VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÝCH SYSTÉMŮ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER SYSTEMS

EXPERIMENTÁLNÍ PŘEKLADAČ Z ČESTINY DO SLOVENŠTINY

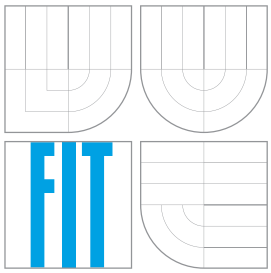
BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE
AUTHOR

PETER KADLEC

BRNO 2015

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÝCH SYSTÉMŮ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER SYSTEMS

EXPERIMENTÁLNÍ PŘEKLADAČ Z ČESTINY DO SLOVENŠTINY

CZECH-SLOVAK MACHINE TRANSLATION

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE
AUTHOR

PETER KADLEC

VEDOUCÍ PRÁCE
SUPERVISOR

Doc. RNDr. PAVEL SMRŽ, Ph.D.

BRNO 2015

Abstrakt

Cílem této bakalářské práce bylo seznámit se s metodami používanými pro automatický strojový překlad, navrhnout a implementovat systém pro překlad českých textů do slovenštiny a na závěr vyhodnotit úspěšnost vytvořeného systému pomocí standardních metrik.

Abstract

Aim of this bachelor thesis was to get familiar with methods used in automatic machine translation, design and implement system for translation from czech to slovak and in the end with help of standard metrics score the created system.

Klíčová slova

štatistický strojový překlad, korpus, zarovnání, moses.

Keywords

statistical machine translation, corpora, alignment, moses.

Citace

Peter Kadlec: Experimentální překladač z čestiny do slovenštiny, bakalářská práce, Brno, FIT VUT v Brně, 2015

Experimentální překladač z čestiny do slovenštiny

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně pod vedením pana Doc. RNDr. Pavel Smrž, Ph.D.

.....
Peter Kadlec
12. května 2015

Poděkování

Touto cestou by som sa chcel poďakovať pánovi docentovi Smržovi za odbornú pomoc pri tvorbe tohoto projektu a Slovenskému národnému korpusu za poskytnutie beletristického paralelného korpusu.

Obsah

Kapitola 1

Úvod

V dnešnej dobe globálnej informačnej siete, v ktorej sú dáta na dosah ruky, narastá problém ako dané dáta interpretovať na užitočné informácie. A teraz nehovorím o zle použitej kódovacej sade, ale o viac pragmatickejšom probléme, ktorým sú ľudské jazyky.

Problematika porozumenia cudziemu jazyku siaha do dávnych čias prvých civilizácií. Porozumenie, podanie a prijatie informácií je kľúčové pre rozvoj, obchod a vytváranie väzieb s okolitým svetom. A v dnešnej dobe globálnych dát avšak nie globálneho jazyka je táto téma viac než aktuálna.

Ako však zabezpečiť aby bol text zrozumiteľný pre každého keď na svete existuje viac než 6000 jazykov ? Ideálne by bolo keby všetci vedeli používať jednotný jazyk, avšak táto myšlienka je pre najbližšie storočia prinajmenšom utopistická. Áno je pravda že anglický jazyk sa v dnešnej dobe považuje za svetový no rozhodne ho neovláda každý. Preto neostáva iné východisko ako text prekladať.

Prekladanie cudzích textov je často jediný spôsob ako sprístupniť informácie širokej verejnosti. Najčastejšou metódou je ručný preklad. Jedná sa o najstaršiu a najkvalitnejšiu metódu prekladu. Avšak je aj veľmi pomalá a drahá. V dnešnej dobe rapídnej produkcie textov sú tieto vlastnosti dosť obmedzujúce. Čo z toho že text je prvotriedne preložený, keď jeho aktuálnosť je často nízka aj napriek nemalým nákladom na jeho preklad. Automatizácia sa v tomto ohľade doslova ponúka. Rýchly preklad a nízka cena sú len niektoré z výhod, ktoré strojový preklad ponúka.

Ak sa bavíme o preklade medzi dvoma príbuznými jazykmi akými sú slovenčina a čeština, možno nás napadne otázka, prečo vlastne prekladať texty medzi nimi, keď sú si natoľko podobné, že ľudia týchto dvoch jazykových skupín si rozumejú aj bez potreby sa druhý jazyk učiť ? Dôvodov je viacero. Áno, je pravda, že česko a slovensko boli kedysi jedna krajina, avšak od ich rozdelenia už uplynulo viac ako 20 rokov. Prevažná väčšina obyvateľstva sa stále dokáže dohovoriť, no v oddelených krajinách v ktorých sú aj média výrazne separované, mladšia generácia sa nenachádza v tak jazykovo premiešanom prostredí ako tá pred nimi. Je možné, že tento trend bude pokračovať a zo striedaním generácii sa budú vzďaľovať aj jazyky. V každom prípade už aj v dnešnej dobe porozumenie nie je na sto percentnej úrovni a v oblasti ako je zákonodarstvo, politika a podobných oblastiach, v ktorých je potrebné plné porozumenie, sa nemôžeme spoliehať na príbuznosť jazykov. Existujú mnohé texty, ktoré musia byť zo zákona preložené do rodného jazyka danej jazykovej skupiny. A tak sa znovu dostávame do problematiky prekladania.

Je nutné podotknúť, že nakoľko je strojový preklad rýchlou a lacnou metódou prekladu, v súčasnej dobe stále zaostáva kvalitou za ručným prekladom. Preto nemôžeme zatiaľ brať strojový preklad ako ultimátny nástroj na preklad. Avšak je vhodným nástrojom na va-

lidáciu ľudského prekladu a jeho medziprodukty môžu byť použité napríklad pri tvorbe slovníkov a na ich obohacovanie a kontrolu.

V tomto projekte sú popísané rôzne prístupy k tvorbe prekladových systémov, ich závislosti na vstupných dátach, výhody a nevýhody. Je tu popísaný kompletný postup pri tvorbe štatistického prekladového systému, nástroje použité pri jeho tvorbe, postup úpravy vstupných dát ako aj samotný proces prekladu a použitie štandardným metrík na ohodnotenie jeho kvality. Ku koncu sú popísané mnou vytvorené štatistické prekladové systémy pre pár čeština-slovenčina, je zhodnotená ich kvalita prekladu na testovacom texte a porovnaná s prekladom už existujúcich prekladových systémov ako google translate.

Kapitola 2

Analýza témy

Tato kapitola vysvetľuje pojem strojového prekladu, rôzne postupy pri tvorbe prekladových systémov, čo sú to paralelné korpusy, ako a kde ich získať a aké nástoje sú k dispozícii na ich úpravu do vhodného formátu, pre použitie v tvorbe štatistických prekladových systémov (SMT).

2.1 Strojový preklad

Strojový preklad tiež známy pod skratkou MT z anglického Machine translation, tvorí podskupinu tzv. výpočtovej lingvistiky, ktorá skúma využitie počítačových programov pri preklade hovoreného, alebo písaného slova z jedného prirodzeného jazyka do druhého. Na základnej úrovni MT predstavuje len jednoduché nahrádzanie slov jedného jazyka, za slová s rovnakým významom z jazyka druhého. Preklad je možné vylepšiť tzv. korpusovými metódami. Tieto metódy napomáhajú pri identifikácií fráz, morfológie, idiómov či rôznych anomálií jazyka, a napomáhajú tak výraznejšie zvýšiť kvalitu prekladu. [22]

Mechanizácia prekladu bola dávnym snom ľudstva, ktorý sa stal skutočnosťou v 20. storočí s príchodom výpočtovej technológie. Avšak od perfektného stroja na prekladanie, ktorý by preložil text jednoduchým stlačením pár tlačidiel má strojový preklad ešte ďaleko aj v dnešnej dobe. Mnohý majú pochybnosti či je takýto systém vôbec reálne zostrojiť.

Systémy ktorými disponujeme v súčasnosti sú vo vývoji, no už aj tieto dokážu preložiť text na relatívne rozumnej úrovni. Navyše ak sa po preklade vykonajú dodatočné ručné úpravy takto nahrubo preloženého textu, dostávame kvalitný preklad za kratší čas ako keby mal byť celý text prekladaný ručne. Ďalším spôsobom ako vylepšiť preklad je zamerať sa len na preklad istého typu textov napr. politických, ktoré majú svoj špecifický slovník a skladbu viet.

Prekladové systémy môžu prekladať medzi špecifickou dvojicou jazykov alebo medzi skupinou jazykov a prekladať v jednom smere alebo v oboch.

Čo sa týka rôznych prístupov tvorby prekladových systémov, exitujú štyri základné prístupy. Tie sú nasledovné.

2.1.1 Priami preklad

Prvý a historicky najstarší je systém priameho prekladu (direct translation). Takýto systém je z pravydla vytvorený pre špecifický pár jazykov v jednom smere prekladu. Jedná sa o preklad po jednotlivých slovách bez ohladu na kontext. Pre preklad sa využívajú slovníky

v ktorých sa vyhľadá dané slovo v jednom jazyku a vráti sa jeho ekvivalent v druhom.[13] [22]

2.1.2 Pravidlovo orientovaný preklad

Ide o vylepšenie priameho prekladu o gramatické pravidlá. Narozdiel od iných prístupov, pravidlovo orientovaný preklad vyžaduje viacej informácií o morfológii, syntaktických pravidlách a sémantike oboch jazykov.

Tento typ prekladačov je najčastejšie používaný na tvorbu slovníkov a gramatických programov.

