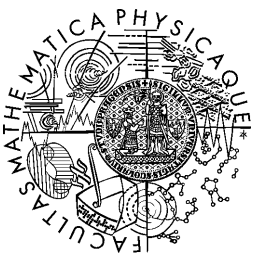
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Eva Viktorinová

Strojový překlad do češtiny přes lematický text

Ústav formální a aplikované lingvistiky

Vedoucí bakalářské práce: RNDr. Ondřej Bojar, Ph.D.
Studijní program: informatika, obecná informatika

2010

Na tomto místě bych ráda poděkovala vedoucímu mé bakalářské práce za pomoc a podnětné rady.

Prohlašuji, že jsem svou bakalářskou práci napsala samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce a jejím zveřejňováním.

V Praze dne Eva Viktorinová

Obsah

Název práce: Strojový překlad do češtiny přes lematický text
Autor: Eva Viktorinová
Katedra (ústav): Ústav formální a aplikované lingvistiky
Vedoucí bakalářské práce: RNDr. Ondřej Bojar, Ph.D.
e-mail vedoucího: Ondrej.Bojar@mff.cuni.cz

Abstrakt: Bakalářská práce se zabývá možností vylepšení překladu z angličtiny do češtiny. Popisuje problém bohaté české morfologie a navrhuje několik metod řešení tohoto problému převedením češtiny na lematický text. Podrobněji zpracovává překlad využitím alternativních dekódovacích cest, kde první cesta překládá z angličtiny do tvarované češtiny a druhá z angličtiny do lematické češtiny.

Klíčová slova: strojový překlad, lema, Moses, poziční značkování, alternativní dekódovací cesta

Title: Machine Translation to Czech via Lemmatized Text
Author: Eva Viktorinová
Department: Institute of Formal and Applied Linguistics
Supervisor: RNDr. Ondřej Bojar, Ph.D.
Supervisor's e-mail address: Ondrej.Bojar@mff.cuni.cz

Abstract: This work tries to improve machine translation from English to Czech language. It describes issue of rich Czech morphology and suggests several methods of solving this problem by lemmatization of Czech text. The translating with alternative decoding paths, where the first path translates from English to Czech and the second one from English to lemmatized Czech, is studied more closely.

Keywords: machine translation, lemma, Moses, positional tags, alternative decoding path

Kapitola 1

Úvod

Cílem bakalářské práce je experimentálně ověřit, zda je možné zlepšit kvalitu *frázového statistického strojového překladu* z angličtiny do češtiny použitím tzv. *lematického textu*, tedy zjednodušené češtiny, kde slova nejsou nijak morfologicky tvarována. Práce navrhuje několik možných variant vylepšení překladu oproti normálnímu překladu pomocí překladače `Moses` [1]. Některé z navržených postupů vyhodnocuje pomocí automatické evaluace kvality překladu a ručním porovnáním malého vzorku výstupů.

1.1 Motivace pro využití lematizace

Čeština patří mezi slovanské jazyky s velmi bohatým tvaroslovím a poměrně volným slovosledem. Jednotlivé větné členy tedy nemají pevně daný pořádek a syntaktické vztahy ve větě jsou zachyceny právě pomocí morfologie. Angličtina má naopak velmi jednoduché tvarosloví. Pro porovnání velký označkovaný český korpus obsahuje asi 2000 morfologických značek, zatímco English Penn Treebank je označkován pouze 50 značkami. [2]

Rozdílné velikosti slovníku pro paralelní český a anglický text Informace z tabulky 1.1 demonstrují problém různých velikostí slovníku u paralelního českého a anglického korpusu. Při srovnatelném počtu tokenů je na použitých datech počet různých slov v anglickém textu a počet různých slov v českém textu téměř dvojnásobný. Právě tento problém se snaží práce odstranit použitím lematického textu, protože počet různých slov v anglickém a českém textu převedeném na základní tvary je téměř stejný.

Výhodou lematizace je výrazná redukce tvaroslovné bohatosti češtiny a tedy menší nároky na objem trénovacích dat.

	# tokenů	# různých slov
anglický text	1,8M	67,9k
anglický lematizovaný text	1,8M	54,0k
český text	1,6M	127,4k
český lematizovaný text	1,6M	65,8k

Tabulka 1.1: Srovnání velikosti anglického a českého slovníku u paralelních korpusů

Ředění dat Bez převedení na základní tvary dochází k zbytečnému „ředění" trénovacích dat. Protože pokud anglickému slovu odpovídá n českých tvarů, musí bez převedení stroj získat slovník s n-násobným počtem položek. V praxi paralelní data objemem nestačí, a proto systém při předem daném trénovacím textu uvidí každou překladovou dvojici n-krát méně často, resp. mnoho variant slov bude ve slovníku chybět. Srovnání počtu slovníkových hesel v běžné a lematické frázové tabulce je vidět v tabulce 1.2.

	# slovníkových hesel
angličtina → čeština	2,167M
angličtina → lematická čeština	2,010M

Tabulka 1.2: Srovnání počtu slovníkových hesel v běžné a lematické frázové tabulce

Poměry OOV Využití lematického českého textu jako další překladové tabulky snižuje nepatrně OOV cílového jazyka, jak je vidět v tabulce 1.3. Toto OOV udává procento n-gramů cílového jazyka, které se vyskytly v testovacích datech, ale nevyskytly se v trénovacích datech (tj. slova, která `Moses` nemohl vygenerovat, ale požadovalo se to po něm).

Poměry OOV nenaznačují, že by překlad přidáním lematické češtiny nabízel výraznou možnost zlepšení. Ale OOV lematické češtiny proti lematizovanému referenčnímu překladu je mnohem blíže OOV pro angličtinu. Překlad doplněný o lemata bude tedy pro uživatele pravděpodobně příjemnější.

		n-gramy			
trénovací data	**testovací data**	**n=1**	**n=2**	**n=3**	**n=4**
čeština	**čeština**	4,3 %	38,5 %	72,1 %	86,3 %
čeština s lemat. češtinou	**čeština**	4,1 %	37,4 %	71,5 %	86,1 %
angličtina	**angličtina**	1,9 %	22,2 %	57,1 %	80,0 %
lematická čeština	**lematická čeština**	2,4 %	26,9 %	63,2 %	82,0 %

Tabulka 1.3: Hodnoty OOV pro korpusy

Překlad z angličtiny do češtiny bez morfologie Lematizace tedy dává velkou možnost na zlepšení překladu z angličtiny do češtiny. V tabulce 1.4 je vidět, že pokud bychom v překladu neuvažovali tvary, došlo by k výraznému zvýšení kvality překladu měřeno automatickou metrikou založenou na podobnosti hypotézy a referenčního překladu. Tabulka obsahuje výsledky automatického vyhodnocení normálního překladu z angličtiny do češtiny a překladu natrénovaném na anglickém a českém lematizovaném korpusu, kde výstup překladu byl vyhodnocen oproti lematizovanému referenčnímu překladu.

	NIST skóre	**BLEU skóre**
normální překlad	4.3449	0.1724
překlad bez ohledu na slovní tvary	5.6500	0.2411

Tabulka 1.4: Porovnání BLEU skóre normálního překladu a překladu do lemat vyhodnocenému oproti lematizovanému referenčnímu překladu

Bohatá česká morfologie je tedy očividně velkým problémem při překladu mezi angličtinou a češtinou. Výsledky statistického překladu do morfologicky bohatých jazyků dávají při použití automatické metriky horší výsledky než v opačném směru. [3] Podle [4] je do češtiny posteditovatelných přibližně 30 % vět, do angličtiny podle zdrojového jazyka až 52 %. Je-li zdrojový jazyk čeština je posteditovatelných jen 25 %.

Tato práce se tedy soustředí na vylepšení překladu v tomto směru - z angličtiny do češtiny.

1.2 Struktura tohoto textu

Bakalářská práce má následující strukturu. Tato první kapitola sestává z úvodu do problému, který práce řeší. Druhá kapitola se zabývá návrhem metod pro vylepšení překladu z angličtiny do češtiny pomocí lematického textu a jejich implementací. Ve třetí kapitole je vyhodnocení jejich úspěšnosti a příklady výstupu překladu. Čtvrtá kapitola obsahuje závěr se stručným zhodnocením. Dále následují přílohy.

