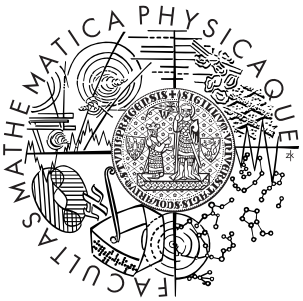
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Adam Liška

Čištění paralelních dat pro strojový překlad

Ústav formální a aplikované lingvistiky

Vedoucí bakalářské práce: RNDr. Ondřej Bojar, PhD
Studijní program: Obecná informatika

2010

Mé poděkování za podnětné rady a připomínky patří vedoucímu mé práce Ondřeji Bojarovi.

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

V Praze dne 28.5.2010 Adam Liška

Obsah

Název práce: Čištění paralelních dat pro strojový překlad
Autor: Adam Liška
Katedra: Ústav formální a aplikované lingvistiky
Vedoucí bakalářské práce: RNDr. Ondřej Bojar, PhD
e-mail vedoucího: bojar@ufal.mff.cuni.cz

Abstrakt: Cílem této práce je návrh, implementace a ruční vyhodnocení filtrů pro čištění paralelních dat, se zaměřením na statistický strojový překlad. Dalším výsledkem je vytvoření anotovaných sad paralelních česko-anglických textů pro využití při vývoji filtrů v budoucnosti. Vyvinuto bylo i několik nástrojů pro práci s těmito sadami a pro automatické vyhodnocení výstupu filtrů.

Klíčová slova: paralelní korpusy, filtrování korpusů, čištění paralelních dat, strojový překlad

Title: Cleaning of Parallel Texts for Machine Translation
Author: Adam Liška
Department: Institute of Formal and Applied Linguistics
Supervisor: RNDr. Ondřej Bojar, PhD
Supervisor's e-mail address: bojar@ufal.mff.cuni.cz

Abstract: The aim of the thesis is to design, implement and manually evaluate filters for parallel data cleaning, focused on statistical machine translation. Annotated sets of parallel texts to be used during development of new filters in the future are another result of this work. Several tools facilitating work with these sets and allowing for automatic evaluation of filter outputs are also developed.

Keywords: paralel corpora, corpus filtering, cleaning of parallel texts, machine translation

Kapitola 1

Úvod

Paralelní data, t.j. texty dostupné současně ve více jazycích, jsou důležitým zdrojem v mnoha oblastech komputační lingvistiky. Zcela nepostradatelnou roli hrají ve statistickém strojovém překladu – systém se z dat sám naučí, jak obsažené fráze překládat. Fráze, které se v těchto datech nevyskytují, ovšem přeložit nedokáže.

Bylo ukázáno, že větší množství paralelních dat zlepšuje kvalitu strojového překladu ([1] a [2]), stejně tak ovšem jejich čistota, tj. absence chybných párů ([3]). Ke zlepšení kvality strojového překladu pracujícího s jedním či dvěma jazyky je tedy velmi důležité, abychom byli schopni získat velké množství paralelních dat a zároveň uměli odfiltrovat chybné větné páry.

Ačkoli publikace se v oblasti filtrování paralelních dat vyskytují (např. [3]), přímo pro český jazyk neexistují. Zároveň není často hodnocena přesnost a pokrytí jednotlivých filtrů.

V této práci je navržena sada filtrů a tyto filtry jsou vyhodnoceny, jednotlivě i v součinnosti. Jako evaluační míry byly zvoleny standardní veličiny přesnost (*precision*) a pokrytí (*recall*). Částečné výsledky byly publikovány již v článku [4].

Text je členěn následovně:

- v kapitole 2 je popsána výchozí situace, motivace a popsán princip statistického strojového překladu,
- kapitola 3 ukazuje různé možnosti filtrování větných párů a jsou v ní popsány nové filtry,

- v kapitole 4 je shrnuto hodnocení jednotlivých filtrů a jejich přínos pro čištění paralelních dat, a zároveň popsáno prostředí pro hodnocení filtrů,
- poslední kapitola 5 shrnuje výsledky práce a vytyčuje cíle do budoucnosti,
- příloha A popisuje použití filtrů z uživatelského hlediska,
- příloha B popisuje filtry z pohledu programátora.

