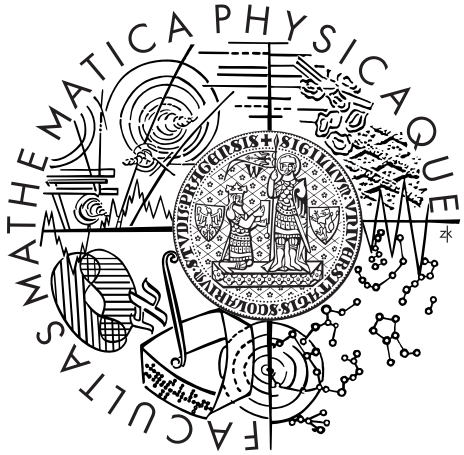
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

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

Důkaz. Neplatné věty nemají důkaz. □

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 pře-

klad 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.

Bez hlubších jazykových znalostí se může jevit úloha překladu snadná, mezi většinou jazyků máme přeci slovník. Ale tak snadné to není. Vezměme si například anglické slovo "house", které bychom chtěli přeložit do češtiny. 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 400 let starý. Vznikají proto nové překlady, které jsou dnešním čtenářům přístupnější. Žádný překlad tedy nelze označit za dokonalý a navždy správný.

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.

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

V této kapitole bude popsána architektura frázových překladových systémů (anglicky PBMT - phrase-based machine translation). Konkrétně bude popsána architektura systému Moses.

2.1 Princip frázového překladu

2.1.1 Jazykový a překladový model

Obecně lze postup při překladu znázornit pomocí Obrázku 2.1. Moses používá k překladu vstupu na výstup překladový a jazykový model. Překladový model zachycuje četnost překladů frází délky n. Tyto fráze jsou také označovány jako n-gramy. Překladový model lze získat z paralelních korpusů, tedy přeložených textů mezi zdrojovým a cílovým jazykem. Kvalita překladů je často úměrná velikosti paralelních korpusů, ze kterých překladový model generujeme. Tato závislost však není lineární, zdvojnásobení korpusu zlepší překladový model jen o málo.

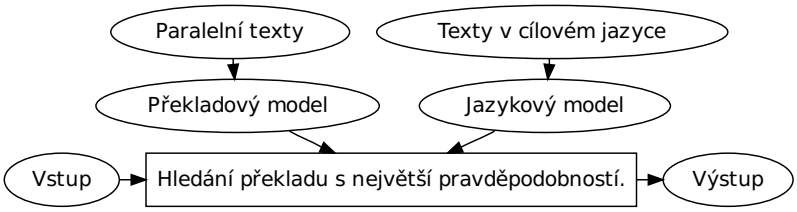
Jazykový model zachycuje četnost výskytu frází v cílovém jazyce. Díky jazykovému modelu dokáže Moses při generování překladu alespoň částečně ohlídat, aby text v cílovém jazyce dával smysl. Kvalitní jazykový model lze získat poměrně snadno - stačí velké množství textů v cílovém jazyce.

Moses pak zjednodušeně hledá překlad tak, že každé části vstupu přidává (vstupním frázím) přiřazuje fráze z překladového modelu. Pomocí jazykového modelu kontroluje alespoň částečně jazykovou korektnost výstupu. Hledání nejlepšího překladu je vlastně hledáním nejpravděpodobnější posloupnosti slov, která pokrývá všechny části vstupu. Možností překladu je obecně exponenciální množství. Přestože se u překladových systémů cení na prvním místě kvalita překladu, získání odpovědi v dostatečném čase je často také důležitou podmínkou. Rychlé hledání v překladových hypotézách je jednou z klíčových vlastností Mosese.