Pravidlovo orientované prekladače delíme na dva typy.[13]

Transferový preklad

Pri tejto metóde sa najprv vykoná morfologická a syntaktická analýza originálneho textu. Z tejto analýzy je získaná syntaktická reprezentácia textu z ktorej sa následne vygeneruje cieľový preklad textu. [13]

Interlingválny preklad

Táto metóda je podobná transferovej s jednou podstatnou zmenou. Zatiaľ čo pri transferovom preklade sa zo zdrojového jazyku po analýze a príslušnej reprezentácií priamo prekladalo do cieľového jazyka, tu je tomu inak. Pri tejto metóde je najprv zdrojový jazyk prevedený do jazyka Interlingua (medzijazyk). Je to umelý jazyk v ktorom je reprezentácia nezávislá na jazyku zdrojovom. Preklad do cieľového jazyka sa generuje z tohoto medzijazyka.Táto metóda má jednu veľkú výhodu a jednu veľkú nevýhodu. Výhodou je, že nie sme závislí na preklade len medzi dvoma jazykmi. Prekladač môže byť jednoducho obohatený o zdrojové i cieľové jazyky, nakoľko v strede sa nachádza nezávislá reprezentácia v jazyku interlingua. Veľkou nevýhodou je však to, že definovanie pravidiel prekladu do a z jazyka interlingua je veľmi náročné pri väčších doménach. Preto je tento preklad vhodný hlavne pre doménovo špecifickú oblasť.[13]

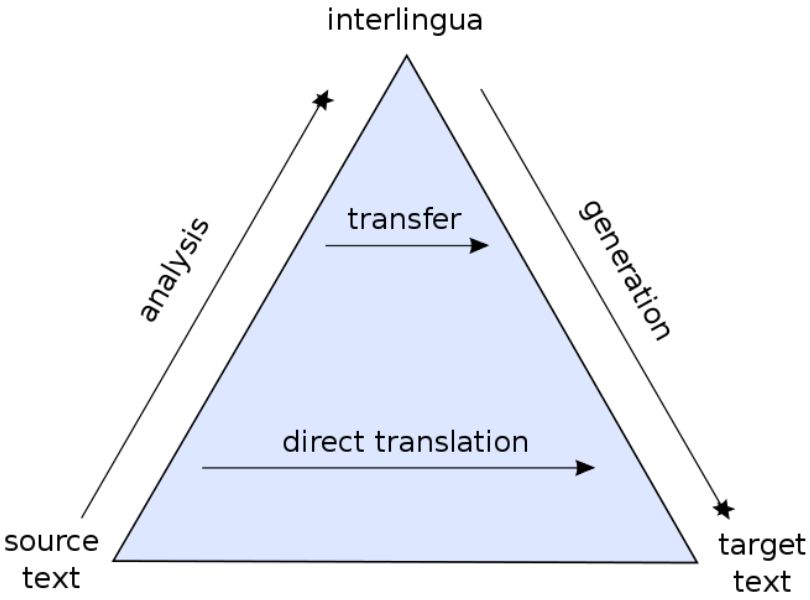
2.1.3 Štatistický preklad

V štatistickom strojovom preklade sú na preklad používané štatistické prekladové modely, ktorých parametre sú odvodené z analýzy dvojjazyčných paralelných textov tzv. korpusov.

Vo svojej podstate je štatistický preklad o výbere najpravdepodobnejšej možnosti prekladu daného slova či frázy. Tieto pravdepodobnosti sa systém učí sám, spracovávaním spomenutých viacjazyčných súborov. Využívajú sa dva modely prekladu. Prvý je založený na preklade samostatných slov. Druhý na preklade väčších celkov akými sú frázy alebo celé vety.

Prvý model prekladá po jednotlivých slovách pričom každé slovo má istú pravdepodobnosť prekladu. Systém vyberá preklad, ktorý má podla analýzy najvyššiu pravdepodobnosť prekladu na dané slovo. Vznikajú tu však isté problémy pri viacslovných výrazoch. Z tohoto dôvodu sa táto metóda veľmi nepoužíva.

Druhý model bol zavedený aby sa predchádzalo takýmto chybám a prekladajú sa nie samostatné slová, ale celé frázy alebo vety. V tomto modeli sa zo zdrojového textu vyberajú na základe pravdepodobnosti skupiny slov, a tie sú následne prekladané do cieľového jazyka.[22]



Obrázok 2.1: Bernard Vauquoisova pyramída, ktorá znázorňuje rôzne úrovne prekladu. Interlinguálny na vrchole nasledovaný transferovo orientovaným v strede a priamym prekladom v dolnej časti.[22]

2.1.4 Preklad založený na príkladoch

Ďalej len EBMT (example-based machine translation). Jedná sa o SP ktorý podobne ako u štatistického prekladu pracuje s multi-jazyčným súborom - korpusom, kde je text preložený do viacerých jazykov. EBMT pracuje na princípe prekladu z predošlých skúseností. Táto metóda zavrhuje myšlienku, že človek pri preklade prekladaný text najprv analyzuje, a až na základe tejto analýzy prekladá. EBMT namiesto toho predpokladá, že človek si pri preklade najprv rozdelí prekladaný text na vety, a tie následne ne frázy. Tieto frázy preloží a následne z nich vhodným spôsobom skomponuje vetu v cieľovom jazyku. Podobne to funguje aj pri EBMT. Systém sa najprv učí prostredníctvom multi-jazyčných súborov a vytvára si slovník fráz a ich prekladov. Následne keď je mu ponúknutý nový text, rozdelí si ho na vety a tie na frázy. Frázy preloží na základe analógie s predošlými prekladmi a napokon skomponuje vetu v cieľovom jazyku. Špecialitou pri EBMT je že korpus musí byť koncipovaný paralelne po vetách. Najprv je uvedená veta v zdrojovom jazyku, a následne veta v cieľovom jazyku.[22]

2.2 Paralelný korpus

Pre tvorbu prekladových systémov založených na štatistike, alebo príkladoch, je dôležité mať k dispozícii rozsiahle paralelné texty tiež známe ako paralelný korpus. Korpus predstavuje špecifický súbor jazykových dát istého prirodzeného jazyka. Obvykle sa skladá z textov rôznych štýlov a žánrov. Ak sa bavíme o paralelnom korpuse, máme na mysli dva korpusy s ekvivalentným textom v dvoch rôznych jazykoch. Aby bol takýto paralelný korpus po-

užiteľný pre trénovanie systémov, musí byť zarovnaný. Zarovnaním sa myslí identifikácia významovo rovnakých slov, fráz, alebo viet a ich umiestnenie na rovnaké miesto v rámci celého textu v oboch korpusoch.

Paralelné texty nie sú žiadna novinka. Ich pomoc bola využívaná už v dávnych časoch. Najznámejším paralelným textom je Rozetská doska, ktorá bola vytvorená roku 196 pnl. Jedná sa o kamennú dosku v ktorej sú vytesané paralelné egyptské a grécke texty a tak jej objav dopomohol k lepšiemu pochopeniu hieroglyfického písma. Ďalším príkladom použitia paralelného textu je napríklad Biblia, ktorá už na počiatku bola preložená do viacerých jazykov. V dnešnej dobe sú presné paralelné texty hlavne z oblasti zákonodarstva a politiky, kde je rovnaké znenie v rôznych jazykoch kľúčové.[22]

2.2.1 Zdroje paralelných textov

V tejto podkapitole sú uvedené rôzne zdroje paralelných textov, ich výhody a nevýhody pri získavaní a použití.

Knihy

Knihy sú vo všeobecnosti veľmi dobrým zdrojom paralelných dát, špeciálne pre účely strojového prekladu. Je to hlavne z dôvodu ich bohatej slovnej zásoby. Avšak získanie takýchto dát je v celku komplikované. Knihy sú v celku dosť ťažko zarovnatelné, pretože často pozostávajú z dlhých súvislých textov bez nejakých záchytných bodov. Môžu sa vyskytnúť aj iné komplikácie ako napríklad vynechanie niektorých viet pri preklade, alebo ich spojenie do jednej. V niektorých prípadoch sú dokonca spojené viaceré vety, alebo vynechané celé pasáže, pretože výsledok znie plynulejšie v danom jazyku. Preklad kníh nemá striktné obmedzenia a tak je dosť závislí na prekladateľovy. Z čoho vyplýva že automatické zarovnanie bude mať s takýmito textami problémy. Zarovnanie je potrebné ručne kontrolovať. [4]

S kniťnými paralelnými korpusmi sa viažu aj problémy s autorskými právami. Vzhľadom na autorské práva je volné šírenie takýchto textov nezákonné. Ich legálne získanie je často založené na písomných zmluvách a niekedy nie je ani možné. Je pravdepodobné, že ak aj takéto texty získame legálnou cestou budú vety musieť byť náhodne poprehadzované aby text ako celok nedával význam. Toto nemusí predstavovať problém, ak bol text pred zamiešaním zarovnaný.

Acquis Communitaire

Pred vstupom do Európskej únie musí každý nový členský štát preložiť a súhlasiť s existujúcou legislatívou, ktorá pozostáva z textov napísaných medzi rokmi 1950 a 2005. Tento súbor textov, ktorý pozostáva odhadom z osemtisíc dokumentov a pokrýva rozmanité témy, sa volá Acquis Communitaire (AC). Korpus obsahuje v súčasnosti texty preložené v dvadsiatich jazykoch zahrnujúc Češtinu aj Slovenčinu.

Dokumenty sú dostupné už v zarovnanej forme na vety alebo dokonca časti viet. Pre zarovnanie boli použité dva zarovnávacie programi; Vannila, ktorý implementuje Church and Gale algoritmus, a HunAlign. Vety jednotlivých párov jazykov sú zarovnané separátne, bez jednotného pivotného jazyka, výsledkom čoho sú kvalitnejšie, párovo špecifické zarovnania.[2]

Nakoľko sa jedná o volne dostupné texty, bez striktných autorských práv, ako tomu bolo pri knihách, ich získanie je bezproblémové.