Kapitola 2

Výchozí situace

2.1 Statistický strojový překlad

Překlad textů mezi různými jazyky je velice žádaný úkol a jeho důležitost v poslední době roste i s rozšiřováním Internetu, díky kterému mají jeho uživatelé přístup k informacím v mnoha jazycích.

Možnosti počítačů v oblasti překladu jsou studovány již několik desetiletí a obecně se počítačem zpracovávaný překlad nazývá strojový překlad. Existují dva hlavní přístupy ke strojovému překladu. První využívá pravidla navržená lingvisty, druhý je založený na strojovém učení a využívá statistické metody a soubory paralelních dat.

V současné době se podstatná část výzkumu zaměřuje na překlad pomocí statistických metod, které dosáhly velkých kvalitativních zlepšení ([1]). Tato orientace s sebou přináší několik výhod. Nástroje potřebné k vytvoření překladového systému jsou povětšinou volně přístupné[1] a v základu jsou jazykově neutrální, tj. nezávislé na vybraném jazykovém páru. Jednoduchý překladový systém lze tedy vytvořit během krátké doby bez hlubších znalostí problematiky.

Úlohou statistického strojového překladu je k větě ve zdrojovém jazyce vytvořit takový překlad do cílového jazyka, který má největší „pravděpodobnost“. Jsou to právě soubory paralelních dat, pomocí kterých jsou vypočítány jednotlivé pravděpodobnosti. Překladový systém na základě parametrů získaných z těchto souborů navrhuje věty v cílovém jazyce a vybírá nejlepší překlad. Během toho využívá různých heuristik, jelikož prostor možných překladů je velice rozsáhlý.

[1] http://www.statmt.org/moses; http://sourceforge.net/projects/joshua

2.2 Paralelní korpusy

Soubory paralelních textů se nazývají paralelní korpusy. V současné době jsou zpracovávány a uloženy zásadně v elektronické podobě, a jsou proto velmi vhodné pro různé statistické analýzy.

Na Internetu je k dispozici mnoho hotových paralelních korpusů pro velké jazyky (angličtina, francouzština atp.). Zdroje pro jejich tvorbu se liší, často ovšem obsahují texty oficiálních dokumentů mezinárodních organizací. Pro překladový systém Google Translate[2] společnosti Google byly například jako základ využity dokumenty Organizace spojených národů, přeložené do všech úředních jazyků organizace: arabštiny, čínštiny, angličtiny, francouzštiny, ruštiny a španělštiny ([1]). Oblíbené jsou také dokumenty Evropského parlamentu[3].

Přípravu paralelního korpusu lze zjednodušeně rozdělit do tří fází:

- získávání paralelních textů;
- větné zarovnání textů;
- zpracování získaných větných párů (přidání lemmatizace atp.).

V případě paralelních korpusů obsahujících jeden či více menších jazyků bývá nejtěžším úkolem první krok. Potřebné texty se získávají z velkého množství zdrojů a ty s sebou přinášejí různé chyby, zejména neodpovídající či neúplné překlady. Tyto je třeba odfiltrovat.

2.3 CzEng 0.9 a příprava korpusu

CzEng 0.9[4] je rozsáhlý česko-anglický korpus vyvíjený na Ústavu pro formální a aplikovanou lingvistiku Matematicko-fyzikální fakulty Univerzity Karlovy v Praze. Korpus se skládá z několika typů textů:

- filmové a seriálové titulky;
- paralelní internetové stránky;
- beletrie;

[2] http://translate.google.com
[3] http://wt.jrc.it/lt/Acquis/
[4] http://ufal.mff.cuni.cz/czeng/czeng09/

- zákony a normy Evropské unie;
- technické dokumentace;
- novinové texty;
- uživatelské překlady k projektu Navajo[5].

Postup zpracování je pro všechny dokumenty v korpusu CzEng 0.9 stejný ([5]). Nejprve jsou převedeny do formátu UTF-8. Segmentace je prováděna trénovatelným tokenizérem a větné zarovnání programem Hunalign[6].