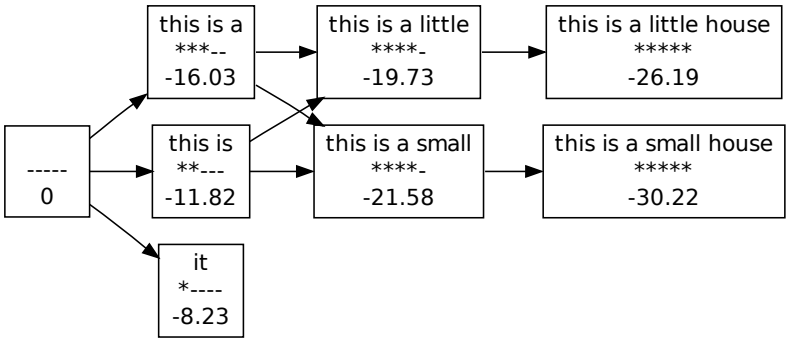
2.1.2 Hledání nejlepších hypotéz

Při generování překladu používá Moses strukturu, která se anglicky nazývá "lattice", neboli graf slov. Jeden takový graf je znázorněn na Obrázku 2.2. Jedná se vlastně o orientovaný graf. Jeho vrcholy jsou částečné hypotézy. Každá částečná hypotéza pokrývá nějakou část vstupu posloupností frází a má skóre určující relativní kvalitu překladu. Graf na obrázku popisuje překlad věty "das ist ein kleines haus" z němčiny do angličtiny. Pro názornost je tento graf velmi malý. Typicky by tento graf měl mnohem více vrcholů i hran.

Hrany v tomto grafu znázorňují jednotlivé kroky algoritmu známého pod názvem "beam search". Tento algoritmus implementovaný v Mosesovi se snaží postupně rozvíjet částečné hypotézy. Obrázek 2.3 opět popisuje překlad věty "das ist ein kleines haus" z němčiny do angličtiny. V tmavě označené vrcholu máme hypotézu, která pokrývá první dvě slova na vstupu, tedy "das ist" frází "this is". Algoritmus se pokusí tuto hypotézu rozvinout. Z překladového modelu získá dvě možnosti překladu fráze "ein kleines" — buď jako "little house", nebo jako "small house". V dalších krocích tyto hypotézy dále rozvíjí podobným způsobem



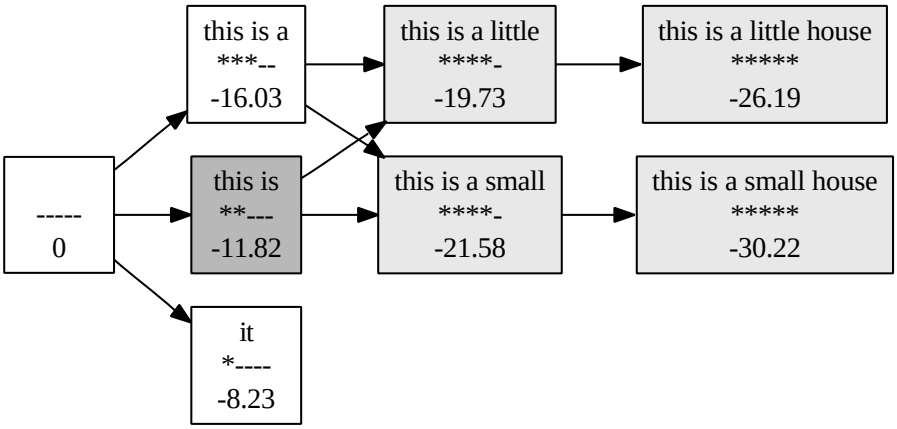
Obrázek 2.1: Architektura frázového překladového systému.



