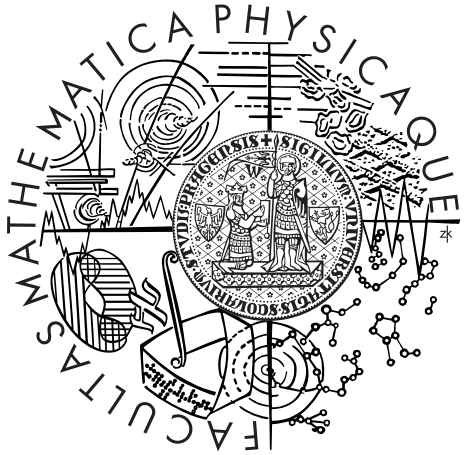
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Ondřej Odcházel

AJAX CAT - webový editor s podporou pro překlad

Ústav formální a aplikované lingvistiky

Vedoucí diplomové práce: RNDr. Ondřej Bojar, Ph.D.
Studijní program: Informatika
Studijní obor: Programování

Praha 2011

Poděkování.

Prohlašuji, že jsem tuto diplomovou práci vypracoval samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ..... Podpis autora

Název práce: AJAX CAT - webový editor s podporou pro překlad

Autor: Ondřej Odcházel

Katedra: Ústav formální a aplikované lingvistiky

Vedoucí diplomové práce: RNDr. Ondřej Bojar, Ph.D., Ústav formální a aplikované lingvistiky

Abstrakt:

Klíčová slova:

Title:

Author: Jméno a příjmení autora

Department: Název katedry či ústavu, kde byla práce oficiálně zadána

Supervisor: Jméno a příjmení s tituly, pracoviště

Abstract:

Keywords:

Obsah

Úvod

Díky novým statistickým přístupům zažívá obor strojového překladu jazyka v posledních letech opět velký rozvoj. Nové výpočetní i algoritmické možnosti umožňují vytváření stále lepších jazykových překladů. Stále se však nepodařilo vytvořit univerzální překladový systém, který by dokázal nahradit lidské překladatele, ani v jednom běžném jazykovém páru. Je stále otázka zdali se podobný překladový systém v budoucnosti lidstvu podaří sestavit. Již dnes jsou ale překladové systémy na úrovni, která sice nedokáže překladatele nahradit, ale v mnoha odvětvích usnadňuje jejich práci. Překladové systémy již nyní poskytují alespoň nápovědu, jak daný text přeložit. Překladatel však stále musí výstup z takového systému kontrolovat a editovat. Každý z těchto editačních zásahů představuje pro překladatele komplikaci a pokud je množství nutných zásahů nad nějakou hranicí, překladatel raději místo editování výstupu překladového systému vytvoří překlad sám. Pro zjednodušení překladatelské práce je tedy potřeba nejen zlepšovat tyto překladové systémy, ale také software kteří překladatelé pro interakci s překladovým systémem používají. Překladový software, který využívá pro nápovědu překladatelům strojový překlad je speciální případem CAT (computer–aided translation) systému.

Cílem této bakalářské práce je implementace jednoho CAT systému. Pro podporu překladu bude systém využívat překladový systém Moses. Celý projekt je rozdělen do dvou částí — serverová a klientská část. Implementací serverové části bude vytvořen HTTP server. Tento server bude spouštět Mosese a skrz HTTP požadavky bude poskytovat klientovi odpovědi. Požadavky budou dvojího typu. Klient se může systému zeptat na překlad věty v určité jazyce. Server pak odpoví tabulkou překladových možností. Sloupce této tabulky jsou jednotlivé úseky ve zdrojovém překladovém úseku (typicky slova ve větě). V řádcích tabulky jsou pak přeložené úseky textu v cílovém jazyce. Úseky jsou seřazeny v tabulce tak, že čím výše je daný úsek, tím větší je pravděpodobnost toho, že se jedná o "správný překlad". Taková to tabulka je jedním ze základů implementovaného CAT systému.