V textu byly použity následující textové konvence

- pojmy jsou v textu zvýrazněny *kurzívou*
- názvy souborů, programů a skriptů jsou v textu napsány `tímto písmem`
- použité nástroje jsou uvedeny **tučně**

1.3 Používané pojmy

V této podkapitole následuje vysvětlení v práci dále používaných pojmů:

statistický strojový překlad
: Statistický strojový překlad je překlad textů v jednom lidském jazyce do druhého pomocí počítače, který se naučil překládat z velkého množství paralelních přeložených textů.

paralelní korpus
: Paralelním korpusem se rozumí dvojice dokumentů (nebo množina dvojic dokumentů), které jsou v různých jazycích a jsou sobě (velmi pravděpodobně) překladem. Standardní formát pro dvojici dokumentů je: věty jsou na samostatných řádcích, i-tá věta v prvním dokumentu je překladem i-té věty ve druhém dokumentu, věty jsou tokenizované

token, tokenizace
: Tokeny jsou prvky věty, tedy slova, čísla nebo interpunkce. Tokenizace je proces, během kterého se vstupní text rozdělí na tokeny, které budou mezi sebou odděleny mezerou. V této práci jsou dále výrazy token a slovo používány jako zaměnitelné.

frázový překlad, fráze
: Frázový překlad probíhá následujícím postupem. Vstupní věta je rozdělena do frází, tedy do posloupnosti po sobě jdoucích slov. Každá fráze je přeložena do cílového jazyka a fráze v cílovém jazyce mohou být přeuspořádány. Fráze ve frázovém překladu je tedy jakákoli souvislá posloupnost tokenů i zcela bez lingvistického opodstatnění.

lematizace
: Lematizace je proces, ve kterém jsou slova převedena na svůj základní tvar, tedy podstatná jména, přídavná jména, zájmena a číslovky na první pád čísla jednotného rodu mužského, přídavná jména a příslovce na první stupeň, slovesa na infinitiv, atd.

lematický text
: Lematický text je text vzniklý lematizací původního tvarovaného textu. Slova v něm nejsou nijak morfologicky tvarována.

n-gram
: n-gram je posloupnost po sobě jdoucích slov délky n.

faktor
: V překladači `Moses` může být každé slovo reprezentováno posloupností faktorů, například tvarem slova, základním tvarem, větným členem, poziční značkou. Jednotlivé faktory jsou ve vstupním textu odděleny pomocí oddělovače `|`.

automatické ohýbání
: Automatické ohýbání slov je problém z oblasti počítačové lingvistiky, kdy na vstupu je zadána lematická věta a úkolem je najít jí odpovídající zohýbanou (vyskloňovanou, vystupňovanou a vyčasovanou) větu.

poziční značkování
: Poziční značkování je způsob, jak u tvaru slova zaznamenat morfologické informace [5]. Značka má tvar řetězce 15 znaků, kde každá pozice znamená jednu morfologickou kategorii (např. pád, číslo, čas, osoba), pokud má kategorie pro dané slovo smysl. Např. poziční značka pro slovo `auty` je `NNNP7-----A----` říká, že jde o podstatné jméno středního rodu, množné číslo, v 7. pádě

OOV, neboli *out-of-vocabulary rate*
: Je definováno jako:

  - procento n-gramů cílového jazyka, které se vyskytly v testovacích datech, ale nevyskytly se v trénovacích datech
  - procento n-gramů zdrojového jazyka, které se vyskytly v testovacích datech, ale nevyskytly se v trénovacích datech

posteditace
: Posteditace je dodatečná úprava automaticky zpracovaného textu živou osobou.

Kapitola 2

Návrh a vlastní implementace metod pro vylepšení překladu z angličtiny do češtiny

V následujících podkapitolách jsou navrženy metody pro vylepšení překladu z angličtiny do češtiny použitím lematického textu.

2.1 Postup při normálním překladu z angličtiny do češtiny

Bakalářská práce se snaží vylepšit úspěšnost překladu z angličtiny do češtiny oproti následujícímu postupu, který je podrobně popsán v [6]:

- **trénování modelu angličtina - tvarovaná čeština** - sběr překladů frází z trénovacího paralelního anglického a českého tvarovaného korpusu
- **doladění vah modelu** - doladění vah v modelu (po natrénování jsou v `moses.ini` defaultní hodnoty) pomocí metody MERT (minimum error rate training [7]) na menších datech - na vývojovém paralelním anglickém a českém korpusu

  Při trénování se vytvoří následující modely:

  - **frázová překladová tabulka** - zajišťuje, že české a anglické fráze jsou sobě navzájem dobrým překladem
    $\phi(f|e)$ s váhou $weight_\phi$

  - **jazykový model** - zajišťuje, že na výstupu překladu bude plyná čeština
    $LM(e)$ s váhou $weight_l$
  - **slovosledný model (distortion model)** - umožňuje přerovnání vstupní věty
    $D(e, f)$ s váhou $weight_d$
  - **slovní penalizace** - zajišťuje, že překlady nejsou příliš dlouhé nebo příliš krátké
    $W(e)$ s váhou $weight_w$

  Pravděpodobnost, že anglická věta `f` bude přeložena do češtiny jako věta `e`, lze tedy matematicky vyjádřit jako:

$$p(e|f) = \phi(f|e)^{weight\_\phi} LM(e)^{weight\_l} D(e, f)^{weight\_d} W(e)^{weight\_w} \quad (2.1)$$

- **samotný překlad** - přeložený text bude vyhodnocen na úspěšnosti pomocí BLEU [8] a NIST skóre [9] na testovacích datech - překlad zadaného anglického testovacího textu pomocí překladače `Moses` se porovná s referenčním překladem

Průběh normálního překladu z angličtiny do češtiny je názorně zobrazen na obrázku 2.1.

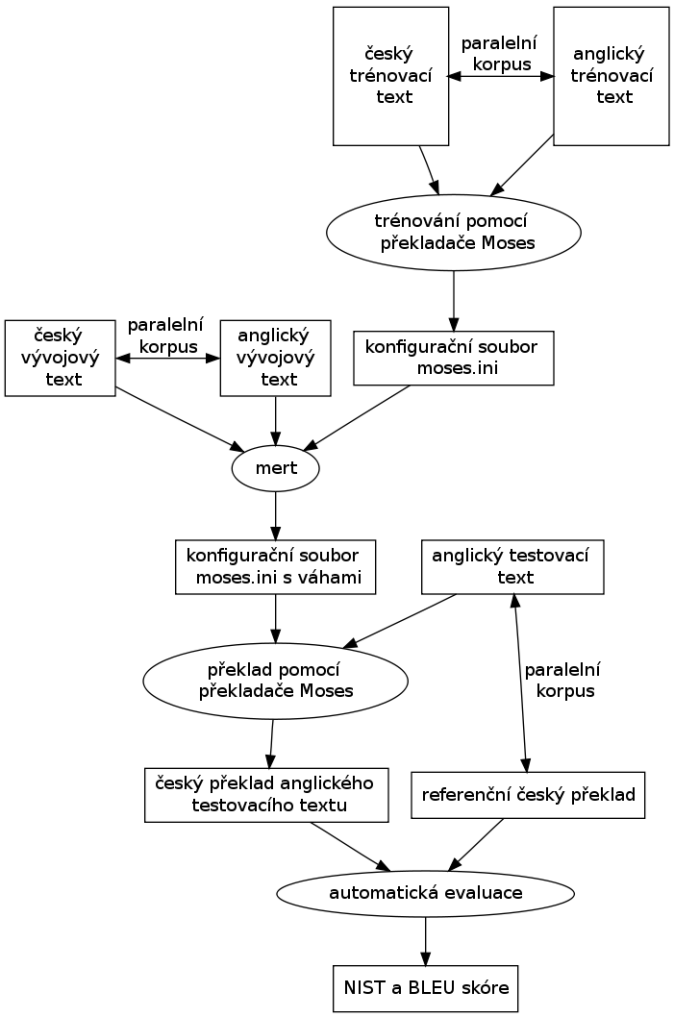
2.2 Překlad pomocí dvojího běhu překladače

První navrženou možností, jak využít lematizaci češtiny pro překlad z angličtiny do češtiny, je dvoufázový překlad. V prvním kroku se anglický text přeloží z angličtiny do lematické češtiny a poté se lematická čeština převede na tvarovaný text.