Ec-Europa-Eu

Ďalším zdrojom paralelných textov môžu byť webové stránky samotné. Príkladom je oficálna stránka Európskej komisie. Táto webová lokalita pozostáva zo stránok v rôznych európskych jazykoch, zahŕňajúc aj češtinu a slovenčinu. Stránky sa líšia príponou v URL adrese a stránky s rovnakými názvami by mali obsahovať rovnaký text. Preto zarovnanie na úrovni dokumentu je v podstate jednoduché. Ak zapojíme do získavania takýchto textov jednoduchý web crawler, veľmi rýchlo získame niekoľko desiatok tisíc riadkov paralelného textu.

Táto stránka, a aj jej podobné, sú zväčša preložené z jedného zdrojového jazyka do ostatných. V tomto prípade to bol anglický jazyk čo môže spôsobiť isté nekonzistentnosti pri porovnaní dvoch cieľových prekladov. Ďalšou vadou s ktorou by sme mali počítať je nie vždy kompletný preklad zo zdrojového jazyka a tak preklady môžu obsahovať fragmenty nepreloženého textu v anglickom jazyku.[4]

Eur-LEX

Jedná sa o oficiálny denník európskej únie. Tento zdroj poskytuje značné množstvo paralelných dát. Čo sa týka obsahu, text je podobný korpusu Acquis a tak má aj podobné vlastnosti. České a slovenské preklady sú preložené z anglického jazyka. Dokumenty v tomto korpuse sú v XML formáte a tak je potrebné tieto texty konvertovať na jednoduchý text.[4]

Slovníky

Ako súčasť paralelného korpusu môžeme použiť aj dvojjazyčné slovníky. Ak sa slovníky správne zarovnajú, preklad jednotlivých slov má najvyššiu presnosť prekladu zo všetkých paralelných korpusov. Avšak takéto texty postrádajú kontext a slová sú často len v základnom tvare. Ďalšou nevýhodou je, že česko slovenské slovníky nie sú veľmi časté. Je však možné vytvoriť si vlastný česko slovenský slovník za použitia tretieho jazyka, ktorým môže byť napríklad angličtina. Príkladom je, že si vezmeme dva slovníky, anglicko-český a anglicko-slovenský a za použitia angličtiny ako pivotného jazyka vytvoríme česko slovenský slovník. Niektoré slovníky sú vo formáte XML čo nám značne uľahčuje orientáciu v dokumente a následné párovanie slov.

EMEA

Jedná sa o paralelný korpus vytvorený z PDF dokumentov Európskej liekovej agentúry. Texty sú prevedené do jednoduchého textu pomocou programu pdftotext. [15]

InterCorp

Jedná sa o projekt Filozfickej fakulty Univerzity Karlovy s cieľom vytvoriť paralelný korpus, ktorý by pokrýval veľké množstvo jazykov, prevažne tých ktoré sa na Fakulte Filozofickej študujú. Ako pivotný jazyk je používaná čeština. Texty sú zarovnávané prevažne ručne, ale obsahuje aj texty zarovnané automaticky. Korpus sa stále rozrastá a približne každý rok vychádza nová verzia. Celý projekt je akademický a nekomerčný, avšak prístup verejnosti ku korpusu je len pomocou internetového rozhrania a vyhľadávacieho enginu. Získanie korpusu vo forme textového súboru je možné len po podpísaní zmluvy. [19]

Europarl

Korpus rokovaní európskeho parlamentu medzi rokmi 1996-2011. Tento korpus bol získaný pomocou web crawleru zo stránky parlamentu európskej únie. Cieľom bolo vytvoriť korpus vhodný pre štatistické strojové prekladače. Texty Europarl korpusu sú dostupné vo viac ako 21 jazykoch v zarovnanej podobe na vety.[5][4]

OPUS

OPUS je rastúcou kolekciou voľne dostupných textov na webe. Je projektom, ktorý má za cieľ konvertovať a zarovnať volne dostupné dáta, pridávať lingvistické anotácie a poskytnúť širokej verejnosti paralelné korpusy. Všetky akcie vykonané na korpuse boli uskutočnené automaticky. Neboli podrobené žiadnej manuálnej úprave. Z týchto faktov vyplýva, že kvalita môže značne kolísať.[21] [12]

SNK

Slovenský národný korpus je elektronická databáza primárne obsahujúca slovenské texty od roku 1955 z rôznych štýlov, žánrov, vedeckých oblastí atď. K dispozícii je aj paralelný slovensko-český korpus. Texty sú automaticky zarovnané na vety. Korpus sa skladá z dvoch častí: podkorpus beletrie a podkorpusu voľne dostupných textov. Podkorpus voľne dostupných textov obsahuje prevažne preklady právnych textov a správ Európskej únie, počítačových a iných manuálov. V podkorpuse beletrie sa nachádza aj populárno-vedecká literatúra, literatúra faktu atď. Prístup k beletristickému podkorpusu je možný len cez webové rozhranie a pomocou vyhľadávacieho enginu. Stiahnutie textového dokumentu nie je verejnosti umožnené, vzhľadom na autorské práva, ktoré sa viažu s týmito textami. [14]

2.2.2 Proces zarovnania

Vo všeobecnosti proces zarovnania pozostáva z štyroch krokov:

1. Rozdelenie skupiny textov do dvoch jazykov
2. Zarovnanie textov na úrovni dokumentu
3. Zarovnanie textov na úrovni viet
4. Zarovnanie textov na úrovni slov

Rozdelenie skupiny textov do dvoch jazykov

Prepokladajme že máme k dispozícii skupinu textov v dvoch neznámych jazykoch bez informácie ktorý text patrí ku ktorému jazyku. Ako prvý problém s ktorým sa budeme potýkať je rozdelenie týchto textov do dvoch skupín podla príslušného jazyka.

Pri riešení tohoto problému využijeme faktu, že dokumenty napísané v rovnakom jazyku budú mať skupinu opakujúcich sa rečových jednotiek, bez ohľadu na to ako veľmi tematicky rozdielne tieto dokumenty sú. [9]

Zarovnanie na úrovni dokumentu

Často sa stáva, že dokumenty, ktoré máme k dispozícii, nie sú rozdelené na originál a preklad. tento stav nastáva hlavne ak dané dokumenty boli získané automatickým prehľadávaním webu pomocou crawlerov.

Našťastie ak existuje viacero verzii rovnakého dokumentu je zvykom zahrnúť v mene podreťazec, ktorý identifikuje jazyk v ktorom je dokument napísaný: príklad, Portuguese, por alebo pt. Tieto podreťazce podliehajú štandardu ISO-639-2. Ďalším spôsobom ako rozdeliť dokumenty je využiť meta-informácií, alebo adresy na ktorej bol dokument uložený.[9]

Zarovnanie textov na úrovni viet

Existuje viacero zdokumentovaných algoritmov a nástrojov schopných zarovnania na úrovni viet. Vo všeobecnosti ich môžeme rozdeliť do troch kategórií: algoritmy založené na dĺžke, slovníku alebo lexikóne a na čiastočnej podobnosti.

Vo všeobecnosti, zarovnávač vezme vstupný text, v niektorých prípadoch dodatočné informácie ako slovník, ktorý pomáha ustanoviť zhody. Typický algoritmus na zarovnanie viet začína výpočtom zarovnávacieho skóre a snahou nájsť v texte spoľahlivé záchytné body pre počiatočné zarovnanie. Toto skóre môže byť vypočítané na základe podobností v dĺžke, podobnosti slov, lexikálneho alebo syntaktického stromu. Po nájdení záchytných bodov sa proces opakuje. Toto sa väčšinou opakuje, kým nie sú nájdené žiadne ďalšie zhody.

Prvé zdokumentované postupy boli založené na meraní dĺžky viet v oboch dokumentoch. Táto metóda bola založená na myšlienke, že ak sú správne zarovnané vety, slová v nich musia byť tiež správne zarovnané. Algoritmus pracuje s pravdepodobnosťou správneho zarovnania. Je pravdepodobné, že prvá a posledná veta majú vysokú pravdepodobnosť správneho zarovnania na rozdielod ostatných. Následne je vypočítaná distribúcia slov v texte a slová s rovnakou distribúciou sú zvolené za záchytné, čo zúži možnosti pri výbere kandidátnych viet. Algoritmus následne iteruje kým konverguje k minimálnemu riešeniu. Tento systém založený na iterácii však nie je veľmi účinný pri veľkých paralelných korpusoch.[9]

Zarovnanie textov na úrovni slov

Proces zarovnania na slová sa podobá procesu zarovnania na vyšších úrovniach. Avšak jedná sa o omnoho komplexnejší proces vzhľadom na frázy, častejší výskyt opačného poradia slov vo vete a rozdieli v syntaktickej štruktúre. Vzhľadom na tieto okolnosti výskum v tejto oblasti nie je tak ďaleko ako na iných úrovniach zarovnávania.

Najčastejšie používaný model pre zarovnávanie slov je Hidden Markov model (HMM). Ide o generatívny model, ktorého podstatou je, že pravdepodobnosti zarovnania nie sú založené na absolútnej pozícii zarovnania slova, ale na relatívnej. To znamená že berieme do úvahy rozdieli medzi indexami pozícii slov.[9]

Pre jeden pár viet existuje vždy viac ako jedna cesta, ako navzájom poprepájať slová. Niekedy niektoré slová v zdrojovej vete nemajú svoj pár v cieľovejvete, a taktiež pre niektoré slová v cieľovej vete neexistuje žiadna linka na slovov zdrojovej vete. [8]

2.2.3 Existujúce zarovnávacie systémy

Pre zarovnávanie slov či už na úrovni slov alebo viet existuje viacero voľne dostupných nástrojov.

