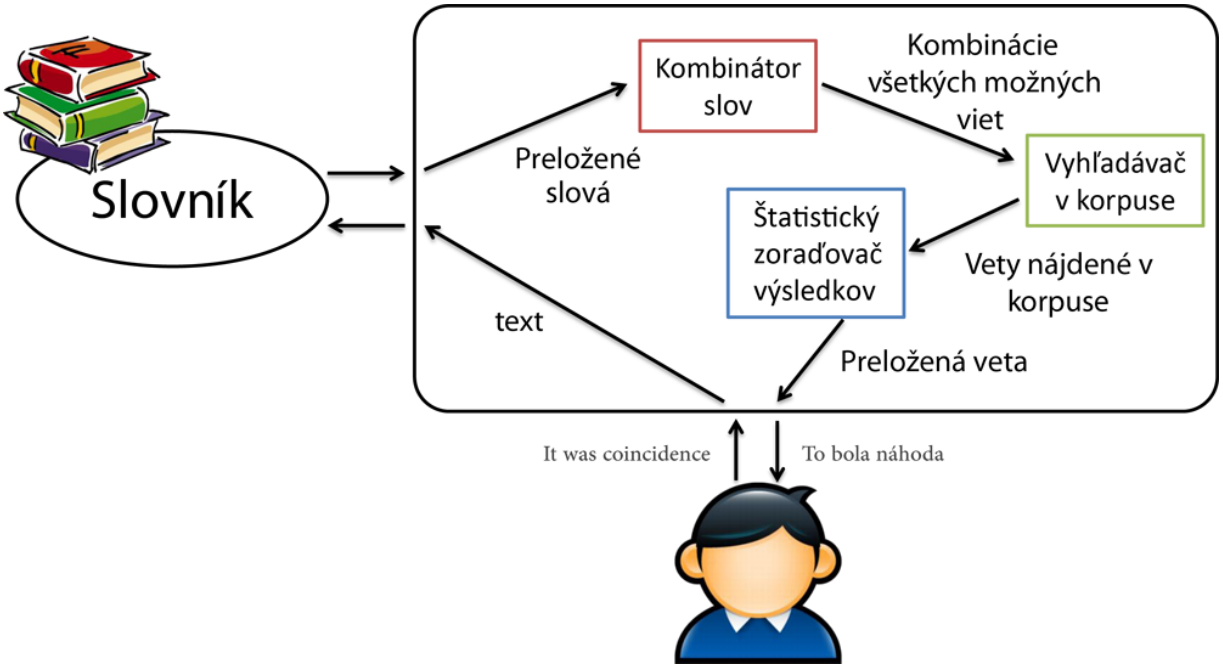
Prehľad riešenia

Nami navrhnutý prekladač sa od ostatných odlišuje absenciou obojstranného korpusu. V našom riešení sa v prvej fáze všetky slová zadaného textu preložia po slovách. Následne sa z týchto prekladov vygeneruje množina prípustných viet. Táto množina je tvorená kombináciami možných prekladov slov z vety. Jednotlivé vety je nutné následne vyfiltrovať tak, aby sa z kombinácií odstránili vety, ktoré majú minimálnu šancu na to, aby sa stali požadovaným prekladom zadanej vety.

Ďalším krokom je vyhľadanie zadaných formulácií v korpuse viet cieľového jazyka. Podľa toho, na koľko sa jednotlivé výsledky približujú vetám v korpuse, sú výsledky vyhodnotené a zoradené. Používateľovi sa teda zobrazia iba tie výsledky, ktoré sú najpravdepodobnejším prekladom ním zadanej vety.

Výhodou nášho riešenia je teda schopnosť s minimálnym úsilím prekladať texty z jedného ľubovoľného jazyka do druhého. Podporu ďalšieho jazyka je možné zabezpečiť pridaním jeho korpusu spolu a slovníka.