První překlad probíhá mezi korpusy se srovnatelnou velikostí slovníku, tím se odstraní problém se zbytečným ředěním dat. V druhém překladu se vygenerují tvary slov na základě kontextu podle druhé natrénované překladové tabulky. Zejména je tu možnost použít jako trénovací data pro druhý model velký jednojazyčný korpus.

Průběh překladu pomocí dvojího běhu překladače je názorně zobrazen na obrázku 2.2 a má následující fáze:

- **vytvoření prvního překladového modelu pro překlad z angličtiny do lematické češtiny**



Obrázek 2.1: Diagram normálního průběhu překladu z angličtiny do češtiny pomocí překladače Moses

  - **trénování prvního modelu angličtina - lematická čeština** - trénování překladače na paralelním anglickém a českém lematickém korpusu, český lematický korpus vznikne lematizací [1] českého korpusu
  - **doladění vah prvního modelu** - doladění vah v prvním modelu pomocí metody MERT na menších datech - na vývojovém paralelním anglickém a českém lematickém korpusu

- **vytvoření druhého překladového modelu pro překlad z lematické češtiny do tvarované češtiny**
  - **trénování druhého modelu lematická čeština - tvarovaná čeština** - trénování překladače na českém lematickém korpusu a českém tvarovaném korpusu
  - **doladění vah druhého modelu** - doladění vah v druhém modelu pomocí metody MERT na menších datech - na vývojovém paralelním českém lematickém a českém tvarovaném korpusu. Vývojový český lematický korpus vznikl lematizací českého tvarovaného korpusu.

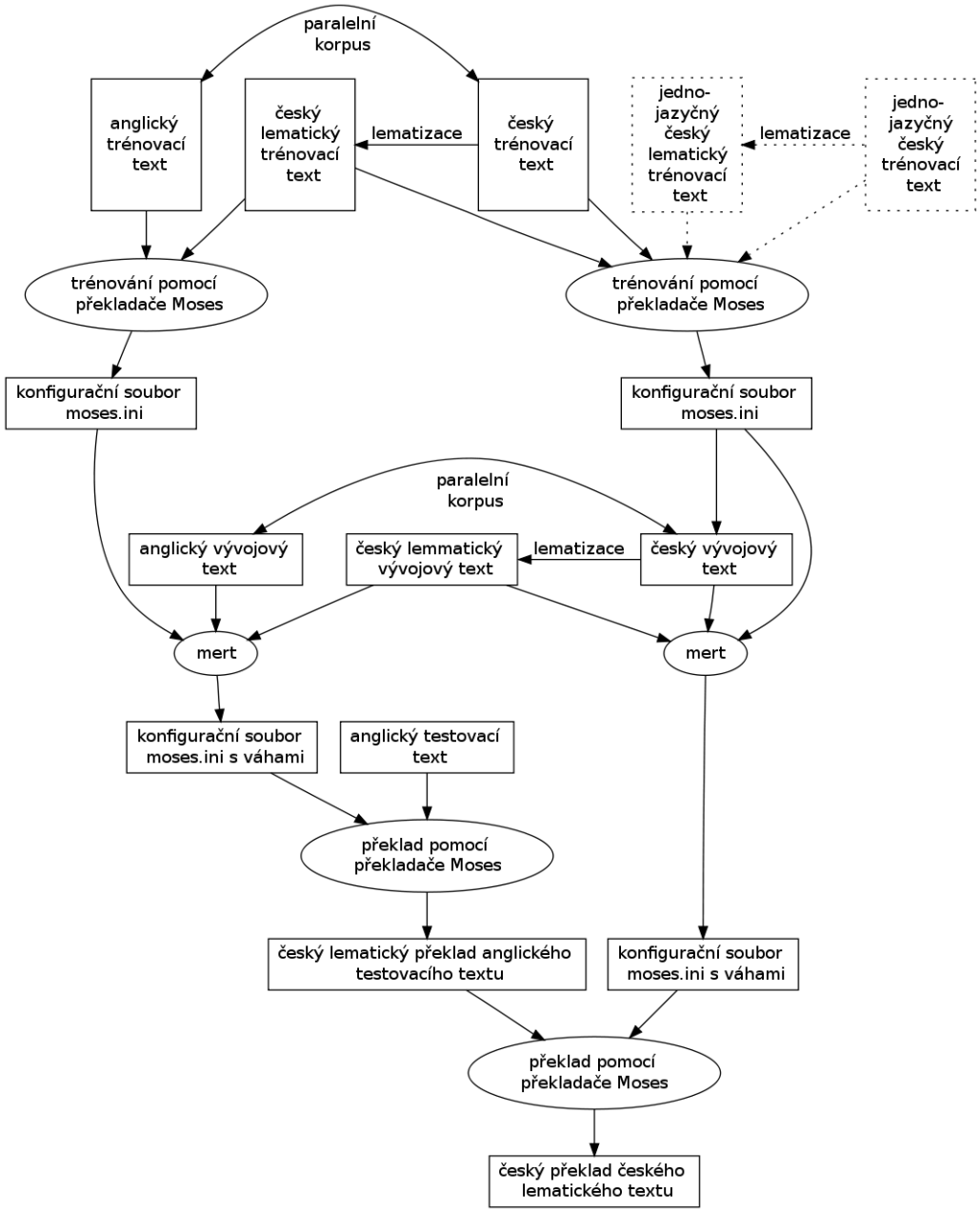
- **dvoufázový překlad**
  - **překlad pomocí prvního modelu** - překlad anglického testovacího textu pomocí prvního modelu
  - **překlad pomocí druhého modelu** - překlad výstupu prvního překladu pomocí druhého modelu do českého tvarovaného textu

- **vyhodnocení překladu** - vyhodnocení úspěšnosti překladu pomocí BLEU a NIST skóre na testovacích datech - porovnání výstupu druhého překladu s referenčním překladem

Na obrázku 2.2 je zachycena i možnost trénování druhého modelu na velkém jednojazyčném korpusu, která v této práci nebyla zkoumána.

Tento postup při základním pokusu bez užití dodatečných jednojazyčných dat nevedl ke zlepšení kvality překladu, jak je vidět v tabulce 3.2. BLEU i NIST skóre jsou nižší. Proto se práce dále více zabývala druhou navrženou metodou.

[1] lemata mohou být již k dispozici (v případě anotovaného korpusu), nebo lze text lematizovat pomocí nástroje TectoMT [10]



Obrázek 2.2: Diagram průběhu překladu z angličtiny do češtiny pomocí dvojího běhu překladače