GIZA++

GIZA++ je nadstavbou staršieho zarovnávacieho nástoja GIZA, ktorý bol súčasťou SMT súboru nástrojov EGYPTH, ktorý bol vyvinutý tým v centre pre spracovanie jazyka a reči na univerzite Johns-Hopkins. GIZA++ obsahuje mnoho dodatočných vlastností, ktoré vytvoril Franz Josef Och. Implementuje HMM zarovnávací model: Baum-Welch tréning, Forward-Beckward algoritmus, prázdne slová, závislosť na slovných triedach, atď. Ide o štatistický zarovnávací systém, zarovnávajúci na úrovni slov a fráz. Vstupom musia byť súbory zarovnané na vety. [10] [16]

HunAlign

HunaAlign zarovnáva bilingválne texty na úrovni viet. V najjednoduchšom prípade je výstupom sekvencia bilingválnych párov viet. Ak je pri zarovnaní prítomný aj slovník, HunAlign ho použije a tak využije jeho informácie spolu s informáciami o dĺžke viet. Ak slovník nie je k dispozícii, najprv sa text zarovná podla dĺžok viet a následne sa z tohoto zarovnania vytvorí automatický slovník ktorý sa použije v druhej iterácii na zlepšenie zarovnania.

Ako mnohé ďalšie zarovnávacie systémy, hunalign nie je schopný sa vysporiadať so zmenou poradia viet: prekríženie segmentov A a B v jednom jazyku, ktoré sú v poradí B, A v druhom jazyku.

HunaAlign je implementovaný v jazyku C++ a je spustiteľný na ľubovolnom operačnom systéme. [3] [18]

NATools

Je skupina nástrojov pre spracovanie paralelných korpusov. Mimo iného obsahuje aj nástroj na zarovnávanie na úrovni viet, zarovnávač na úrovni slov a súbor ďalších nástrojov pre študovanie zarovnaných paralelných korpusov. [20]

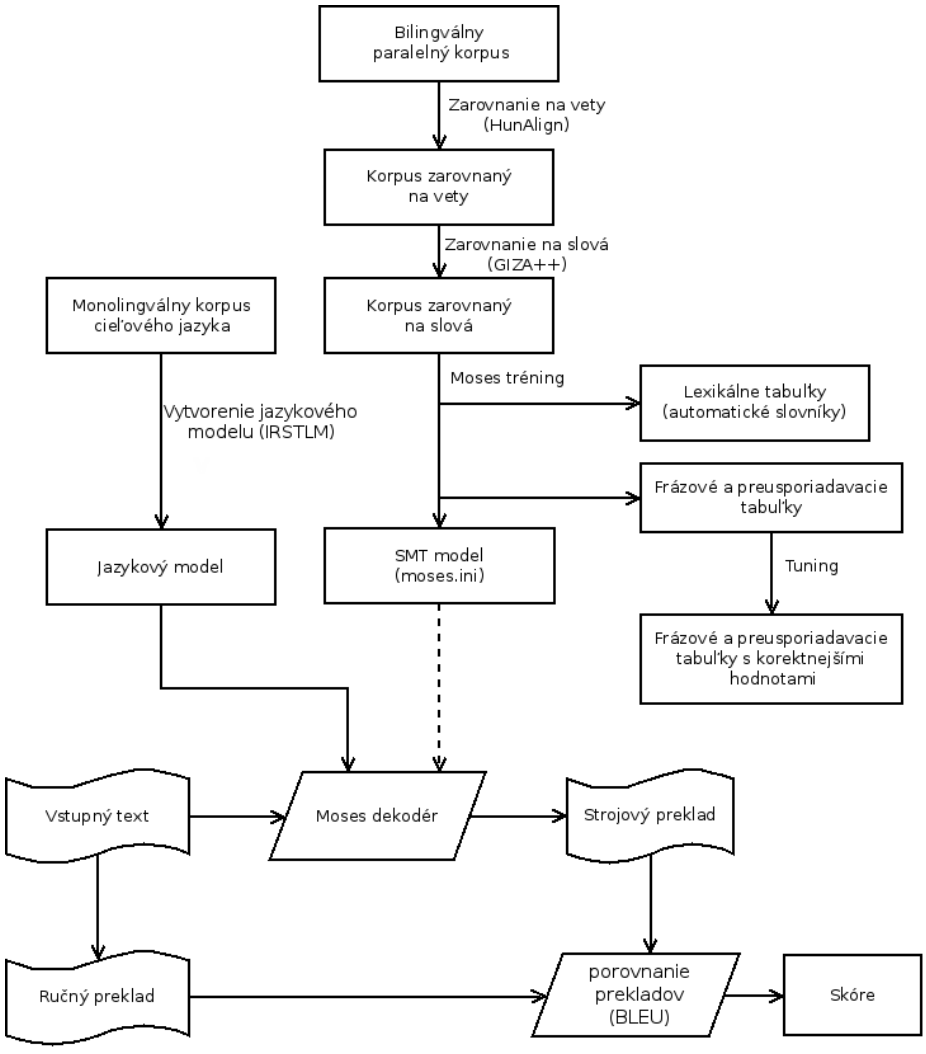
GMA

Geometric Mapping and Alignment je softwarový balíček ktorý implementuje SIMR algoritmus pre mapovanie bilingválných textov a Geometric Segment Alignment (GSA) pre konvertovanie bitextových máp na monotónne segmentové zarovnanie.[17]

Kapitola 3

Postup

V tejto kapitole je vysvetlený postup, kroky a technológia použitá pri vytváraní prekladového systému od korpusu až po ohodnoteniu prekladu.



Obrázok 3.1: Postup pri tvorbe a hodnotení systému

3.1 Moses

V tejto sekcii je popísaný systém pre vytváranie prekladových systémov Moses, ktorý som sa rozhodol použiť v tomto projekte.

Moses je systém pre štatistický strojový preklad, ktorý nám dovoľuje automaticky trénovať prekladové modely pre ľubovolný pár jazykov. Všetko čo potrebujeme je kolekcia preložených textov (korpus). Po vytvorení modelu výkonný vyhľadávací algoritmus dokáže rýchlo nájsť preklad s najvyššou pravdepodobnosťou spomedzi veľkého množstva možností.[6]

3.1.1 Technológia

Moses je implementáciou štatistického (dátami riadeného) prístupu k strojovému prekladu. Tento prístup je dominantným prístupom, ktorý využívajú online-ové prekladové systémy vytvorené spoločnosťami ako je Google alebo Microsoft. Tento prístup využíva veľké kvantá paralelných dát (tiež známych ako bitextov) zarovnaných na vety a rovnako tiež veľké kvantá jednojazyčných textov, na ktorých sa systém učí ako by mal cieľový jazyk vyzerať.

Proces trénovania vezme tieto paralelné dáta a používa vzájomný výskyt (coocurance) slov a segmentov (fráz) pre odvodenie prekladových vzťahov medzi dvoma jazykmi, s ktorými pracujeme. Vo frázovo orientovanom strojovom preklade sú tieto vzťahy jednoducho medzi sekvenciami slov. V hierarchickom frázovo orientovanom strojovom preklade alebo syntakticky orientovanom sú odvodené vzťahy viac štrukturované. Moses tiež implementuje rozšírenie frázovo orientovaného prekladu známeho ako faktorovaný preklad, ktorý umožňuje pridávanie dodatočných lingvistických informácií do frázovo založených systémov.[6]

3.1.2 Komponenty

Dva hlavné komponenty v Mosese sú tréningový pipeline a dekodér. Samozrejme sú tu aj iné pomocné nástroje, ktoré moses v procese môže používať. Tréningový pipeline je kolekciou nástrojov (poväčšine napísaných v perli a C++), ktoré berú vstupný paralelný korpus zarovnaný na úrovni viet a vracajú prekladový model. Dekodér je samostatná C++ aplikácia, ktorá potrebuje trénovaný prekladový model, jazykový model a vstupný text určený pre preklad a vracia jeho preklad v cieľovom jazyku.[6]

3.2 Získanie vstupných dát

Kľúčom k dobrému SMT systému je veľa kvalitných dát. Na výber je veľký počet zdrojov paralelných dát popísaných vyššie. Ďalším faktorom, ktorý zohráva rolu pri kvalitnom preklade je blízkosť trénovacích dát k cieľovým textom, ktoré bude náš systém prekladať. Toto je jedna z výhod prečo používať open-source nástroje ako je Moses. Ak máme vlastné dáta môžeme si vytvoriť prekladový systém na mieru, ktorý bude vyhovovať našim potrebám a potencionálne dostaneme systém, ktorý prekladá lepšie ako systémy so všeobecným zameraním. [6]

3.3 Predpripravenie korpusu

Predtým ako začneme korpus používať na trénovanie musíme ho upraviť do vhodného tvaru, v ktorom sa bude dobre analyzovať a v ktorom ho dokážu použité aplikácie správne interpretovať. Preto podrobíme paralelný korpus nasledujúcim štyrom úpravám:

1. Zarovnanie na vety
2. Tokenizácia
3. Truecasovanie
4. Čistenie korpusu

3.3.1 Zarovnanie na vety

Pre zarovnanie na úrovni viet môžeme použiť niektorý z vyššie uvedených programov. V tomto projekte bol použitý HunAlign. Výsledný formát paralelného korpusu je nasledovný: každý riadok je jedná veta, pričom každá veta v jednom dokumente má na rovnakom mieste svoj preklad v druhom dokumente.

3.3.2 Tokenizácia

Tokenizácia je dôležitým krokom z hľadiska správnej interpretácie slov. Spočíva v dosadení medzier medzi slová a interpunkčné znamienka. Pre tokenizáciu bol použitý skript tokenizer.perl, ktorý je súčasťou Mosesa.