Ze zarovnaných párů jsou ponechány jen zarovnání 1-1, která představují přibližně 82% všech zarovnání. Toto je první významný krok filtrace. Následně je na česko-anglické páry aplikována první sada filtrů, která má za cíl odfiltrovat špatně zarovnané či jinak nevhodné páry. Aby byl větný pár přidán do korpusu, musí projít všemi filtry.

I přes použití výše zmíněných filtrů obsahuje korpus stále chybné páry, a proto je třeba vyvíjet filtry nové. Staré i nové filtry jsou popsány v následující kapitole. Jejich ohodnocení se nachází v kapitole 4.

[5]http://www.navajo.cz
[6]http://mokk.bme.hu/resources/hunalign

Kapitola 3

Možnosti filtrování větných párů

Tato kapitola shrnuje starší filtry z korpusu CzEng 0.9 a podrobně popisuje nové filtry implementované v této práci.

3.1 Nové a původní filtry

Nejvíce aktivními filtry implementované před vydáním korpusu CzEng 0.9 jsou mimo jiné následující ([4]):

- český i anglický segment jsou identické řetězce (např. nepřeložený text z internetových stránek),
- významný rozdíl v délkách vět (obvykle kvůli špatné segmentaci či větnému zarovnání),
- v českém segmentu se nevyskytuje české slovo či naopak v anglickém segmentu se nevyskytuje anglické slovo (s využitím seznamu slov z Českého a Britského národního korpusu),
- výskyt podezřelého znaku či sekvence opakujících se znaků,
- pozůstatky HTML tagů,
- různá záhlaví a zápatí legislativních textů EU, textů z Project Gutenberg atp.,
- délka věty přesahuje daný limit.

Na základě analýzy špatných párů, které ovšem prošly všemi výše zmíněnými filtry, v této práci navrhujeme následující filtry.

- Číselný filtr se snaží větné páry filtrovat na základě použití čísel v obou větách. Pro všechna čísla v anglickém segmentu se snaží najít ekvivalent v segmentu českém.
- Častá chyba objevující se v korpusu CzEng vzniká již při získávání paralelních textů - některé z nich jsou jen z části dvojjazyčné a obsahují na anglické straně nepřeložené české segmenty (tato chyba je obzvláště častá u „dvojjazyčných“ internetových stránek). Tyto odstraní filtr, který se snaží veškeré znaky mimo znakovou sadu ASCII v anglickém segmentu potvrdit na české straně.
- Jinou možností, jak přistupovat k problému popsanému v předchozím bodě, je zkontrolovat, zdali anglický segment neobsahuje pro anglický jazyk atypické sekvence znaků. Z anglických textů se spočítá statistika výskytu n-gramů písmen a pomocí této statistiky filtr ohodnotí jednotlivé anglické věty.
- Pro každý větný pár je možné určit slovní zarovnání a jeho pravděpodobnost, na jejímž základě je možné porovnávat *pravděpodobnost překladu* jednotlivých větných párů.
- Další metoda určuje podíl slov v české větě, které jsou překladem některého ze slov v anglickém segmentu, k počtu všech slov v české větě.

Jediný filtr specifický pro konkrétní jazyk (v tomto případě pro angličtinu) je filtr pracující se znakovou sadou ASCII. Ostatní filtry jsou jazykově nezávislé a pro práci s jinými jazykovými páry nejsou nutné žádné modifikace.

3.2 Číselný filtr

Implementace tohoto filtru je vcelku přímočará. Filtr nejprve načte všechna čísla z anglické věty a hledá jejich ekvivalenty v české větě. Jako ekvivalent se ovšem nepovažuje pouze identické číselné vyjádření, ale i slovní opis – často je totiž číselný výraz v anglické větě vyjádřen v české větě slovem. Filtr tedy využívá externí slovník těchto překladů.

V mnoha případech bývá formát čísel nekonzistentní (např. v českém segmentu se vyskytují mezery mezi trojicemi čísel – `6 049` – a v anglickém nikoliv). Pokud v anglickém segmentu existuje číslo, ke kterému se filtru nepodařilo najít ekvivalent v českém segmentu, nemusí to znamenat, že je dané zarovnání špatné. Je možné, že formát čísel byl nekonzistentní. Proto v tomto případě mohou nastat dvě situace:

- v české i anglické větě jsou použity stejné cifry a pár je označen jako správný,
- v české a anglické větě jsou použity různé cifry a pár je označen jako špatný.