2.3 Překlad s více překladovými tabulkami

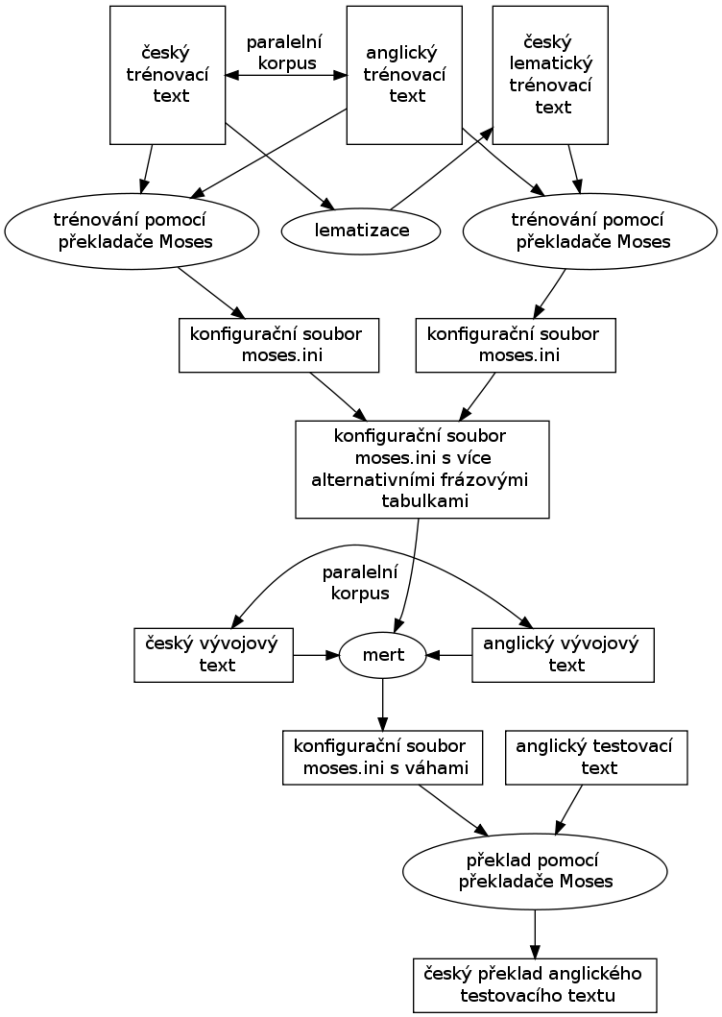
Další možností využití českého lematického textu je paralelní překlad. Vytvoří se dvě frázové překladové tabulky, jedna z angličtiny do tvarované češtiny, druhá z angličtiny do lematické češtiny. Tato metoda zachovává výhody překladu normálním postupem a navíc by mohla přeložit slova, která se v trénovacím korpusu nevyskytují příliš často.

Překlad s alternativními překladovými tabulkami je názorně zobrazen na obrázku 2.3 a má následující fáze:

- **trénování modelu angličtina - tvarovaná čeština** - trénování modelu na paralelním anglickém a českém korpusu
- **trénování modelu angličtina - lematická čeština** - trénování modelu na paralelním anglickém a českém lematickém korpusu
- **úprava konfiguračního souboru moses.ini pro překlad s více překladovými tabulkami** - konfigurační soubor `moses.ini` je potřeba upravit, aby pro překlad používal obě natrénované překladové tabulky. Způsob úpravy je popsán v následující podkapitole 2.3.1.
- **doladění vah modelu** - doladění vah v modelu pomocí metody MERT na menších datech - na vývojovém paralelním anglickém a českém korpusu
- **samotný překlad** - překlad anglického testovacího textu
- **vyhodnocení překladu** - vyhodnocení úspěšnosti překladu pomocí BLEU a NIST skóre na testovacích datech - překlad zadaného anglického testovacího textu pomocí překladače `Moses` se porovná s referenčním překladem

2.3.1 Úprava konfiguračního souboru pro překlad s více frázovými tabulkami

Do konfiguračního souboru `moses.ini` je nutné přidat druhou frázovou tabulku, prodloužit váhový vektor a určit, jakým způsobem budou frázové tabulky používány. Tabulky mohou být při překladu využívány buď *a*) všechny najednou, nebo *b*) pouze jedna z nich



Obrázek 2.3: Diagram průběhu překladu z angličtiny do češtiny pomocí alternativních překladových tabulek

Použití všech tabulek najednou Při překladu se musí překladové možnosti nacházet ve všech frázových tabulkách. Překladová možnost bude ohodnocena podle vah všech tabulek. V konfiguračním souboru se uvede následující pořadí operací:

```
[mapping]
T 0
T 1
```

T značí operaci překladu, číslo za operací určuje, jaká frázová tabulka má být k překladu použita.

Alternativní dekódovací cesty Překladové možnosti jsou vybrány z jedné tabulky a doplněny o další možnosti ze zbývajících překladových tabulek. Ohodnocení překladové možnosti se kombinuje z obou tabulek. Pořadí operací je v konfiguračním souboru následovné:

```
[mapping]
0 T 0
1 T 1
```

Číslo před operací T udává dekódovací cestu. Definují se tedy alternativní dekódovací cesty.

První možnost v této práci není vhodná, protože druhá tabulka je tvořena pouze lematickými frázemi. Při překladu by byly uvažovány pouze překladové možnosti, které se nacházejí v obou tabulkách. Tedy takové, kdy český základní tvar je identický s tvarem slova. Na výstupu překladu by byly pouze základní tvary. Vedlo by to tedy k opačnému efektu, než který potřebujeme. Alternativní dekódovací cesty naopak umožňují přímý překlad z tvarované angličtiny do tvarované češtiny doplněný o překlady do českých základních tvarů.

2.3.2 Pokusy s pozičním značkováním

V této práci byly vyzkoušeny následující konfigurace pro druhou překladovou tabulku:

1. v trénovacím korpusu s českým lematickým textem jsou pouze čistá lemata

2. v trénovacím korpusu s českým lematickým textem jsou lemata s příznakem, že jde o lemata, ale bez jakékoliv morfologické informace - tedy s prázdnou poziční značkou

3. v trénovacím korpusu s českým lematickým textem jsou lemata s částí morfologické informace

Čistý český lematický text V základní variantě je celá morfologická informace odstraněna. Z první frázové tabulky budou vybrány možnosti překladu z angličtiny do tvarované češtiny. Druhá frázová tabulka doplňuje o možnosti překladu z angličtiny do základního tvaru. Ve výstupu překladu bude většina slov z první překladové tabulky, tedy vytvarovaná slova. Pouze pokud se překladová dvojice nebude vyskytovat v první tabulce, bude vybrána ze druhé, tedy tabulky se základními tvary. Výstup překladu se dále nijak nezpracovává, lemata jsou ponechána v základním tvaru. Tato varianta se opírá a předpoklad, že uživatel na výstupu uvidí raději nevytvarované české slovo, než původní nepřeloženou angličtinu.

Český lematický text s příznaky, že se jedná o základní tvar V trénovacím korpusu se nezachovávají žádné morfologické kategorie, poziční značka s morfologickou informací je odstraněna celá. Základní tvar oproti výše popsané konfiguraci dostane příznak, že se jedná o základní tvar. Na výstupu překladu je tedy možné zjistit, že byl načten z druhé překladové tabulky. Překlad je dále zpracováván pomocí programu `zohybej` [2], který bude brát v potaz pouze slova s příznakem a pokusí se je podle kontextu vytvarovat.

Český lematický text s částí morfologické informace Lematický text, který slouží pro natrénování druhé frázové tabulky, má ponechanou část poziční značky. Při překladu přes pouhá lemata by totiž mohlo dojít ke ztrátě významu, pokud by se ztratily některé morfologické informace. U neohebných slovních druhů, předložek, spojek, částic a citoslovcí, nemá smysl, žádné morfologické informace zachovávat, protože se neohýbají. V tabulce 2.1 jsou přehledně zobrazeny kategorie, které se v této variantě překladu uvažují. Výstup překladu je dále zpracován pomocí programu `zohybej`, který uvažuje pouze ty tvary slov, které odpovídají poziční značce.

[2] program slouží k automatickému ohýbání slov a vznikl jako ročníkový projekt autora této práce

slovní druh	zachované kategorie
podstatná jména	číslo, rod
přídavná jména	číslo, negace
zájmena	-
číslovky	-
slovesa	negace, slovní poddruh (SubPOS)
příslovce	-
předložky	-
spojky	-
částice	-
citoslovce	-
interpunkce	-

Tabulka 2.1: Zachované morfologické kategorie pro jednotlivé slovní druhy

2.4 Implementace

Trénování modelu, doladění jeho vah, samotný překlad a následné vyhodnocení úspěšnosti pro jednotlivé pokusy je realizováno pomocí bashových skriptů. Ty a další pomocné skripty a programy jsou k práci přiloženy na DVD. Zde následuje seznam skriptů a popis experimentu, který realizuje:

`01normal.sh` - normální postup překladu z angličtiny do češtiny, podrobněji popsaný v podkapitole 2.1

`02twoStep.sh` - dvoufázový překlad, nejdříve je vytvořen model pro překlad z angličtiny do lematické češtiny, poté model pro překlad z lematické češtiny do tvarované češtiny, podrobněji popsáno v podkapitole 2.2

`03multipleDecodingPathNoFlagNoTag.sh` - překlad s alternativními dekódovacími cestami, kde první frázová překladová tabulka obsahuje angličtinu a tvarovanou češtinu, ve druhé tabulce jsou překlady anglických frází do českých čistých lematických frází.