3.3.3 Truecasovanie

Truecasovane je ďalším krokom zabezpečujúcim lepšiu interpretáciu. Rieši problém veľkého písmena na začiatku vety v slove, ktoré nie je vlastným podstatným menom. Tieto písmená sú zmenená na malé. Avšak na začiatku slova sa môže vyskytnúť aj vlastné podstatné meno, v ktorom by sa nemalo prvé písmeno zamieňať za malé. Preto je pred samotným truecasovaním pomocou analýzy vstupného paralelného korpusu vytvorený tzv. truecase-model. Je to súbor v ktorom sú uvedené v každom riadku početnosti daného slova s malým písmenom na začiatku a s veľkým. Podľa tohoto modelu a teda vyššej početnosti daného tvaru slova sa následne rozhodne či sa písmeno zmení, alebo zostane veľké. Pre vytvorenie truecase-modelu bol použitý skript train-truecaser.perl a pre samotné truecasovanie truecase.perl. Oba skripty sú súčasťou Mosesa. [6]

3.3.4 Čistenie korpusu

Poslednou fázou predúpravy korpusu je dodatočné odstránenie prázdnych riadkov, prebytočných medzier ale aj odstránenie príliš krátkych, dlhých alebo zle zarovnaných viet, nakoľko tieto môžu spôsobovať problémy pri trénovaní. Pre tieto úpravy bol použitý skript clean-corpus-n.perl, ktorý je súčasťou Mosesa.[6]

3.4 Vytrvorenie jazykového modelu

Modelovanie jazyka sa zaoberá problematikou predikcie nasledujúceho slova na základe slov predchádzajúcich. Tento druh úlohy je známi aj pod pojmom Shannonová hra.

Úlohu predikcie nasledujúceho slova môžeme chápať ako pokus o určenie pravdepodobnosti P:

$$P(w_n|w_1, ..., w_{n-1})$$

Pri takomto stochastickom probléme používame predchádzajúce slová, históriu, pre predpoveď slova nasledujúceho. Na základe toho, že sme preštudovali veľa textu v danom jazyku,

vieme ktoré slová zvyknú nasledovať predchádzajúce. Je teda potrebné zaviesť istý systém, ktorý bude klasifikovať rovnaké úseky histórie a slová ktoré za nimi budú nasledovať. Možnou metódou ako toto docieliť je tzv. Markovov predpoklad. Tento predpoklad hovorí, že len niekoľko predchádzajúcich slov ovplyvňuje nasledujúce slovo. Ak preto zostrojíme model v ktorom rovnakých n-1 predchádzajúcich slov je umiestnených do rovnakej skupiny rovnosti, dostávame n-1 stupňový Markov model, alebo teda n-gram-ový model, v ktorom posledné n-té slovo je dôsledkom n-1 predchádzajúcich. Väčšinou je n=2,3,4 ktoré sa nazývajú bigram, trigram a four-gram-ové modely. [1]

V tomto projekte boli tvorené trigram-ové jazykové modely pomocou programu IRSTLM. [7]

3.5 Tréning

Proces vytvorenia prekladového systému pozostáva z viacerých krokov, ktoré sú implementované ako pipeline pričom mozes umožňuje jednoduché vkladanie externých nástrojov to tréningovej pipeline. Proces trénovania môžeme rozdeliť do týchto deviatich krokov:[6]

1. Pripravenie dát
2. Beh GIZA++
3. Zarovnanie slov
4. Vytvorenie lexikalnej prekladovej tabuľky
5. Extrahovanie fráz
6. Ohodnotenie fráz
7. Vytvorenie lexikalizovaného preusporiadavacieho modelu
8. Vytvorenie generovacieho modelu
9. Vytvorenie konfiguračného súboru

3.5.1 Pripravenie dát

Nakoľko budeme na zarovnanie používať GIZA++, tak musíme vstupné dáta previesť do formátu vhodného pre spracovanie. Sú vygenerované dva súbory so slovnou zásobou a paralelný korpus je prevedený do číselného formátu.

Pre obe polovice korpusu je vytvorený samostatný súbor so slovnou zásobou, pozostávajúci z troch stĺpcov. V prvom je číselný identifikátor slova, ktoré sa nachádza v druhom stĺpci. V treťom stĺpci sa nachádza počet výskytov daného slova v rámci danej polovice korpusu.

Prevedením korpusu do číselného formátu rozumieme nasledovné. Každá veta ja reprezentovaná tromi riadkami čísel. V prvom riadku je jedno číslo, ktoré reprezentuje počet výskytov danej vety v rámci textu. V druhom a treťom riadku sa nachádza sekvencia čísel, ktorá je reprezentáciou originálnej vety v zdrojovom jazyku (druhý riadok) a v cieľovom jazyku (tretí riadok), ale slová v nich boli nahradené číselnými identifikátormi zo súborov so slovnou zásobou.

3.5.2 Beh GIZA++

GIZA++ je voľne dostupná implementácia IBM modelu pre zarovnanie textu na úrovni slov. Proces zarovnávania na slová pomocou GIZA++ je z celého trénovania modelu najzdĺhavejší proces. Rovnako je náročný aj na pamäť.

Slová zarovnávame v dvoch samostatných behov zarovnania, kedy v jednom zarovnávame text v cieľovom jazyku podla zdrojového a v druhom naopak. Príklad obsahu súboru so zarovnaním v jednom aj druhom smere:

cs-sk:

```
# Sentence pair (11923429) source length 5 target length 4 alignment score :
1.30047e-06
nevšímej si Protivy .
NULL ({ }) nezahadzuj ({ 1 }) sa ({ 2 }) so ({ }) Zloduchom ({ 3 }) . ({ 4 })
```

sk-cs:

```
# Sentence pair (11923429) source length 4 target length 5 alignment score :
9.39915e-08
nezahadzuj sa so Zloduchom .
NULL ({ }) nevšímej ({ 1 }) si ({ 2 }) Protivy ({ 3 4 }) . ({ 5 })
```

3.5.3 Zarovnanie slov

Výsledné zarovnanie vytvoríme z prieniku týchto dvoch súborov, plus sa používajú rôzne heuristiky na vytvorenie nových a lepších zarovnaní.

3.5.4 Vytvorenie lexikálnej prekladovej tabuľky

Keď už máme text zarovnaný na slová, je v celku jednoduché odvodiť si pravdepodobnosti prekladov slov. Z týchto pravdepodobností sa vytvoria dva súbory v oboch smeroch prekladu, v ktorých sa nachádzajú tri stĺpce. Prvý je slovo v jednom jazyku, v druhom jeho preklad a v treťom pravdepodobnosť tohoto prekladu. Nižšie môžeme vidieť obsahu takéhoto súboru na príklade desiatich najpravdepodobnejších prekladov slova zeměkoule.

```
zeměkoule zemeguľa 1.0000000
zeměkoule zem 0.0103448
zeměkoule zem 0.0036397
zeměkoule Zeme 0.0073171
zeměkoule zemeguli 0.0909091
zeměkoule zemegule 0.7500000
zeměkoule zemeguľa 0.3384615
zeměkoule trvá 0.0002910
zeměkoule svet 0.0002571
zeměkoule sveta 0.0009529
```

Takéto lexikálne tabuľky nachádzajú uplatnenie aj mimo štatistického strojového prekladu, ako základ pre autoamticky generované bilingválne slovníky.

3.5.5 Extrahovanie fráz

V tomto kroku sú všetky frázy zo zarovnaného korpusu uložené do jedného súboru s tromi stĺpcami. V prvom je fráza v jednom jazyku v druhom jej možný preklad v druhom jazyku a v treťom sú body zarovnanie tejto frázy. Následne je vytvorený inverzovaný súbor, v ktorom sú stĺpce jedna a dva prehodené.

```
že ačkoli mohly být její pochybnosti ohledně ||| že hoci sa jej pochybnosti o
||| 0-0 1-1 3-2 4-3 5-4 6-5
že ačkoli mohly být její pochybnosti ||| že hoci sa jej pochybnosti ||| 0-0 1-1
3-2 4-3 5-4
že ačkoli mohly být její ||| že hoci sa jej ||| 0-0 1-1 3-2 4-3
že ačkoli mohly být ||| že hoci sa ||| 0-0 1-1 3-2
že ačkoli mohly ||| že hoci ||| 0-0 1-1
```

3.5.6 Ohodnotenie fráz

Následne je vytvorená prekladová tabuľka z uložených párov preložených fráz z predchádzajúceho kroku. Rovnako ako v kroku 4, každá fráza musí mať svoju pravdepodobnosť prekladu. Aby sme tieto pravdepodobnosti získali postupujeme nasledovne: V prvom kroku je súbor s frázami usporiadaný. Toto zaistí, že frázy v prvom jazyku sú pod sebou. Následne sú spočítané rovnaké frázy s rovnakým prekladom a je odvodená pravdepodobnosť prekladu. Pre opačnú pravdepodobnosť je rovnako spracovaný súbor s invertovanými stĺpcami fráz z predchádzajúceho kroku. Výsledná prekladová tabuľka fráz vyzerá nasledovne:

```
v evropě ||| v európe ||| 0.5 0.409578 1 0.716958 ||| 0-0 1-1 ||| 4 2 2 ||| |||
v Evropě ||| v európe ||| 0.25 0.204789 9.53925e-05 5.37718e-05 ||| 0-0 1-1 |||
4 10483 1 ||| |||
v Evropě , ||| v európe ||| 0.25 0.0283772 0.000737463 5.37718e-05 ||| 0-0 1-1
||| 4 1356 1 ||| |||
```

3.5.7 Vytvorenie lexikalizovaného preusporiadavacieho modelu

Tento model je potrebný pri preusporiadávaní slov vo vetách v ktorých sa mení slovosled pri preklade do iného jazyka.

3.5.8 Vytvorenie generovacieho modelu

Tento model je vytvorený za pomoci jednojazyčného textu cieľového jazyka. Takýmto je cieľová časť nášho vstupného paralelného korpusu.

3.5.9 Vytvorenie konfiguračného súboru

Na záver je vygenerovaný konfiguračný súbor pre dekodér s cestami ku všetkým potrebným súborom, ktoré boli vytvorené v predošlých krokoch.