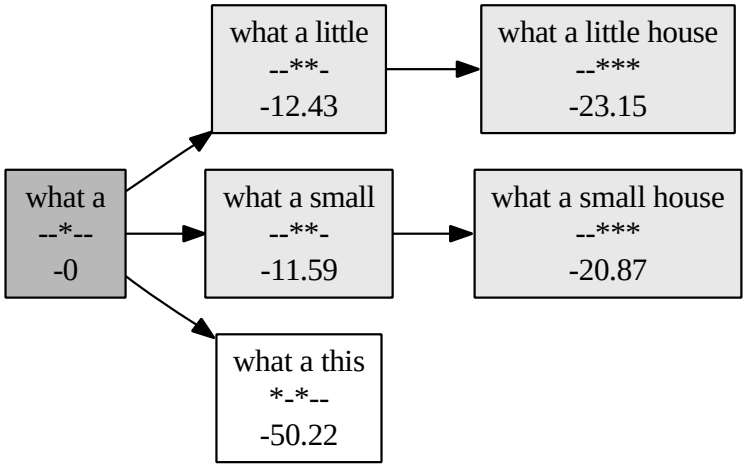
Obrázek 2.2: Graf slov, ve kterém Moses hledá nejlepší překlad.

a na konci má dvě překladové hypotézy, které pokrývají všechna vstupní slova — "this is a little house" a "this is a small house". Tyto hypotézy mají skóre, takže je lze seřadit a jako výstup nabídnout hypotézu s nejvyšším skóre. Pomocí tabulky částečných překladů dokáže Moses nabídnout i hypotézy, které při běhu algoritmu nezískaly nejvyšší skóre. Často se totiž hodí znát i jiné možnosti překladu, než jen tu, kterou Moses označil za nejlepší. Graf je podobný orientovanému stromu. Kořenem je počáteční prázdná hypotéza a listy jsou hypotézy, které už nejdou rozšířit. O strom se však nejedná kvůli možnosti spojení hypotéz. Často se může stát, že k jedné hypotéze dopracujeme pomocí více posloupností částečných hypotéz. Na Obrázku 2.3 to lze vidět na hypotéze "this is a little", která rozšiřuje dvě jiné hypotézy.

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.



Obrázek 2.3: Moses rozšiřuje částečnou hypotézu.



Obrázek 2.4: Moses hledá hypotézu z neprázdné počáteční hypotézy.

2.2 Rozšíření Mosese

Přireálném překladu se často může stát, že překladatel použije překlad, který neodpovídá žádné částečné hypotéze v Mosesovi. Jedním z cílů tohoto projektu je tedy i rozšíření Mosese tak, aby mohl generovat nápovědu z částečných hypotéz, které mu poskytne překladatel. Pokud zůstaneme u překladu věty "das ist ein kleines haus", může jeho překlad začít překladatel tak, že přeloží slovo "ein" frází "what a". Poté se chce překládat dále, ale neví jak, zeptá se tedy Mosese a řekne mu které části vstupu přeložil a jak je přeložil. Na Obrázku 2.4 je ukázka toho, jak by měl Moses na tento druh dotazu zareagovat. Místo prázdné hypotézy použije na začátku algoritmu hypotézu s parametry od uživatele.

3. Implementace serverové části

3.1 Rozšíření Mosese

3.1.1 Jak funguje Moses

Program Moses můžeme spustit v příkazové řádce typicky způsobem.
$ echo "das ist ein kleines haus" — moses -f moses.ini
Moses dostane text ve zdrojovém jazyce

4. Implementace serverové části

Na přiloženém CD je

4.1 Instalace

-make, g++, libtool, autoconf
svn checkout https://gnunet.org/svn/libmicrohttpd/
soubor INSTALL
c++0x gnu svn checkout svn://gcc.gnu.org/svn/gcc/trunk

4.2 Virtuální systém

Celý překladový systém lze spustit.

4.3 Instalace Mosese

4.4 Instalace serverové části

4.5 Instalace klientské části

Závěr

Seznam použité literatury

Seznam tabulek

Seznam použitých zkratek

Přílohy

Seznam použité literatury

[KH09] Philipp Koehn and Barry Haddow. Interactive Assistance to Human Translators using Statistical Machine Translation Methods. In *MT Summit XII*, 2009.