`04multipleDecodingPathFlag.sh` - překlad s alternativními dekódovacími cestami (podrobněji v 2.3), kde první frázová překladová tabulka ob-

sahuje angličtinu a tvarovanou češtinu, ve druhé tabulce jsou překlady anglických frází do českých lematických frází s příznakem.

`05multipleDecodingPathFlagDifferentWeights.sh` - překlad s alternativními dekódovacími cestami, kde první frázová překladová tabulka obsahuje angličtinu a tvarovanou češtinu, ve druhé tabulce jsou překlady anglických frází do českých lematických frází s příznakem, váhy pro konfigurační soubor `moses.ini` jsou použity z pokusu č. 3

`06multipleDecodingPathFlagZohybej.sh` - překlad s alternativními dekódovacími cestami, kde první frázová překladová tabulka obsahuje angličtinu a tvarovanou češtinu, ve druhé tabulce jsou překlady anglických frází do českých lematických frází s příznakem, váhy pro konfigurační soubor `moses.ini` jsou použity z pokusu č. 3, výstup překladu je zohýbán pomocí programu `zohybej`

`07multipleDecodingPathPartTag.sh` - překlad s alternativními dekódovacími cestami, kde první frázová překladová tabulka obsahuje angličtinu a tvarovanou češtinu, ve druhé tabulce jsou překlady anglických frází do českých lematických frází s příznakem a částí morfologické značky.

`08multipleDecodingPathPartTagDifferentWeights.sh` překlad s alternativními dekódovacími cestami, kde první frázová překladová tabulka obsahuje angličtinu a tvarovanou češtinu, ve druhé tabulce jsou překlady anglických frází do českých lematických frází s příznakem a částí morfologické značky, váhy pro konfigurační soubor `moses.ini` jsou použity z pokusu č. 3

`09multipleDecodingPathPartTagZohybej.sh` - překlad s alternativními dekódovacími cestami, kde první frázová překladová tabulka obsahuje angličtinu a tvarovanou češtinu, ve druhé tabulce jsou překlady anglických frází do českých lematických frází s příznakem a částí morfologické značky, váhy pro konfigurační soubor `moses.ini` jsou použity z pokusu č. 3, výstup překladu je zohýbán pomocí programu `zohybej`

Způsob spuštění a význam jednotlivých vstupních parametrů jsou vysvětleny v příloze A.2.

Kapitola 3

Vyhodnocení úspěšnosti navržených metod

V této kapitole jsou uvedeny důležité statistiky pro použitá data, výsledky automatického vyhodnocení jednotlivých metod překladu a ruční porovnání malého vzorku překladů.

3.1 Použitá data

Pro práci byla použita data z **CzEng 0.9** [11]. CzEng obsahuje automaticky označkovaná data z několika různých typů zdrojů (filmové titulky, legislativa EU, technická dokumentace, beletrie, paralelní webové stránky, noviny, Project Navajo).

Jako trénovací data byla použita sekce 00 a 01, vývojová data ze sekce 80 (prvních 2000 vět) a testovací data ze sekce 90 (prvních 1000 vět). Trénovací data pro program `zohybej` jsou ze sekce 00, 01, 02, 10, 11 ... 19.

Velikosti dat, na kterých byly prováděny pokusy, jsou zachyceny v tabulce 3.1.

3.2 Vyhodnocení úspěšnosti překladu pomocí dvojího běhu překladače

Následující podkapitoly obsahují vyhodnocení úspěšnosti překladu pomocí dvojího běhu překladače.

	# vět	# tokenů
anglický trénovací	160,9k	1,8M
český trénovací	160,9k	1,6M
anglický vývojový	2k	23,3k
český vývojový	2k	20,9k
anglický testovací	1k	11,3k
český testovací	1k	9,7k
zohybej trénovací	1M	10,7M

Tabulka 3.1: Velikosti použitých trénovacích, vývojových a testovacích dat

3.2.1 Výsledky automatického vyhodnocení

V tabulce 3.2 jsou zachyceny výsledky automatického vyhodnocení této metody v porovnání s normálním překladem. Jak bylo již dříve zmíněno, tento postup nebyl pro pokles skóre dále rozpracován.

	# experimentu	NIST skóre	BLEU skóre
normální překlad	1	4.3449	0.1724
dvojitý překlad	2	4.1346	0.1663

Tabulka 3.2: Porovnání výsledků překladu normálním způsobem a překladu pomocí dvojího běhu překladače přes lematický text

3.2.2 Příklady výstupu překladu

V tabulce 3.3 následují okomentované příklady překladu, které demonstrují nedostatky této metody. Nejdřívě je uveden anglický originál (**src**), dále referenční český překlad (**ref**), výstup překladu normálním postupem (**nor**), poté výstup překladu pomocí dvojího běhu překladače (**two**). Nakonec následuje komentář (**pozn**).

src	I knew it .
ref	Já to věděla .
nor	Já to věděl .
two	Já to vím .
pozn	*při dvojím překladu došlo ke ztrátě informace o čase slovesa, výstupní věta je správně zohýbaná podle kontextu, ale v nesprávném čase*
src	He ca n't go far .
ref	Nebude daleko
nor	Nemůže jít daleko .
two	Mohu jít dál .
pozn	*při překladu došlo ke ztrátě negace a osoby*

Tabulka 3.3: Příklady výstupu překladu normálním postupem překladu a překladu pomocí dvojího běhu překladače

Tato metoda nezohledňuje morfologické informace z českého textu. Provádí plnou lematizaci a tedy nezachovává např. osobu, číslo, čas nebo negaci. Jako možné vylepšení by se mohla u českého lematického textu zachovávat částečná morfologická informace.

3.3 Vyhodnocení úspěšnosti překladu pomocí alternativních dekódovacích cest

V následujících podkapitolách je uvedeno vyhodnocení úspěšnosti překladu pomocí alternativních dekódovacích cest.

3.3.1 Výsledky automatického vyhodnocení

V tabulce 3.4 jsou zachyceny výsledky různých variant této metody v porovnání s normálním překladem.

Překlad pomocí alternativních dekódovacích cest je již v základním pokusu nepatrně lepší než normální postup překladu. BLEU skóre se zvyšuje,

	# experimentu	NIST skóre	BLEU skóre
normální překlad	1	4.3449	0.1724
lemata bez příznaku	3	4.3439	0.1738
lemata s příznakem	4	4.4071	0.1728
lemata s příznakem, váhy bez příznaku	5	4.3627	0.1750
lemata s příznakem, následné zohýbání	6	4.3601	0.1750
lemata s částí značky	7	4.3812	0.1692
lemata s částí značky, váhy bez příznaku	8	4.3610	0.1743
lemata s částí značky, následné zohýbání	9	4.3601	0.1743

Tabulka 3.4: Porovnání výsledků překladu normálním způsobem a překladu pomocí alternativních překladových tabulek

ale NIST skóre se snižuje. V dalších pokusech se experimentuje s lematizací druhé frázové tabulky - k lematům je přidán příznak nebo část morfologické informace. Doladění vah pomocí metody MERT probíhá na anglických a českých tvarovaných datech, proto je skóre v překladu s použitím lemat s příznakem a lemat s částí morfologické značky nižší než v základním pokusu. Použijí-li se váhy z pokusu č. 3 (alternativní frázová tabulka pouze s čistými lematy), skóre v pokusu č. 5 a č. 8 výrazně stoupne. Překlad pomocí alternativní lematické frázové tabulky umožní navíc překlad málo častých slov nebo slov, která v trénovacích datech nebyla v podobném spojení použita. Dochází k nárůstu počtu překladových možností, tedy někdy dojde k špatnému přeuspořádání věty či zkrácení překladu oproti normálnímu postupu překladu. Nejlepších výsledků bylo dosaženo v pokusu č. 5.