3.3 Použití znaků mimo znakovou sadu ASCII

Implementace tohoto filtru je také přímočará. Z anglické věty jsou pomocí regulárního výrazu načteny všechny znaky mimo znakovou sadu ASCII (s výjimkou pomlček, různých typů uvozovek a znaku evropské měny). Filtr vychází z předpokladu, že anglický text nepotřebuje až na výše uvedené výjimky žádné znaky mimo ASCII. Jedině v případě, že by uváděl cizojazyčné vlastní jméno – ovšem to by mělo být uvedeno i v českém ekvivalentu dané věty.

Filtr tedy uvažuje dvě možnosti:

- všechny znaky mimo ASCII jsou potvrzeny v české větě a pár je označen jako správný,
- v anglické větě existují znaky mimo tabulku ASCII, které se nevyskytují v české větě, a pár je označen jako špatný.

3.4 N-gramy písmen

Tento filtr nejprve z dostatečně dlouhého anglického textu získá pravděpodobnosti bigramů a trigramů znaků. Pro získávání těchto modelů a práci s nimi je použita sada nástrojů SRILM[1].

[1] http://www-speech.sri.com/project/srilm

Nechť $e_1...e_n$ je anglická věta. Ta se pomocí získané statistiky ohodnotí tímto způsobem:

$$P(e_1^n) = \prod_1^n P(e_k|e_1^{k-1})$$

Aby bylo možné porovnávat získané hodnoty, je třeba je znormalizovat. Skóre věty, které se porovnává s určeným limitem, je následující:

$$score(e_1^n) = \frac{\log P(e_1^n)}{n}$$

Zde ovšem dochází ke zkreslení při krátkých segmentech, proto všechny segmenty menší než 35 znaků jsou automaticky označeny jako správné a filtr se zabývá jen delšími segmenty. Všechny segmenty se skóre nižším než daný limit jsou označeny jako špatné. Pravděpodobnosti bigramů a trigramů znaků byly získány z části `news` korpusu CzEng 0.9, která je relativně čistá.

3.5 Pravděpodobnost překladu

Tento filtr určí pravděpodobnost překladu daného segmentu, porovná ji s limitní hodnotou, a pokud je nižší, označí daný segment jako špatný.

Pravděpodobnost překladu je získána na základě pravděpodobnosti slovního zarovnání dvojjazyčného páru. Slovní zarovnání mezi dvěma větami určuje korespondenci jednotlivých slov v překladu. Výsledkem je bipartitní graf, jehož vrcholy jsou slova z jednotlivých vět a jehož hrany spojují ta slova, která si v překladu odpovídají.

Pro natrénování slovního zarovnání je použito standardního nástroje v této oblasti, GIZA++[2]. Vstupem pro tento proces je paralelní korpus. Výstupem je kolekce souborů, z nichž jeden (`A3.final`) popisuje slovní zarovnání jednotlivých vět a pravděpodobnost tohoto zarovnání pro každou větu.

Například pro větný pár *Iraq's voters have spoken. - Iráčtí voliči promluvili.* má výstup následující tvar:

```
# Sentence pair (548) source length 4 target length 6 alignment score
: 4.84659e-06
```

[2]http://code.google.com/p/giza-pp

```
Iraq 's voters have spoken .

NULL ({ }) Iráčtí ({ 1 2 }) voliči ({ 3 }) promluvili ({ 4 5 }) .
({ 6 })
```

První číslo určuje pořadové číslo větného páru v korpusu, dále jsou uvedeny délky obou vět a skóre (pravděpodobnost) daného zarovnání. V tomto případě je česká věta zdrojovou větou a anglická věta cílovou. Každé slovo ze zdrojové věty může být zarovnáno na žádné, jedno či více slov v cílové větě; na každé slovo v cílové větě může být zarovnáno nanejvýše jedno slovo ze zdrojové věty.