Dalším typem klientského požadavku bude jakási lokální nápověda během překladu. Jedná se o podobný druh nápovědy jakou nám poskytují například intenetové vyhledávače. V nich často uživatel nemusí psát celý vyhledávací dotaz a může využí nápovědy, která mu nabízí nejběžnější podobné dotazy. Podobně i implementovaný server bude dávat nápovědu, jak dále pokračovat s překladem. Aby překladový systém mohl tuto nápovědu poskytnout, pořebuje znát tři parametry. Text ve zdrojovém jazyce, vektor určující, které úseky jsou již přeloženy a již přeložená část věty v cílovém jazyce. Cílem této práce je i rozšířit možnosti Mosese tak, aby s těmito třemi parametry dokázal pracovat a vygeneroval nápovědu, jak v překladu pokračovat.

Samotná serverová část tak bude moci fungovat jako komponenta samostatně a poskytovat nápovědu k překladu i jiném CAT systému.

Druhou částí práce je implementace klientské části CAT systému. Tato část bude sloužit k interakci překladatele s překladový systémem. Tato interakce by měla být co nejvíce přátelská k uživateli. Ten typicky zadá zdrojový text pro překlad a zdrojový a cílový jazyk překladu. CAT systém tento zdrojový text rozseká

do bloků (typicky vět) a ke každé větě překladateli nabídne nápovědu generovanou v serverové části. Součástí implementace klientské části bude i jednoduchý systém pro zprávu obsahu, aby překladatel mohl pokračovat v překladu i po znovuotevření aplikace. Klientská část bude podobně jako serverová část fungovat sama o sobě. Pokud tedy nebude napojena na server, může pracovat sama o sobě jako systém pro správu překladů.

1. Překlad a strojový překlad

1.1 Překladové problémy

Překlad je proces přenesení významu z textu ve zdrojovém jazyce do jazyka cílového. Úloha překladu je složitá i tím, že žádný výsledek nejde označit za nejlepší. Že neexistuje dokonalý překlad lze ilustrovat na překladech knih, nebo divadelních her. Hry Williama Shakespearea byly z angličtiny do češtiny přeloženy mnohokrát, přesto jsou stále inscenovány hry s různými překlady.

Bez hlubších jazykových znalostí se může jevit úloha překladu snadná, mezi většinou jazyků máme přeci slovník. Ale pokud chceme přeložit anglické slovo "house" do češtiny, můžeme mít problém. Ve většině případů lze toto slovo přeložit jako "dům". Pokud ale překládáme text o anglické královně, kde se objeví sousloví "House of Windsor", zřejmě není řeč o domu, kde bydlí Windsorové, ale o "rodu Windsorů". Překladatel tedy při textu potřebuje znát kontext ve kterém je slovo použito a často také potřebuje mít odborné znalosti z oboru překládaného textu.

Jazyk není neměnný a v průběhu času se vyvíjí. Můžeme to vidět například na Bibli. Její nejznámější překlad, Bible Kralická je přes 300 (?) let starý. Vznikají proto nové překlady, které jsou dnešním čtenářům přístupnější. Žádný překlad tedy nelze označit za dokonalý a navždy správný.

1.2 Historie strojového překladu

Na počátku dějin strojového překladu stála, podobně jako v mnoha jiných oborech, armáda. Spojené státy Americké byli v padesátých letech ve Studené válce se Sovětským svazem. V této válce beze zbraní sehráli velkou úlohu i výzvědné služby, které zachytávali velké množství nepřátelských zpráv. Tyto zprávy bylo nutné co nejrychleji přeložit. A právě v této době se zrodila myšlenka použít k tomuto účelu počítače, které byli produktem předchozího válečného konfliktu, 2. světové války. (http://www.hutchinsweb.me.uk/GU-IBM-2005.pdf) Významnou demonstrací použití strojového překladu se v roce 1954 stal takzvaný Georgetownský experiment. Pro tento experiment vyvinula Georgetownská univerzita spolu s firmou IBM překladový systém pro překlad z ruštiny do angličtiny. Tento systém používal slovník 250 slov a 6 gramatických pravidel. Jeho doménou byly zejména překlady v oblasti chemie. Během experimentu bylo přeloženo více než 60 vět. Experiment byl všeobecně přijat jako úspěch, což donutilo americkou vládu investovat v následujících letech do oblasti strojového překladu.