3.3.2 Výsledky ručního porovnání malého vzorku překladů

Pro ověření výsledků automatického hodnocení došlo k lidskému posouzení překladu na malém vzorku 60 vět. Aby bylo ověření nezávislé, bylo z výstupů překladů pomocí některých metod vybráno náhodně 60 vět. Pro posouzení

byla hodnotiteli předložena originální anglická věta, její referenční překlad a překlady v náhodném pořadí, které bylo vytvořeno pomocí nástroje `quick-judge`. [1] Zde následují výsledky ručního porovnání několika metod.

Jak je vidět v tabulce 3.5 překlad přes lematický text dává pro uživatele lepší výsledky. Takový výsledek potvrzuje očekávání, že uživatel na výstupu raději vidí přeložená slova (třeba i netvarovaná) než původní anglická slova. V porovnávání nebyly uvažovány identické překlady. Z výsledků je vidět, že dodatečné zohýbání kvalitu překladu dále zvyšuje.

	výrazně lepší	**lepší**	**stejně dobré**	**špatné**	**horší**	**výrazně horší**
čistá lemata	2	22	22	8	5	1
zohýbaná lemata s příznakem	2	26	22	4	6	0

Tabulka 3.5: Výsledky ručního porovnání malého vzorku výstupů normálního postupu překladu a překladu přes lematický text pomocí alternativních dekódovacích cest

3.3.3 Příklady výstupu překladu

V tabulce 3.6 následují okomentované příklady překladu, které demonstrují vylepšení a nedostatky překladu přes základní tvary. Nejdříve je uveden anglický originál (**src**), dále referenční český překlad (**ref**), výstup překladu normálním postupem (**nor**), výstup překladu s využitím alternativních dekódovacích cest, kde druhá překladová tabulka je natrénována na čistém lematickém textu (**mult**), poté výstup překladu, kde druhá překladová tabulka je natrénována na lematickém textu s příznakem a následně zohýbána (**tag-zo**). Nakonec následuje komentář (**pozn**).

Tabulka 3.6: Příklady výstupu překladu normálním postupem překladu a překladu pomocí alternativních dekódovacích cest - ve variantě čistý lematický text a zohýbaný lematický text s příznakem - viz následující strany

[1] `http://ufal.mff.cuni.cz/euromatrix/quickjudge/`

src	He 's like fine shrimp .
ref	Má žaludek na vodě .
nor	Je fajn shrimp .
alt	Je v pořádku . Krevetka Severní
tag-zo	Je v pořádku . Shrimp
pozn	*shrimp je správně navíc přeloženo jako kreveta, ale v kontextu překlad nedává smysl*
src	This decision is not to apply to liechtenstein ,
ref	Toto rozhodnutí by se nemělo vztahovat na Lichtenštejnsko ,
nor	Toto rozhodnutí není chcete použít pro Lichtenštejnsko ,
alt	Toto rozhodnutí se použije na Lichtenštejnsko ,
tag-zo	Toto rozhodnutí se použije na Lichtenštejnsko ,
pozn	*překladem přes lemata došlo ke ztrátě negace*
src	Measures decided upon by the Commission may be referred to the Council by any Member State within three working days of the day on which they were communicated .
ref	Opatření , o kterém rozhodne Komise může kterýkoli členský stát do tří dnů ode dne jeho sdělení předložit Radě k přezkoumání .
nor	Opatření rozhodnout Komise může být uvedené Rada členský stát do tří pracovních dnů od dne přenést .
alt	Opatření rozhodnout Komise může být uvedené Rada členský stát do tří pracovních dnů od dne . Sděleny
tag-zo	Na opatření Komise může uvedené Rada členský stát do tří pracovních dnů od dne . Sděleny
pozn	*they were communicated je v normálním překladu přeloženo jako přenést, v alternativním jako sděleny, druhý překlad dává větší smysl, ale je horší větný pořádek*

src	It was drowned out in a burst of drunken laughter , then one of the men stiffened .
ref	Zbytek věty splynul s výbuchem opileckého smíchu , a pak jeden z mužů zpozorněl .
nor	Bylo utopená v výtrysk bolesti opilý smích , potom jeden z mužů stiffened .
alt	Bylo utopená v výtrysk bolesti opilý smích , pak jeden z mužů ztuhnout .
tag-zo	Bylo utopená v výtrysk bolesti opilý smích , pak jeden z mužů ztuhl .
pozn	*anglické slovo stiffened bylo správně přeloženo jako ztuhnout, i když je v základním tvaru, uživatel raději uvidí jeho překlad do češtiny, aby alespoň tušil význam,* `zohybej` *poté správně zohýbalo jako ztuhl*
src	How does it feel to be my personal property ?
ref	Jaké to je , být můj osobní majetek ?
nor	Jaké je to být moje osobní majetek .
alt	Jaké je to být moje osobní vlastnost ?
tag-zo	Jaké je to být moje osobní vlastnost ?
pozn	*property bylo správně přeloženo jako vlastnost, ale v kontextu dává větší smysl překlad jako majetek*
src	And I mean that literally .
ref	Nemusíš doslovně !
nor	A myslím , že literally .
alt	A myslím , že doslova .
tag-zo	A myslím , že doslova .
pozn	*literally je správně přeloženo jako doslova, v referenčním překladu je chyba*
src	Show Time
ref	Zobrazit čas
nor	Jsem nahej
alt	Zobrazit čas
tag-zo	Zobrazit čas
pozn	*při překladu normální metodou dojde k nesprávnému překladu, v jiném kontextu by tento překlad byl přijatelný*

Při překladu věty **it was drowned out in a burst of drunken laughter , then one of the men stiffened** dojde navíc při použití alternativních

dekódovacích cest k správnému přeložení slova **stiffened**. To je umožněno různým obsahem frázových tabulek. V příkladu 3.1 a 3.2 je uvedena část frázových tabulek s tímto anglickým slovem. Frázová tabulka natrénovaná na tvarované češtině neobsahuje frázi **stiffened** → **ztuhl**, proto při překladu nebude tato překladová možnost načtena a slovo se jako neznámé nepřeloží.

```
' suddenly he stiffened , and | " náhle strnul , shýbl se a
he stiffened , and | strnul , shýbl se a
he stiffened as | ztuhl , jako
he stiffened | ztuhl ,
he stiffened | ztuhl
stiffened , and | strnul , shýbl se a
suddenly he stiffened , and | náhle strnul , shýbl se a
then he stiffened as | pak ztuhl , jako
then he stiffened | pak ztuhl ,
then he stiffened | pak ztuhl
```

Příklad 3.1: Obsah frázové tabulky angličtina - čeština

```
he stiffened as if | ztuhnout , jako být on
stiffened , and stooping | strnout , shýbnout se
stiffened , and stooping | strnout , shýbnout
stiffened as if ||| ztuhnout , jako být
stiffened | ztuhnout ,
stiffened | ztuhnout
suddenly he stiffened , and stooping | náhle strnout , shýbnout
then he stiffened as if ||| pak ztuhnout , jako být on
```

Příklad 3.2: Obsah frázové tabulky angličtina - lematická čeština

Kapitola 4

Závěr

Tato bakalářská práce se zabývala možností zlepšení kvality statistického frázového překladu z angličtiny do češtiny využitím lematizace české části paralelního korpusu. V práci byly vyzkoušeny dva různé přístupy.

První přístup pomocí dvojího běhu překladače nevedl k lepším výsledkům. Při automatickém ohodnocení dával výrazně nižší skóre.