3.6 Tuning

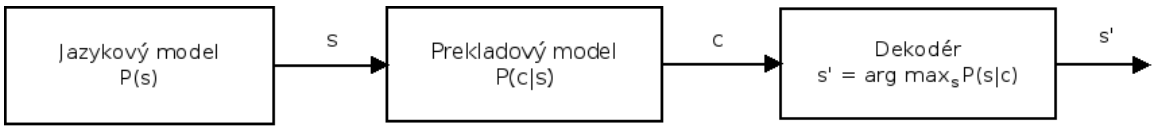
Tuning (ladenie) je nepovinnou a najpomalšou časťou tvorby prekladového systému. Systém bude fungovať aj bez tohoto kroku, ale jeho kvalita môže byť značne nižšia.

Podstatou tuningu je vylepšenie pravdepodobností vo frázových a preusporiadavacích tabuľkách. Toto je docielené opakovaným automatickým prekladaním malého trénovacieho paralelného korpusu a na základe kvality jeho prekladu sú následne odvodené nové hodnoty pravdepodobností v tabuľkách.

3.7 Preklad

Pri štatistickom strojovom preklade používame model tzv. zašumeného prenosového kanálu. Jeho podstatou je, že máme istý prenosový kanál, ktorý má tendenciu prenášanú informáciu pozmeniť a preto je nutné používať dekodér, ktorý nám danú informáciu dokáže upraviť do tvaru podobného tomu, ktorý bol vysielaný. Dekodér funguje za predpokladu, že máme informácie o spôsobe enkódovania správy na strane odosielateľa a informácie o spôsobe akým prenosový kanál mení prenášanú informáciu.

V SMT je vytvorená paralela na tento model, kde našou vstupnou informáciu je náš cieľový jazyk, ktorý bol však pri „prenose" pozmenený na náš zdrojový jazyk a preto za pomoci dekodéra sa snažíme z tejto pozmenenej informácie získať pôvodnú. Čiže v podstate náš prekladaný text chápeme ako zašumený preklad. Informácie pre dekodér o spôsobe enkódovania správy reprezentuje jazykový model a informácie o prenosovom kanále a jeho funkcii pozmenenia prenášanej informácie reprezentuje prekladový model, ktorý je výsledkom tréningu.[1]



Obrázok 3.2: Model zašumeného komunikačného kanálu. Text v češtine je reprezentuje c, text v slovenčine s, a preklad s'.

3.7.1 Dekodér

Úlohou dekodéra je vyvodiť čo najpravdepodobnejší tvar vstup z daného výstupu. Ten sa dá odvodiť pomocou Bayesovho teorému.

$$s' = argmax_s P(s|c) = argmax_s \frac{P(s)P(c|s)}{P(c)} = argmax_s P(s)P(c|s)$$

P(c) môžeme v druhej úprave odstrániť, nakoľko jeho hodnota je fixná.

3.7.2 Postup prekladu

Využívajú sa dva modely prekladu. Prvý je založený na preklade samostatných slov. Druhý na preklade fráz. V našom prípade je použitý model prekladu fráz. Pri použití takéhoto

modelu sú v prvom kroku najprv preložené všetky slová a frázy na množiny ich najpravdepodobnejších ekvivalentov v cieľovom jazyku. Z týchto množín sa následne vyberie ich najpravdepodobnejšia kombinácia. V tejto fáze je uplatňovaný aj preusporiadavycí model podla ktorého sú preložené frázy a slová usporiadané do ich najpravdepodobnejšieho poradia vzhľadom na slovosled v danom jazyku.[1]

3.8 Hodnotenie kvality prekladu

Po preložení je posledným krokom určiť aký dobrý je náš preklad v porovnaní s ľudským. Hodnotenie správnosti prekladu je záludné, nakoľko ani dvaja ľudia nemusia preložiť daný text na identický preklad aj keď oba preklady majú rovnaký význam. Exitujú však štandardné metriky, ktoré dokážu s istou presnosťou správnosť prekladu odhadnúť.

3.8.1 BLEU

BLEU (bilingual evaluation uderstudy) je algoritmus pre ohodnotenie kvality textu, ktorý bol strojovo preložený z jedného prirodzeného jazyka do iného. Kvalita je vnímaná ako mira zhody medzi strojovým prekladom a ručným prekladom. Čím je preklad bližšie k profesionálnemu ľudskému prekladu, tým je lepší. Bleu bola jednou z prvých metrík, ktorá dosiahla vysokú zhodu s ľudským úsudkom o kvalite. BLEU bol navrhnutý aby sa zhodoval s ľudským úsudkom na úrovni korpusu. Na úrovni viet už nevykazujú vysokú zhodu s človekom.

Kvalita prekladu je meraná na stupnici od 0 po 100. Čím sú si dva texty podobnejšie, tým je výsledné skóre bližšie k 100. Avšak málo dvoch ľudských prekladov sa priblíži alebo dosiahne hodnotenie 100. Aby bola výsledná hodnota 100 musia byť dva kandidátne preklady identické. [11]

Pre spresnenie ohodnotenia sú referenčné aj kandidátne preklady predkladané BLEU v tokenizovanej podobe.

3.9 Zhrnutie

Pri vytváraní prekladových systémov v tomto projekte som sa držal vyššie uvedeného postupu a každý systém prešiel nasledovnými fázami.

1. Získanie paralelných textov
2. Vytvorenie korpusu
3. Zarovnanie na úrovni viet
4. Tokenizácia, truecasovanie, čistenie
5. Vyrvorenie jazykového modelu
6. Spustenie trénovacej pipeline
7. Tuning
8. Preklad
9. Ohodnotenie

Pre zjednodušenie práce bolo vytvorených niekoľko skryptov v shelly a pythone. Pre vytvorenie základného systému bol vytvorený shell skript, ktorý vykonal fázy 3 až 6. Vytvorený netrénovaný systém bol odtestovaný pre funkčnosť. Pre veľkosť frázových a preusporiadavacích tabuliek museli byť tieto pri väčších korpusoch binarizované, nakoľko pri spustení systému sa inak načítavajú do pamäte, ktorej veľkosť však na používanom zariadení nepostačovala. Nevýhodou je že takto binarizované tabuľky zaberajú veľa miesta na pevnom disku. Následne boli vytvorené trénované verzie tohoto systému, bol na nich vykonaný preklad a následne zmeraná ich kvalita.

Ďalej boli vytvorené pomocné skripty ako napríklad skript na spojenie anglicko-českého a anglicko-slovenského slovníka, úpravu formátu textov, skript na porovnanie automaticky získaných slovníkov atd.

Kapitola 4

Výsledky a štatistiky

V tejto kapitole sa budem zaoberať konkrétnou prácou vykonanou na tomto projekte. Je tu popísaný cieľ ktorý som sa snažil dosiahnuť, štatistické údaje a bližší popis konkrétnych použitých korpusov a výsledky vytvorených prekladových systémov.

4.1 Cieľ

Ako cieľ som si stanovil vytvoriť viacero prekladových systémov, nielen jeden, ktoré boli vytvorené rôznymi spôsobmi. Takto som chcel zistiť, aká kombinácia vykazuje najlepšie výsledky. Z analýzy prác, ktoré sa zaoberali štatistickým prekladom som vypozoroval, že hlavným faktorom, ktorý vo veľkej miere ovplyvňuje kvalitu prekladového systému, sú vstupné bilinguálne dáta. Ďalším krokom, ktorý môže zmeniť kvalitu prekladu, je ladenie (tuning) výsledného modelu získaného procesom trénovania. V tomto procese sa používajú malé paralelné dáta a práve to aké môže rôzne ovplyvniť kvalitu prekladu.

4.2 Korpusi

V tejto sekcii sú uvedené rôzne korpusi, ktoré, boli použité pre vytvorenie systémov a ich štatistické údaje a veľkosti.

4.2.1 Acquis, Ec-Europa-Eu, EMEA, Eur-LEX, Europarl-v6

Tieto korpusi sú bližšie popísané v analýze témy. V skratke ide o politické a medicínske texty, ktoré som získal už zarovnané na úrovni viet. V nasledujúcej tabuľke sú uvedené ich štatistiky.

	MB	**viet**	**slov**	
Acquis	151.4	926 082	20 228 962	
Ec-Europa-Eu	3.0	24 190	378 226	
EMEA	85.8	1 067 905	11 533 146	
Eur-LEX	372.2	3 078 210	4 5648 278	
Europarl-v6	67.9	459 089	9 015 624	

Tabuľka 4.1: Tabuľka veľkostí korpusov pre českú časť

	MB	viet	slov	
Acquis	152.5	926 082	20 384 971	
Ec-Europa-Eu	3.0	24 190	375 092	
EMEA	83.9	1 067 905	11 401 202	
Eur-LEX	372.8	3 078 210	45 648 278	
Europarl-v6	68.4	459 089	9 110 973	

Tabuľka 4.2: Tabuľka veľkostí korpusov pre slovenskú časť

4.2.2 Slovníky

V tomto projekte boli použité aj dva slovníky. Jeden priamo česko-slovenský, ktorý bolo treba len rozdeliť na dva súbory. Nakoľko slová boli pod sebou tu problém nebol. Druhý slovník bol vytvorený z dvoch slovníkov anglicko-českého a anglicko-slovenského. Slovníky boli v xml formáte a pre ich spojenie som použil vlastný skript v pythone. Výsledný korpus je spojením oboch slovníkov a nachádza sa v ňom 580 062 preložených slov a fráz. Aby slovníky mohli byť použité pri trénovaní, bol každý riadok ukončený bodkou, aby boli brané ako vety.