V předcházejícím příkladě bylo zarovnání bezchybné, např. české slovo „voliči“ skutečně odpovídá třetímu anglickému slovu „voters“. Kvalita zarovnání ovšem závisí na mnoha faktorech, jedním z nichž je např. velikost korpusu. Většinou se v získaném zarovnání vyskytují chyby, jako v následujícím příkladě:

```
# Sentence pair (106) source length 5 target length 7 alignment score
: 7.00829e-10

But the environment has been suffering .

NULL ({ 2 }) Životní ({ 1 4 5 }) prostředí ({ 3 }) ovšem ({ }) trpí
({ 6 }) . ({ 7 })
```

Slovo „Životní“ je zarovnáno na slova „But“, „has“, a „been“. Získané skóre nicméně i tak koreluje s *pravděpodobností překladu*. Pro získání přesnějších výsledků se zarovnání získají pro oba směry a v kombinaci dávají výsledné skóre pro daný větný pár:

$$Score\left(e_1^J, f_1^I\right) = \frac{1}{J}\log\left(p\left(\boldsymbol{e}, a \mid \boldsymbol{f}\right)\right) + \frac{1}{I}\log\left(p\left(\boldsymbol{f}, a \mid \boldsymbol{e}\right)\right)$$

Fungování filtru lze popsat následovně:

- pomocí nástroje GIZA++ se získají slovní zarovnání pro oba směry,
- každému slovnímu páru se na základě jeho slovních zarovnání přiřadí skóre pravděpodobnosti překladu,
- empiricky se určí hranice a všechny slovní páry s nižším skóre se označí jako špatné.

3.6 Slovníkový filtr

Tento filtr pomocí slovníku získá překlady slov v anglické větě a určí podíl českých slov, která jsou překladem nějakého slova v anglické větě, k počtu všech slov v české větě. Pokud je tento podíl příliš nízký, je daný větný pár označen jako špatný.

U tohoto filtru je důležité používat slovník obsahující co nejvíce slovních párů. Jako základní slovník byl použit anglicko-český slovník přístupný pod licencí GNU Free Documentation License na adrese `slovnik.zcu.cz`.

Slovník je ovšem možné získat i ze slovního zarovnání. Jedním z výstupů programu GIZA++ je soubor `t3.final`, který obsahuje překlady slov a jejich pravděpodobnosti. Tento soubor nicméně obsahuje mnoho chybných slovních párů a těch je třeba se nejprve zbavit. Prvním krokem je natrénovat slovní zarovnání v obou směrech a použít jen překlady, které jsou potvrzeny v obou směrech. Tato množina je dále zmenšena (a vyčištěna) vybráním jen jednoho, nejpravděpodobnějšího překladu pro každé anglické slovo. Pro natrénování slovního zarovnání byly použity větné páry z části `news` korpusu CzEng 0.9.

Aby nebylo nutné zabývat se bohatou morfologií českého jazyka, je třeba před použitím tohoto filtru provést lemmatizaci větných páru. Ze stejného důvodu je i slovní zarovnání natrénováno na lemmatech.

Kapitola 4

Hodnocení filtrů

4.1 Prostředí pro hodnocení filtrů

Během vývoje a hodnocení filtrů bylo připraveno prostředí s anotovanými texty a nástroji pro automatické hodnocení filtrů. Součástí tohoto prostředí jsou čtyři sady paralelních textů, z nich tři jsou anotovány.

4.1.1 Soubory paralelních textů

Všechny sady jsou vybrány z korpusu CzEng 0.9 a jsou ve formátu CzEng Export Format. Každý větný pár je reprezentován jedním řádkem, který daný větný pár popisuje: obsahuje lemmatizaci, morfologické značky apod. Pro podrobný popis formátu, viz [5]. Pokud daná sada obsahuje anotaci, je obsažena v samostatném souboru s příponou `anot` v následujícím formátu:

anotace anglický_segment český_segment

Anglické a české segmenty obsahují pouze text větných párů. Řádky těchto dvou souborů si navzájem odpovídají. Jednotlivá pole jsou od sebe oddělena tabulátorem. Pro anotaci existují pouze dvě značky: *x* pro špatné zarovnání, *ok* pro správné zarovnání.

Sada `devset` obsahuje 1000 anotovaných vět. Tuto sadu je možno využít při vývoji nových filtrů, nastavování jejich parametrů atp.

Sada `testset` je složena z 2200 vět obsažených v korpusu CzEng 0.9. Tato sada byla využita při konečném vyhodnocení jednotlivých filtrů.