Proto se práce více věnovala druhému přístupu - překladu pomocí alternativních dekódovacích cest, kde první frázová tabulka byla z angličtiny do tvarované češtiny a druhá frázová tabulka z angličtiny do lematické češtiny. Tato metoda již při základním experimentu dávala nepatrné zvýšení BLEU skóre. Zvýšení kvality překladu se potvrdilo i ručním ohodnocením malého vzorku překladů. Při použití lematické češtiny jako alternativní dekódovací cesty jsou navíc přeložena i slova, která se v anglickém trénovacím textu příliš často nevyskytují. Automatické i ruční ohodnocení potvrdilo předpoklad, že uživatel na výstupu překladu uvidí raději nezohýbané české slovo než nepřeloženou angličtinu.

Bylo vyzkoušeno několik variant lematizace druhé frázové tabulky. V základní variantě byla v druhé frázové tabulce pouze čistá lemata, v další variantě lemata s příznakem a poté lemata s částí morfologické značky. Nejlepších výsledků při automatickém ohodnocení bylo dosaženo ve druhé variantě při použití vah z první varianty. To bylo ověřeno i ručním ohodnocením.

Bakalářská práce ověřila, že použití lematizace českého textu při překladu z angličtiny do češtiny pomáhá odstranit problém bohaté české morfologie.

Práci by bylo možné dále rozšířit např. o vyhodnocení metod na větší množině trénovacích dat, zapojit jednojazyčná trénovací data do překladu pomocí dvojího běhu překladače nebo místo povrchové morfologie použít

část rysů z hloubkové syntaxe.

Literatura

[1] Koehn P., Hoang H., Birch A., Callison-Burch C., Federico M., Bertoldi N., Cowan B., Shen W., Moran C., Zens R., Dyer C., Bojar O., Constantin A., Herbst E.: *Moses: Open Source Toolkit for Statistical Machine Translation*, Annual Meeting of the Association for Computational Linguistics (ACL), Praha, 2007

[2] Bojar 0., Hajič J.: *Phrase-Based and Deep Syntactic English-to-Czech Statistical Machine Translation*, Proceedings of the Third Workshop on Statistical Machine Translation, Columbus, Ohio, June 2008

[3] Koehn P.: *Europarl: A Parallel Corpus for Statistical Machine Translation*, Proceedings of MT Summit X, 2005

[4] Callison-Burch C., Koehn P., Monz C., Schroeder J.: *Findings of the 2009 Workshop on Statistical Machine Translation*, Proceedings of the Fourth Workshop on Statistical Machine Translation, Athens, Greece, 2009

[5] Hajič J.: *Disambiguation of Rich Inflection - Computational Morphology of Czech*, Charles University Press - Karolinum, 2004

[6] Koehn P.: *Statistical Machine Translation*, Cambridge University Press, 2010

[7] Bertoldi N., Haddow B., Fouet J.-B.: *Improved Minimum Error Rate Training in Moses*, Proceedings of 3rd MT Marathon, Prague, Czech Republic, 2009

[8] Papineni K., Roukos S., Ward T., Zhu W.: *Bleu: a Method for Automatic Evaluation of Machine Translation*, Computer Science, 2001

[9] Doddington G.: *Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics*, Proceedings of the second international conference on Human Language Technology Research, San Diego, California, 2002

[10] Žabokrtský Z., Bojar O.: *TectoMT, Developer's Guide*, ÚFAL/CKL Technical Report TR-2008-38

[11] Bojar 0., Žabokrtský Z.: *CzEng 0.9: Large Parallel Treebank with Rich Annotation*, Prague Bulletin of Mathematical Linguistics, **92**, 2009.

Dodatek A

Přílohy

Následující kapitola obsahuje popis příloh k této bakalářské práci.

A.1 Přiložené DVD

K bakalářské práci je přiloženo DVD, které obsahuje text této práce, použitá data, výstupy překladu, použité skripty a program `zohybej`.

A.2 Uživatelská dokumentace k použitým programům a skriptům

V následující podkapitole je popsán způsob spuštění skriptů a programů, které byly v této práci použity.

A.2.1 Skripty s experimenty

```
01normal.sh -t configFile train dataDir workDir
01normal.sh -u configFile tune dataDir workDir
01normal.sh -e configFile eval train dataDir workDir
```

- `-t` přepínač pro trénování modelu
- `-u` přepínač pro doladění vah modelu
- `-e` přepínač pro překlad pomocí natrénovaného modelu a vyhodnocení úspěšnosti překladu

- `configFile` soubor s konfigurací pro překladač `Moses` - obsahuje cestu k adresáři s překladačem `Moses` a pomocným skriptům
- `dataDir` adresář s daty - trénovacím a vývojovým korpusem a testovacími daty
- `workDir` pracovní adresář pro experiment
- `train` název trénovacího paralelního korpusu
- `tune` název vývojového paralelního korpusu
- `eval` název testovacích dat

```
02twoStep.sh -t configFile train dataDir workDir
02twoStep.sh -u configFile tune dataDir workDir
02twoStep.sh -e configFile eval train dataDir workDir
```

- `-t` přepínač pro trénování modelu
- `-u` přepínač pro doladění vah modelu
- `-e` přepínač pro překlad pomocí natrénovaného modelu a vyhodnocení úspěšnosti překladu
- `configFile` soubor s konfigurací pro překladač `Moses` - obsahuje cestu k adresáři s překladačem `Moses` a pomocným skriptům
- `dataDir` adresář s daty - trénovacím a vývojovým korpusem a testovacími daty
- `workDir` pracovní adresář pro experiment
- `train` název trénovacího paralelního korpusu
- `tune` název vývojového paralelního korpusu
- `eval` název testovacích dat

```
03multipleDecodingPath-onlyLemma.sh -t configFile train
    dataDir workDir
03multipleDecodingPath-onlyLemma.sh -u configFile tune
    dataDir workDir
03multipleDecodingPath-onlyLemma.sh -e configFile eval train
    dataDir workDir
```

- `-t` přepínač pro trénování modelu
- `-u` přepínač pro doladění vah modelu
- `-e` přepínač pro překlad pomocí natrénovaného modelu a vyhodnocení úspěšnosti překladu
- `configFile` soubor s konfigurací pro překladač `Moses` - obsahuje cestu k adresáři s překladačem `Moses` a pomocným skriptům
- `dataDir` adresář s daty - trénovacím a vývojovým korpusem a testovacími daty
- `workDir` pracovní adresář pro experiment
- `train` název trénovacího paralelního korpusu
- `tune` název vývojového paralelního korpusu
- `eval` název testovacích dat

```
04multipleDecodingPath-flag.sh -t configFile train dataDir
    workDir
04multipleDecodingPath-flag.sh -u configFile tune dataDir
    workDir
04multipleDecodingPath-flag.sh -e configFile eval train
    dataDir workDir
```

- `-t` přepínač pro trénování modelu
- `-u` přepínač pro doladění vah modelu
- `-e` přepínač pro překlad pomocí natrénovaného modelu a vyhodnocení úspěšnosti překladu

- `configFile` soubor s konfigurací pro překladač `Moses` - obsahuje cestu k adresáři s překladačem `Moses` a pomocným skriptům
- `dataDir` adresář s daty - trénovacím a vývojovým korpusem a testovacími daty
- `workDir` pracovní adresář pro experiment
- `train` název trénovacího paralelního korpusu
- `tune` název vývojového paralelního korpusu
- `eval` název testovacích dat

```
05multipleDecodingPath-flag-weights.sh -t configFile train
   dataDir workDir
05multipleDecodingPath-flag-weights.sh -u configFile workDir
   workDir3 workDir4
05multipleDecodingPath-flag-weights.sh -e configFile eval
   train dataDir workDir
```

- `-t` přepínač pro trénování modelu
- `-u` přepínač pro doladění vah modelu
- `-e` přepínač pro překlad pomocí natrénovaného modelu a vyhodnocení úspěšnosti překladu
- `configFile` soubor s konfigurací pro překladač `Moses` - obsahuje cestu k adresáři s překladačem `Moses` a pomocným skriptům
- `dataDir` adresář s daty - trénovacím a vývojovým korpusem a testovacími daty
- `workDir` pracovní adresář pro tento experiment
- `workDir3` pracovní adresář s experimentem č. 3
- `workDir4` pracovní adresář s experimentem č. 4
- `train` název trénovacího paralelního korpusu
- `eval` název testovacích dat

```
06multipleDecodingPath-flag-zohybej.sh -e configFile eval
   dataDir workDir work5 zohybej zohybejModel
```