4.2.3 SNK

Pre potreby rozšírenia slovnej zásoby systému a z toho vyplývajúceho lepšieho prekladového modelu, je vhodné mať k dispozícii beletristické paralelné texty. Pre potreby tohoto projektu, bol Slovenský Národný korpus tak ústretový, že mi poskytol ich beletristickú časť paralelného česko slovenského korpusu, ktorý je inak nedostupný. Vety v ňom museli byť z dôvodu autorských práv zamiešané. Nakoľko však zarovnanie na vety nebolo porušené, je táto zmena, pre potreby tohoto projektu, zanedbateľná. V tomto korpuse sa nachádza 136 kníh rôznych žánrov, poviedky a básne.

	MB	viet	slov	
česká časť	53.8	653 472	9 525 324	
slovenská časť	53.3	653 472	9 510 313	
spolu	107.1	1 306 944	19 035 637	

Tabuľka 4.3: Veľkosť korpusu od SNK

4.2.4 Ostatné zdroje

Ďalším zdrojom beletristických paralelných textov, ktoré sa mi podarilo získať, je séria kníh Herry Potter. Séria obsahuje 7 kníh, ktoré som získal v pdf formáte ako e-booky. Tento formát so sebou prináša mnohé komplikácie. V prvom rade pdf formát na musí byť konvertovaný na klasický textový. K tomuto som využil open-source nástroj príkazovej riadky pdftotext. Takáto konverzia však nie je dokonalá a ak pdf obsahovalo obrázky, umelecké nadpisy alebo číslovania strán, tak takéto objekty navyše pri konverzii vytvoria vo výslednom textovom súbore rôzne artefakty, ktorých je potrebné sa zbaviť. V tomto prípade vznikli artefakty len v slovenskej časti korpusu a to konkrétne pri nadpisoch kapitol. Nakoľko by bolo dosť náročné postihnúť regulárnym výrazom všetky tieto artefakty a ich počet nebol priveľký, tak boli odstránené ručne. Narozdiel od toho číslovania strán mali uniformnú notáciu a tak ich odstránenie pomocou regulárneho výrazu bolo jednoduché.

Ďalším nedostatkom takto získaných textov je chýbajúce zarovnanie na úrovni viet. Pre zarovnanie som si zvolil program HunAlign avšak ten vyžaduje isté vlastnosti zarovnávaného textu, ktoré musia byť splnené aby zarovnanie bolo úspešné.

Jednou z vlastností je, že každý riadok je jedna veta. Formát pdf je dosť špecifický čo sa týka formátovania textu. Nespolieha sa na zobrazovací software že zalomí riadky. Namiesto toho pdf explicitne špecifikuje miesta zalamovania riadkov, čo sa odrazí aj v konvertovanom textovom formáte, kde sa následne môže objaviť koniec riadka aj uprostred vety. Odstránenie prebytočných zakončení riadkov bolo vykonané v textovom editore emacs za pomoci funkcie vyhľadanie a nahradenia regulárneho výrazu jednoduchou medzerou. Následne tým istým spôsobom bol za každú bodku umiestnený znak konca riadku. Po týchto úpravách dostávame formát vhodný pre spracovanie HunAlignom.

Je tu však ešte jeden problém, na ktorý som narazil počas samotného zarovnávania. Zdá sa totiž, že HunAlign má problémy pri spracovávaní veľkého paralelného korpusu. Keď som mu predložil paralelný korpus pozostávajúci zo všetkých siedmich kníh, tak uprostred procesu vracal HunaLign segmentation fault. Preto boli knihy zarovnané samostatne a až po zarovnaní boli spojené do jedného korpusu.

Pri zarovnávaní je vhodné poskytnúť HunAlignu aj prekladový slovník, ktorý vylepší presnosť zarovnania. Ako slovníkové dáta bol použitý korpus slovníkov popísaný vyššie. Jediná zmena bola v neukončovaní riadkov bodkou a spojení českej a slovenskej časti do jedného dokumentu kde originál a preklad sú oddelené zavináčom.

	MB	**viet**	**slov**	
česká časť	7.1	56 680	1 228 203	
slovenská časť	6.3	56 680	1 110 042	
spolu	13.4	113 360	2 338 245	

Tabuľka 4.4: Veľkosť korpusu Harry Potter po zarovnaní

4.3 Test sety

Pre odtestovanie a ohodnotenie prekladového systému potrebujeme malý súbor paralelných dát, ktorý predložíme systému na preloženie. Tieto dáta musia byť zarovnané aspoň na úrovni viet a aby systém nemal problém pri interpretovaní textu s čiarkami, bodkami atď. je vhodné tieto dáta predkladať systému v tokenizovanej a truecasovanej podobe.

4.3.1 WMT 2011

Hlavným testovacím súborom, ktorým som testoval preklad, je test set z podujatia WMT 2011, ktorý obsahuje 3003 viet v angličtine a ich ručné preklady v češtine a slovenčine. Jedná sa predovšetkým o politický text.

WMT je jeden z najprestížnejších podujatí v rámci výpočtovej lingvistiky, zameraný predovšetkým na strojový preklad. V rámci tohoto podujatia je súťaž o najlepší preklad stanoveného textu. A práve tento text z ročníka 2011 budem používať na ohodnotenie kvality prekladu mojich systémov.

Pri kontrole tohoto textu som však narazil na nekonzistentnosť v podobe anglických úvodzoviek (“”) v českej časti. Pri tokenizácii sa anglické úvodzovky nahradzujú retazcom &quot avšak v slovenskej časti boli úvodzovky slovenského typu („”), ktorý sa však pri

tokenizácii nezmenil. Preto boli dodatočne v slovenskej časti pomocou regulárneho výrazu všetky dolé a horné úvodzovky nahradené reťazcom &quot .

4.4 Ladiace paralelné dáta

Pre ladenie (tuning) boli použité tieto paralelné dáta. Samotný SNK bol používaný až po odobratí daného množstva dáta pre tuning preto sa medzi ladiacimi a trénovacími dátami nenachádza prienik.

- Ec-Europa-Eu
- Tuning set 2 - pozostáva z prvých 3000 viet Ec-Europa-Eu
- Tuning set 3 - pozostáva z prvých 3000 viet korpusu od SNK
- Tuning set 4 - kombinácia test setu 2 a 3, 6000 viet

4.5 Vytvorené systémi

Jednotlivé vytvorené prekladové systémy boli pomenovávané podla korpusu, ktorý bol použitý na ich trénovanie. AIO (all in one) je pracovné pomenovanie pre trénovacie korpusi, ktoré vznikli kombináciou viacerých menších podkorpusov.

4.5.1 Acquis

Systém trénovaný pomocou korpusu Acquis bol vytvorený ako prvý a zároveň prototypový prekladový systém. Systém bol testovaný bez odladenia a aj s odladením na korpuse Ec-Europa-Eu. Ako testovací korpus bol použitý WMT 2011. Kvalita prekladu bola zmeraná pomocou BLEU.

	WMT 2011	
neodladený	35.14	
Ec-Europa-Eu	40.65	

Tabuľka 4.5: Kvalita prekladu systému Acquis zisťená pomocou BLEU

4.5.2 Systém AIO

Korpus AIO v sebe kombinuje všetky politické korpusy, ktoré som mal k dispozícii okrem korpusu Ec-Europa-Eu, ktorý som kvôli svojej veľkosti vyčlenil na iné účeli. Obsahuje: Acquis, EMEA, Europarl-v6 a Eur-LEX.

	MB	**viet**	**slov**
česká časť	677.5	5 531 286	86 426 010
slovenská časť	677.5	5 531 286	86 514 850
spolu	1355	11 062 572	172 940 860

Tabuľka 4.6: Veľkosť AIO

	WMT 2011	
neodladený	39.62	

Tabuľka 4.7: Kvalita prekladu systému AIO zisťená pomocou BLEU

4.5.3 Systém AIO2

AIO2 stavia na korpuse pre AIO, ale navyše obsahuje aj beletristické korpusy od SNK a korpus Harry Potter.

	MB	viet	slov
česká časť	738.6	6 380 799	96 995 354
slovenská časť	737.21	6 284 885	96 950 283
spolu	1475.81	12 665 684	193 945 637

Tabuľka 4.8: Veľkosť AIO2

	WMT 2011	
neodladený	42.47	
tuning set2	45.19	
tuning set3	40.30	
tuning set4	43.11	

Tabuľka 4.9: Kvalita prekladu systému AIO2 zisťená pomocou BLEU

4.5.4 Systém AIO3

AIO3 používa korpus AIO2 rozšírený o korpus slovníkov.

	MB	viet	slov
česká časť	746.93	6 957 861	97 831 780
slovenská časť	745.41	6 861 947	97 761 239
spolu	1492.34	13 819 808	195 593 019

Tabuľka 4.10: Veľkosť AIO3

	WMT 2011	
neodladený	42.45	
tuning set2	48.31	
tuning set3	50.31	
tuning set4	49.72	

Tabuľka 4.11: Kvalita prekladu systému AIO3 zisťená pomocou BLEU

4.5.5 Systém AIO4

Korpus AIO4 je rozšírením korpusu AIO2 o slovníky, ktoré sú ale na rozdiel od korpusu AIO3 pridané 10 krát. Cieľom bolo zvýšiť početnosť slovníkových prekladov o ktorých vieme, že majú vysokú presnosť a tak zvýšiť aj pravdepodobnosť, že pri preklade sa použije práve tento preklad.

	MB	**viet**	**slov**
česká časť	824.21	12 178 419	105 750 430
slovenská časť	821.40	12 082 505	105 450 264
spolu	1645.61	24 260 924	211 201 694

Tabuľka 4.12: Veľkosť AIO4

	WMT 2011	
neodladený	42.84	
tuning set2	48.16	
tuning set3	50.58	
tuning set4	50.07	

Tabuľka 4.13: Kvalita prekladu systému AIO4 zisťená pomocou BLEU

4.5.6 Systém AIO5

Pri tvorbe predchádzajúcich kombinovaných korpusoch vznikla chyba v zarovnaní na vety, ktorú môžeme vidieť aj v štatistických údajoch o korpusoch, kde sa počet viet nezhoduje a preto bol vytvorený korpus AIO5, v ktorom bola táto chyba opravená.