Predpokladáme, že náš prekladač by bolo možné s približne rovnakou úspešnosťou rozšíriť nielen o ľudské jazyky, ale aj tie programovacie. Výsledkom by bola transformácia zdrojového kódu z jedného programovacieho jazyka do druhého.



Zhrnutie súčasného stavu riešenia projektu

Naše riešenie je schopné prekladať jednoduchšie a kratšie texty. Aktuálne je podporovaný iba preklad z anglického jazyka do slovenského a späť. Našim zámerom bolo hlavne vytvoriť kvalitný produkt, ktorý bude neskôr jednoducho rozšíriteľný o iné jazyky.

Základom riešenia je anglický slovník a korpus slovenského jazyka, ktorý je získaný z dvoch zdrojov. Prvým je portál sme.sk, ktorý je zložený z množstva článkov napísaných v publicistickom štýle. Druhým sú titulky k filmom, ktoré reprezentujú hovorový jazyk.

Pre efektívnu prácu s korpusom sme hľadali technológiu, ktorá nám umožní rýchle prehľadávanie záznamov. Po podrobnej analýze dostupných technológii sme nakoniec sa rozhodli použiť ElasticSearch. Táto technológia okrem rýchleho vyhľadávania poskytuje aj číselné ohodnotenie zhody výsledku so zadaným vzorom a tak vieme jednoducho zoradiť dosiahnuté výsledky prekladov.

Na to aby, sme prekladač mohli testovať a verejne poskytovať sme potrebovali vytvoriť webovú službu, ktorá by zastrešovala jej možnosti. Po uskutočnení analýzy tejto oblasti sme sa rozhodli pre najmodernejšiu formu webových služieb a to službu typu REST. Používateľovi je tak poskytnuté vyhľadanie najúspešnejšieho prekladu, relevantných prekladov a vyhľadávanie všetkých možných prekladov danej vety.

Plán ďalšieho postupu

Už v týchto chvíľach usilovne pracujeme na ďalších možnostiach nášho prekladača. Podstatnou časťou zlepšovania prekladu je zrýchlenie jeho odozvy. V dnešnej dobe je prekladanie dosť pomalé, čo by pri dlhších vetách a textoch mohlo odradiť potenciálnych požívateľov. Zrýchlenie sa budeme snažit dosiahnuť v dvoch smeroch:

1. zakomponovaním moderných technológií - tu by sme chceli využiť už používaný ElaticSearch, ktorý poskytuje rôzne možnosti distribuovaného nasadenia a tak dosiahnutie väčšej rýchlosti vyhľadávania. Na to by nám aktívne mohla pomôcť aj technológia Hadoop, ktorá je na fakulte zastrešená, a ktorú sa hodláme vo finálnom stave používať.
2. vylepšením algoritmu prekladu - algoritmus prekladača sa snažíme zlepšovať každým dňom a v nasledujúcom období chceme vniesť do prekladu ešte viac štatistiky. Budeme sa snažiť ohodnotiť jednotlivé slová tak, aby sme vedeli hneď po vygenerovaní kombinácií viet vyfiltrovať tie, ktoré sú najmenej pravdepodobné. Okrem toho by sme chceli zaradiť do prekladu pravdepodobnosti n-gramov, pričom by sme hneď vedeli určiť, ktoré slová sa spolu najčastejšie vyskytujú. Týmto spôsobom by sme zvýšiť skóre viet, ktoré takéto kombinácie obsahujú.

V ďalšej časti semestra by sme sa radi venovali kontextu prekladu. V tejto časti by sme chceli identifikovať, či v kontexte vystupuje v hlavnej roly žena alebo muž, a tak prispôsobiť výsledky viac reálnemu textu. Neskôr by sme chceli priniesť aj možnosť prepínania korpusu. V prípade, že by používateľ chcel prekladať odborný text mal by sa ako korpus použiť zdroj odborných článkov a kníh. V prípade hovorového štýlu by to mali byť napríklad titulky, príspevky na rôznych portáloch a podobne.