- `-e` přepínač pro překlad pomocí natrénovaného modelu a vyhodnocení úspěšnosti překladu
- `configFile` soubor s konfigurací pro překladač `Moses` - obsahuje cestu k adresáři s překladačem `Moses` a pomocným skriptům
- `dataDir` adresář s daty - trénovacím a vývojovým korpusem a testovacími daty
- `workDir` pracovní adresář pro tento experiment
- `workDir5` pracovní adresář s experimentem č. 5
- `zohybej` adresář s programem `zohybej`
- `zohybejModel` natrénovaný model pro program `zohybej`
- `eval` název testovacích dat

```
07multipleDecodingPath-tag.sh -t configFile train dataDir
   workDir
07multipleDecodingPath-tag.sh -u configFile tune dataDir
   workDir
07multipleDecodingPath-tag.sh -e configFile eval train
   dataDir workDir
```

- `-t` přepínač pro trénování modelu
- `-u` přepínač pro doladění vah modelu
- `-e` přepínač pro překlad pomocí natrénovaného modelu a vyhodnocení úspěšnosti překladu
- `configFile` soubor s konfigurací pro překladač `Moses` - obsahuje cestu k adresáři s překladačem `Moses` a pomocným skriptům
- `dataDir` adresář s daty - trénovacím a vývojovým korpusem a testovacími daty

- `workDir` pracovní adresář pro experiment
- `train` název trénovacího paralelního korpusu
- `tune` název vývojového paralelního korpusu
- `eval` název testovacích dat

```
08multipleDecodingPath-tag-weights.sh -t configFile train
   dataDir workDir
08multipleDecodingPath-tag-weights.sh -u configFile workDir
   workDir3 workDir7
08multipleDecodingPath-tag-weights.sh -e configFile eval
   train dataDir workDir
```

- `-t` přepínač pro trénování modelu
- `-u` přepínač pro doladění vah modelu
- `-e` přepínač pro překlad pomocí natrénovaného modelu a vyhodnocení úspěšnosti překladu
- `configFile` soubor s konfigurací pro překladač `Moses` - obsahuje cestu k adresáři s překladačem `Moses` a pomocným skriptům
- `dataDir` adresář s daty - trénovacím a vývojovým korpusem a testovacími daty
- `workDir` pracovní adresář pro tento experiment
- `workDir3` pracovní adresář s experimentem č. 3
- `workDir7` pracovní adresář s experimentem č. 7
- `train` název trénovacího paralelního korpusu
- `eval` název testovacích dat

```
09multipleDecodingPath-tag-zohybej.sh -e configFile eval
   dataDir workDir workDir8 zohybej
zohybejModel
```

- `-e` přepínač pro překlad pomocí natrénovaného modelu a vyhodnocení úspěšnosti překladu
- `configFile` soubor s konfigurací pro překladač `Moses` - obsahuje cestu k adresáři s překladačem `Moses` a pomocným skriptům
- `dataDir` adresář s daty - trénovacím a vývojovým korpusem a testovacími daty
- `workDir` pracovní adresář pro tento experiment
- `workDir8` pracovní adresář s experimentem č. 8
- `zohybej` adresář s programem `zohybej`
- `zohybejModel` natrénovaný model pro program `zohybej`
- `eval` název testovacích dat

A.2.2 Program zohybej

Program `zohybej` lze spustit z příkazové řádky ve dvou režimech, trénovacím a vyhodnocovacím. Trénovací režim se zapíná pomocí přepínače `-t` nebo `--train` a vyhodnocovací pomocí přepínače `-e` nebo `--eval`.

Trénovací režim Parametry pro spuštění v trénovacím režimu jsou následující:

```
zohybej -t -n N -c corpusFileName -m modelName
zohybej --train --ngram N --corpus corpusFileName --model
   modelName
```

- `-t` nebo `--train` přepínač do trénovacího režimu
- `-n N` nebo `--ngram N` velikost N-gramů, se kterými bude program pracovat
- `-c corpusFileName` nebo `--corpus corpusFileName` název souboru s korpusem pro natrénování

- `-m modelName` nebo `--model modleName` název souborů s natrénovaným modelem

Všechny tyto parametry trénovacího režimu jsou povinné. Vstupní trénovací data jsou zohýbané věty. Věty musí být na samostatných řádcích a jednotlivá slova jsou oddělena mezerami.

Vyhodnocovací režim Parametry pro spuštění ve vyhodnocovacím režimu jsou následující:

```
zohybej -e -n N -d dictionaryFileName -p dictionaryFilePath -
    m modelName lemma [lemma] [-o
outputFileName]
zohybej --eval --ngram N --dictionary dictionaryFileName --
    path dictionaryFilePath --model
modelName lemma [lemma] [--output outputFileName]
```

- `-e` nebo `--eval` přepínač do vyhodnocovacího režimu
- `-n N` nebo `--ngram N` velikost N-gramů, se kterými bude program pracovat
- `-d dictionaryFileName` nebo `--dictionary dictionaryFileName` název souboru se slovníkem pro `Free Morphology`
- `-p dictionaryFilePath` nebo `--path dictionaryFilePath` cesta k adresáři se slovníkem pro `Free Morphology`
- `-m modelName` nebo `--model modelName` název natrénovaného modelu
- `lemma [lemma]` lematická věta, která má být zohýbána
- `-o outputFileName` nebo `--output outputFileName` název výstupního souboru, kam bude zapsána zohýbaná věta

```
zohybej -e -n N -d dictionaryFileName -p dictionaryFilePath -
    m modelName -i inputFileName [-o
outputFileName]
zohybej --eval --ngram N --dictionary dictionaryFileName --
    path dictionaryFilePath --model --input
inputFileName [--output outputFileName]
```

- `-e` nebo `--eval` přepínač do vyhodnocovacího režimu
- `-n N` nebo `--ngram N` velikost N-gramů, se kterými bude program pracovat
- `-d dictionaryFileName` nebo `--dictionary dictionaryFileName` název souboru se slovníkem pro `Free Morphology`
- `-p dictionaryFilePath` nebo `--path dictionaryFilePath` cesta k adresáři se slovníkem pro `Free Morphology`
- `-m modelName` nebo `--model modelName` název natrénovaného modelu
- `-i inputFileName` nebo `--input inputFileName` názve vstupního souboru s lematickým textem
- `-o outputFileName` nebo `--output outputFileName` název výstupního souboru, kam bude zapsána zohýbaná věta

Všechny parametry, kromě parametru `-o`, jsou povinné. Výstup programu je vypsán na standardní výstup a, pokud je zadán parametr `-o` do výstupního souboru. Vstupem programu pro vyhodnocovací režim je lematický text. Může být zadán přímo z příkazové řádky, nebo načten ze souboru, kde jednotlivé věty musí být na samostatných řádcích a jednotlivá slova oddělena mezerami. Věta má například následující podobu:

```
na/lemma stůl/lemma/* je kniha/lemma/NNNS1*
```

Jednotlivé možnosti, jak zadat slova, jsou:

- `na/lemma` - slovo v základním tvaru bez poziční značky - odpovídá všem tvarům
- `stůl/lemma/*` - slovo v základním tvaru s poziční značkou, která odpovídá všem tvarům
- `je` - slovo již na vstupu správně zohýbané
- `kniha/lemma/NNNS1*` - slovo v základním tvaru s poziční značkou - program uvažuje pouze tvary, které odpovídají této značce

Program využívá pro zjištění všech možných tvarů k lematům program `Free Morphology` [1], konkrétně perlovský skript `FMGenerate.pl`. Uživatel tedy musí mít pro správné používání programu nainstalován Perl.

Implementace programu je podrobněji rozebrána v dokumentu `documentation.pdf`, který je k této práci přiložen na DVD.

[1] `http://ufal.mff.cuni.cz/pdt/Morphology_and_Tagging/Morphology/`