Následovaly léta práce zejména v SSSR a USA na systémech pro automatické překlady zejména mezi ruštinou a angličtinou. Žádný dobře použitelný systém, který by poskytoval uspokojivé výsledky, však nevzniknul. Pochybnosti ohledně možností strojového překladu vyjádřil na konci padesátých let lingvista Yehoshua Bar-Hillel. Ten argumentoval pomocí anglické věty "The box was in the pen." Překlad této věty by mohl být: "Pero bylo v ohradě." Jelikož anglické slovo "pen" znamená "pero" i "ohrada", musí mít překladový systém, který chce větu přeložit správně, sémantickou informaci, která by mu napověděla, že krabice nemůže být peru, tedy že správným překladem slova "pen" do češtiny je v tomto kontextu slovo "ohrada".

http://www.hutchinsweb.me.uk/ALPAC-1996.pdf I z tohoto důvodu bylo vytvoření komplexního překladového systému v té době zřejmě nemožné. Americká vláda však dále pokračuje ve financování výzkumu. V roce 1966 vyšla zpráva skupiny ALPAC (Automatic Language Processing Advisory Committee). Která doporučovala americké vládě další postup při financování překladového výzkumu. Zpráva zmenšovala optimismus, vyvolaný zejména Georgetownským experimentem, že se v dohledné době podaří vytvořit kvalitní systém pro strojový překlad. Výsledkem bylo téměř úplné zastavení financování výzkumu americkou vládou. Výzkum dále pokračoval zejména v Evropě, nebo Kanadě. Právě v kanadském Montrealu vznikl systém METEO. Ten byl v letech 1981 až 2001 používán pro překlad meteorologických zpráv mezi angličtinou a francouzštinou. Právě omezená překladová doména systému umožnila nabízet kvalitní překlady předpovědí počasí.

1.3 Současnost strojového překladu

V posledních letech spolu s pokračující globalizací světa a stále vyšší penetrací internetového připojení se zvyšuje i poptávka po překladech. Nadnárodní firmy potřebují při svém růstu stále více překladů. Dalším impulzem zvyšujícím popávku po překladech je i rozšiřování Evropské Unie. V současnosti unie používá 23 oficiálních jazyků ve kterých musí být přístupné všechny důležité úřední dokumenty. Tvorba tolika překladů je velice pracná a nákladná, což vytváří poptávku po zjednodušením procesu překladu.

Výpočetní výkon počítačů stále roste rychlostí Mooreova zákona (ftp://download.intel.com/
což v posledních letech otevřelo možnosti pro použití statistických metod ve strojovém překladu. Toho využívá statistický překladový systém Moses, open-source překladový systém, který používám i ve svém projektu a cílem projektu byla i implementace nových rozšíření. Dalším známým statistickým překladovým systémem je Google Translate.

Kromě systému využívajících statických postupů se stále vyvíjí i pravidlové překladové systémy. Kromě mnoha proprietálních systémů bych zmínil open-source systém Apertium. Stejně jako u dalších podobných systémů se pro každou dvojici překládaných jazyků musí vytvořit slovníky a překladová pravidla. Je velmi náročné a nákladné tato pravidla vytvořit. Navíc jsou tato pravidla často použitelná pouze pro jeden jazykový pár.

1.4 Computer-aided translation

CAT, neboli computer-aided translation, či computer-assisted translation je zkratka označující systémy pro podporu překladu. Tyto systémy poskytují překladateli podporu při překladu. Mohou to být jak desktopové, tak online aplikace a často se liší stylem, jakým překladatele podporují. Jedním z druhů podpory může být nabídka předchozích překladů z paměti. Této paměti se říká překladová paměť, obsahuje přeložené úseky textu a překladatel si tuto paměť buduje buď sám, nebo může využít nějakou z kolektivních databází. Příkladem desktopové aplikace může být například OmegaT. Ta je určena pro použití profesionálními překladateli, kterým nabízí úseky z překladové paměti. Hledaný úsek nemusí odpovídat

aktuálně překládanému úseku na 100 procent, OmegaT implementuje algoritmus, který pozná i blízké shody.