Sada `trainset` obsahuje 100 000 vět, které nejsou anotovány. Z tohoto souboru jsou brány dodatečné větné páry, jež po doanotování umožní lepší odhad přesnosti filtrů.

Anotované páry ze sady `trainset` jsou poté přesunuty do zvláštní sady `extra`. Tento soubor je anotovaný, nicméně distribuce jednotlivých typů zdrojů je jiná než v korpusu CzEng 0.9. Zkreslení vzniká tím, že obsažené větné páry byly alespoň jedním filtrem označeny jako špatné a anotovány byly z důvodu přesnějšího vyhodnocení filtrů.

4.1.2 Popis nástrojů

Součástí tohoto prostředí je kolekce nástrojů napsaných v jazyce Java, obsažených v balíčku `filtrum`.

Program `GetStats` vyhodnocuje precision (přesnost) a recall (pokrytí) filtrů. Tyto dvě metriky jsou definovány následovně:

$$precision = \frac{tp}{tp + fp},$$

$$recall = \frac{tp}{tp + fn},$$

kde *tp* představuje *true positives*, tedy špatné větné páry označené filtrem jako špatné, *fp* představuje *false positives*, tedy správné větné páry označené filtrem jako špatné, a konečně *fn* představuje *false negatives*, tedy špatné větné páry označené filtrem jako správné.

Program čte ze standardního vstupu data následujícího formátu:

anotace segment1 segment2 $chyba1|chyba2|chyba3|...$

Jednotlivá pole jsou od sebe oddělena tabulátorem. Výstup programu je na standardní výstup a představuje jej tabulka obsahující pro každý filtr jeho název, počet párů, které označil jako špatné, a odpovídající hodnotu precision a recall.

Zároveň jsou filtry vyhodnoceny i v součinnosti. Kombinovaný filtr označuje jako špatné takové páry, jež alespoň jeden z filtrů označil jako špatné. To dává uživateli informaci o celkové přesnosti a pokrytí dané sady filtrů. Tyto informace jsou ve výstupu uvedeny v řádku `combined`.

Druhým nástrojem je program `Combine`, který kombinuje výstupy několika filtrů na stejných datech do jednoho souboru. Jako parametry

spuštění očekává soubory výstupů jednotlivých filtrů, výstup je posílán na standardní výstup.

Často se stává, že filtr na dané anotované sadě vět označí jako špatné jen malé množství vět. To postačuje pro určení pokrytí filtru, nicméně pro bližší určení přesnosti je třeba doanotovat další věty. Pro tento případ byl vyvinut nástroj `SelectForAnnotation`, který při spuštění vyžaduje dva číselné argumenty následované seznamem sledovaných chyb. Ze standardního vstupu čte data stejného formátu jako program `GetStats`, z nichž některé řádky přeposílá na standardní výstup. První číselný argument určuje počet řádků od počátku, které mají být přímo přeposlány na standardní výstup. Pro všechny sledované filtry je udržována informace o počtu řádků, jež označily jako špatné. Jakmile jsou všechny počáteční řádky přeposlány, následující řádky jsou přeposílány pouze v případě, pokud byly označeny alespoň jedním z určených filtrů, jehož počet označených řádků stále nedosáhl na druhý číselný parametr. Jakmile všechny filtry dosáhly tohoto limitu, žádné další řádky již nejsou na výstup posílány.

Pomocí nástroje `ParameterStats` lze vypočítat hodnoty precision a recall pro různé hodnoty dolního prahu skóre. Lze jej tedy využít pro všechny filtry, které jednotlivým větným párům přiřazují jisté skóre a určují dolní mez, stanovující hranici mezi dobrými a špatnými větnými páry. Na standardním vstupu program čte data ve formátu:

skóre anotace

kde tato dvě pole jsou oddělena tabulátorem. Vstup musí být setříděný podle hodnoty skóre. Filtr předpokládá závislost „čím vyšší skóre, tím kvalitnější větný pár“.