	MB	**viet**	**slov**
česká časť	824.13	12 039 058	105 935 613
slovenská časť	821.34	12 039 058	105 635 186
spolu	1645.47	24 078 116	211 570 799

Tabuľka 4.14: Veľkosť AIO5

	WMT 2011	
neodladený	42.74	
tuning set2	48.76	
tuning set3	50.76	
tuning set4	50.35	

Tabuľka 4.15: Kvalita prekladu systému AIO5 zisťená pomocou BLEU

4.6 Existujúce prekladové systémi

Pre porovnanie kvality prekladu mnou vytvorených systémov s inými existujúcimi a používanými prekladovými systémami som preložil testset WMT2011 na prekladačoch Google translate a na systéme Česílko.

4.6.1 Google translate

Je to viacjazyčná služba poskytovaná Googlom pre preklad písaného textu z jedného jazyka do druhého. Podporuje cez 90 jazykov a v súčasnosti je jeden z najpoužívanejších webových prekladačov. Štatistika z roku 2013 uvádza, že deňne Google translate využilo cez 200 miliónov ľudí.

Google translate aplikuje princíp štatistického strojového prekladu. Google translate neprekladá texty priamo z jedného jazyka do druhého, ale používa angličtinu ako medzijazyk. Niektoré jazyky sa dokonca prekladajú cez viacero jazykov do angličtiny a až následne do cieľového jazyka. Takto je to aj v prípade slovenčiny, ktorá sa prekladá do češtiny z nej do angličtiny a až na záver do cieľového jazyka.

	WMT 2011	
Google translate	43.05	

Tabuľka 4.16: Kvalita prekladu systému Google translate zisťená pomocou BLEU

4.6.2 Česílko

Je prekladový systém pre príbuzné jazyky akými sú slovanské jazyky. Systém je plne schopný prekladu z češtiny do slovenčiny. Systém využíva priameho prekladu slovo na slovo a bol vytvorený hlavne ako pomocný nástroj pri ručnom preklade.

	WMT 2011	
Česílko	38.80	

Tabuľka 4.17: Kvalita prekladu systému Česílko zisťená pomocou BLEU

4.7 Chyby

Na hodnotenie BLEU vplývajú rôzne nezhody v ručnom preklade a strojovom. V tejto podkapitole uvediem niektoré z nich, na ktoré som natrafil pri ručnej kontrole prekladu. Príklady strojového prekladu sú výstupom systému AIO5 + tuning set 3.

Častou nezhodou, ktorú bleu ráta ako chybu prekladu, je inak zapísaná veta avšak s rovnakým významom. Česká veta: *Ze slovníčku, jehož výrazy deník Aktuálně.cz zveřejnil letos v červnu.* Ručný preklad: *Zo slovníka zverejneného denníkom Aktuálně.cz v júni tohto roku.* Strojový preklad: *Zo slovníčka, ktorého výrazy denník Aktuálně.cz uverejnil v júni tohto roku.* Je vidno, že veta nie je preložená nesprávne, ale aj tak bude vyhodnotená ako nesprávna pretože sa nezhoduje s ručným prekladom.

Ďalšími chybami sú chyby z nejednoznačnost. Česká veta: *Na škole mi říkali Bříza.* Ručný preklad: *V škole ma volali Breza.* Strojový preklad: *Na škole mi povedali Bříza.* V češtine môže mať slovné spojenie "na škole"dva významy v zmysle spojenom s dochádzkou do danej inštitúcie alebo vo význame na škole ako budove. V slovenčine sa toto rozlišuje. Keď myslíme na dochádzku použijeme "v"a ak myslíme na budovu tak "na". Podobne je to aj so slovom "říkali". V tejto vete sa vyskytuje aj chyba v slove "Bříza", ktoré je síce vlastným podstatným menom, ale je zároveň preložiteľné do slovenčiny vo význame stromu "breza". Je pravdepodobné, že takýto prípad v tréningovom texte nebol a preto systém nemá vo svojich prekladových tabuľkách záznam o tom, ako ho preložiť a preto ho ponechá

v českom tvare. V tomto prípade to nemusí byť chyba, nakoľko je to vlastné podstatné meno, avšak tento typ chyby, keď sa k slovu nepodarý nájsť žiaden preklad, je jeden z najčastejších a práve tieto sú riešiteľné väčšími korpusmi.

4.8 Porovnanie systémov

Na záver uvádzam tabuľku porovnania najlepšieho vytvoreného systému a existujúcich prekladových systémov.

	WMT 2011	
AIO5 + tuning set3	50.76	
Google translate	43.05	
Česílko	38.80	

Tabuľka 4.18: Tabuľka porovnania kvality prekladu systémov v teste BLEU

Kapitola 5

Záver

Zadaním tejto bakalárskej práce bolo zoznámiť sa s metódami používanými pre automatický strojový preklad, navrhnúť a implementovať systém pre preklad českých textov do slovenčiny a na záver vyhodnotiť úspešnosť vytvoreného systému pomocou štandardných metrík.

V tejto práci sú popísané rôzne techniky tvorby prekladových systémov, ich závyslosti na vstupných dátach, techniky obdržania paralelných korpusov a nástroje pre ich spracovanie. Špeciálna časť venujúca sa postupu použitej metódy.

V rámci tohoto projektu bolo vytvorených niekoľko prekladových systémov založených na princípe štatistického strojového prekladu, odlišujúcich sa použitými vstupnými dátami a dátami použitými pri ich ladení. Pri tvorbe tréningových korpusov boli použité aj slovníky, čo sa výrazne odrazilo na zvýšení kvality prekladu, ak bol systém následne odladený. Pre ladenie boli použité 3 rôzne kombinácie dát pričom najlepší výsledok vykazovali systémy odladené pomocou beletristických dát. Ohodnotenie systémov bolo vykonané štadardným nástrojom pre meranie kvality prekladu BLEU. Úspešne bol vytvorený systém s nameranou hodnotou kvality prekladu 50.76 (AIO5 + tunning set3). Preklad bol následne porovnaný s kvalitou prekladu Google transalate, oproti ktorému môj systém získal v teste BLEU o 7.71 bodov viac. Ďalej bol systém porovnaný s prekladom na systéme Česílko, oproti ktorému je môj systém lepši o 11.96 bodov. Pri testovaní bol použitý jednotný postup na rovnakých testovacích dátach.

Literatura

[1] Christopher D. Manning, H. S.: *Foundations of statistical natural language processing*. The Mit press, 1999, iSBN 0-262-13360-1.

[2] Commission, E.: The Acquis Communautaire multilingual parallel corpus and Eurovoc (v 2.2).
URL http://optima.jrc.it/Acquis/JRC-Acquis.2.2/doc/README_Acquis-Communautaire-corpus_JRC.html

[3] D. Varga, P. H., L. Németh: Parallel corpora for medium density languages. In *In Proceedings of the RANLP*, 2005, s. 590–596.

[4] Galuščáková, P.; Bojar, O.: Czech-Slovak Parallel Corpora. In *Proc. of Slovko 2011*, October 2011.

[5] Koehn, P.: Europarl: A Parallel Corpus for Statistical Machine Translation.
URL http://homepages.inf.ed.ac.uk/pkoehn/publications/europarl-mtsummit05.pdf

[6] Koehn, P.: *Moses, Statistical machine translation system, User manualand code guide*. University of Edinburgh, May 2015.
URL http://www.statmt.org/moses/manual/manual.pdf

[7] M. Federico, M. C., N. Bertoldi: *IRST Language Modeling Toolkit*. FBK-irst, Trento, Italy, September 2008.
URL http://hermes.fbk.eu/people/bertoldi/teaching/lab_2010-2011/img/irstlm-manual.pdf

[8] Michalička, J.: *Štatistický strojový preklad veľmi blízkych jazykov*. Diplomová práce, Univerzita Komenského Fakulta matematiky, fyziky a informatiky, 2005.

[9] Nazar, R.: Parallel corpus alignment at the document, sentence and vocabulary levels. Technická zpráva, IULA, Universitat Pompeu Fabra, Roc Boronat 138, 08018, Barcelona, 2011.
URL http://www.upf.edu/pdi/iula/rogelio.nazar/735.pdf

[10] Och, F. J.; Ney, H.: Improved Statistical Alignment Models. Hongkong, China, October 2000, s. 440–447.

[11] Papineni, K.; Roukos, S.; Ward, T.; aj.: BLEU: a Method for Automatic Evaluation of Machine Translation. 2002, s. 311–318.

[12] Tiedemann, J.: Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, editace N. C. C. Chair); K. Choukri; T. Declerck; M. U. Dogan; B. Maegaard; J. Mariani; J. Odijk; S. Piperidis, Istanbul, Turkey: European Language Resources Association (ELRA), may 2012, ISBN 978-2-9517408-7-7.

[13] W. John Hutchins: *An introduction to machine translation*. Academic Press, 1992, iSBN 0-12-362830-X.

[14] web: Slovenský národný korpus.
URL http://korpus.juls.savba.sk/

[15] web: stránka EMEA.
URL http://opus.lingfil.uu.se/EMEA.php

[16] web: stránka GIZA++.
URL http://www-i6.informatik.rwth-aachen.de/Colleagues/och/software/GIZA++.html

[17] web: stránka GMA.
URL http://nlp.cs.nyu.edu/GMA/

[18] web: stránka HunaAlign.
URL http://mokk.bme.hu/en/resources/hunalign/

[19] web: stránka Intercorp.
URL https://ucnk.ff.cuni.cz/intercorp/

[20] web: stránka NATools.
URL http://corpora.di.uminho.pt/natools/

[21] web: stránka OPUS.
URL http://opus.lingfil.uu.se/

[22] wikipedia: Machine translation.
URL http://en.wikipedia.org/wiki/Machine_translation