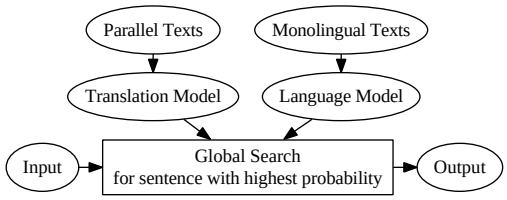
Další ukázkou CAT pomůcky je Google Translator Toolkit (Nástroje pro překladatele). Tato internetová aplikace umožňuje překladateli nahrát si svou překladovou paměť, kterou pak může využít při překladu. Dále nástroj nabízí výsledky překladu z Google Translate, který dále může uživatel editovat.

1.5 Moses

2. Moses

Moses je statistický překladový systém založený na překladu frází (anglicky PBMT - phrase-based machine translation).

2.1 Princip frázového překladu



Obrázek 2.1: Architektura frázových překladových systémů.

Současné nejvyspělejší generické (tedy více či méně nezávislé na jazykovém páru) překladové systémy jsou založeny na překladu frází. Jejich architektura je znázorněna na Obrázku 2.1.

2.2 Jazykový a překladový model

Tyto systémy používají k překladu vstupu na výstup překladový a jazykový model. Překladový model zachycuje četnost překladů frází délky n. Tyto fráze jsou také označovany jako n-gramy. Překladový model lze získat z paralelních korpusů, tedy přeložených textů mezi zdrojovým a cílovým jazykem. Jazykový model zachycuje četnost výskytu frází v cílovém jazyce. Pro jeho získání tedy stačí texty v cílovém jazyce. Díky překladovému a jazykovému modelu lze získat různé překlady frází ve vstupním textu. Pro vygenerování výstupu potřebuje překladový systém nalézt posloupnost frází v cílovém jazyce, které pokrývají všechny části vstupu a mají největší pravděpodobnost výskytu.

2.3 Hledání nejlepších hypotéz

Při generování překladu používá Moses strukturu, která se anglicky nazývá "lattice", neboli graf slov.

Variant překladu může být obecně exponenciální množství. Pro rychlé hledání v těchto variantách implementuje Moses tzv. beam search algoritmus. Postup algoritmu při hledání zobrazuje Obrázek 2.2. Jedná se vlastně o orientovaný graf. Vrcholy grafu znázorňují částečné hypotézy. Každá taková částečná hypotéza pokrývá nějaká slova ze vstupu a má skóre, které určuje kvalitu hypotézy. Na začátku je překladu je prázdná hypotéza, která nepokrývá žádnou část vstupu. Na konci máme hypotézy které pokrývají celý vstup. Během překladu překladový systém postupně rozšiřuje exitující hypotézy o další fráze. Hypotéza s nejvyšším skóre je nabídnuta jako překlad na výstup. Pomocí tabulky překladovýchmožností dokáže Moses poskytnou i další hypotézy seřazené podle pravděpodobnosti.

2.4 Rozšíření Mosese

Pro poskytnutí nápovědy během překladu bylo nutné rozšířit možnosti Mosese tak, aby dokázal generovat hypotézy z neprázdné počáteční hypotézy. Překladatel je uprostřed překladu věty, má přeložená určitá slova a potřebuje radu, jak nejlépe pokračovat dál. Pomocí implementovaného rozšíření se nyní může zeptat Mosese, který může začít generovat překlad navazující na již přeloženou část. K tomu potřebuje vektor označující části vstupní věty, které byly již přeloženy. Kvůli jazykovému modelu, který kontroluje, aby navazující část byla v cílovém jazyce co nejvíce smysluplná, potřebuje Moses znát část přeloženého textu, na kterou se může pokusit navázat. Pomocí tohoto rozšíření může Moses začít rozvíjet hypotézy, které se v prvotním překladu z prázdné hypotézy nemusely vůbec objevit. Toto by mělo přispět k flexibilitě nápovědy. Správný překlad totiž leckdy může být specifický a použitá slova nemusí být v překladovém modelu vůbec použita.

3. Implementace

3.1 Instalace serverové části

3.2 Instalace klientské části

Závěr

Seznam použité literatury

Seznam tabulek

Seznam použitých zkratek

Přílohy