Posledním nástrojem je `ExportFormatProcessor`, pomocí kterého lze zpracovávat CzEng Export Format a vybírat z něj pouze potřebná data. Data čte ze standardního vstupu a výstup se ovládá následujícími přepínači:

- `-plain` – výstupem je text segmentů;
- `-lemma` – výstupem je lemmatizovaný text segmentů;
- `-pseudolemma` – výstupem jsou pseudolemmata;
- `-tag` – výstupem jsou morfologické značky.

Výběr filtrů	Precision	Recall
Nedostatek písmen	94%	7%
Neodpovídající si délky	91%	11%
Opakující se znak	88%	2%
Absence anglického slova	80%	11%
Podezřelý znak	75%	1%
Segmenty jsou identické	72%	26%
Absence českého slova	67%	2%
Příliš dlouhá věta	12%	0%
Nadbytečné záhlaví	2%	0%
Celkem (všechny filtry)	57%	42%
Celkem (anotované filtry)	57%	41%

Tabulka 4.1: Ruční ohodnocení filtrů použitých při přípravě CzEng 0.9.

4.2 Vyhodnocení filtrů

V této části vyhodnocujeme starší a nově implementované filtry proti ruční anotaci. Použity jsou metriky precision a recall, které byly popsány v části 4.1.2.

Postup vyhodnocení je následující. Filtry byly nejprve spuštěny na základní sadě anotovaných vět. Zde byla určena hodnota jejich pokrytí. Jelikož počty segmentů, které jednotlivé filtry označily jako špatné, se lišily a pro některé filtry byly nízké, bylo třeba doanotovat další segmenty tak, aby pro každý filtr byl anotován stejný počet segmentů, jež označil jako špatné. Na těchto rozšířených sadách byla určena přesnost filtrů. Konkrétní velikosti těchto sad jsou uvedeny v následujících pododdílech.

4.2.1 Filtry použité v CzEng 0.9

Filtry již použité při přípravě korpusu CzEng 0.9 byly ohodnoceny na náhodně vybraných 1000 větných párech získaných zpracováním zdrojů korpusu. Některé filtry na této sadě dat označily jako špatné pouze např. tři segmenty, a proto bylo třeba pro získání hodnoty precision evaluační sadu rozšířit. Pro tuto dodatečnou anotaci byly vybrány pouze filtry, které označily alespoň jeden větný pár v původní sadě jako špatný ([4]). Výsledné hodnocení je uvedeno v tabulce 4.1.

Zde je třeba zdůraznit, že nízký recall u jednotlivých filtrů byl očekáván, protože cílem je vytvářet vysoce přesné filtry, které až v *součinnosti*

Filtr	Precision	Recall
Pravděpodobnost překladu (hranice -9.5)	81%	24%
Slovníkový filtr (pokrytí alespoň 0.25)	77%	25%
Číselný filtr	86%	4%
N-gramy písmen (hranice -1.5)	62%	5%
Použití znaků mimo ASCII	82%	5%
Celkem (všechny filtry)	74%	46%

Tabulka 4.2: Ruční ohodnocení nových filtrů

(t.j. segment je označen jako špatný, pokud jej alespoň jeden z filtrů označil jako špatný) dosahují vysokého pokrytí.

Celkové pokrytí filtrů je necelých 60%. Jeden z nejméně přesných filtrů je filtr označující všechny věty delší než 400 slov jako špatné, ačkoliv ve většině případů tyto segmenty byly dle manuální anotace správné. Nicméně strojový překlad natrénovaný na CzEng 0.9 tímto ovlivněn není, jelikož např. nástroje pro slovní zarovnání takto dlouhé věty nezpracovávají.

Nejvyšší pokrytí vykazuje filtr označující identické segmenty jako špatné. V mnoha případech je ponechání identických segmentů korektní, pokud např. obsahují jen jméno či vlastní název. Jejich případné odstranění ovšem opět strojový překlad neovlivní, jelikož většina systémů neznámá slova na vstupu použije ve výstupu v nezměněné podobě.

4.2.2 Nové filtry

Postup hodnocení nových filtrů byl obdobný jako u filtrů již použitých při přípravě CzEng. Recall jednotlivých filtrů byl určen na 2200 anotovaných řádcích sady `testset`. Sada `testset` je vybrána z korpusu CzEng 0.9, a proto obsahuje pouze segmenty, které nebyly označené staršími filtry jako špatné.

Pro určení přesnosti filtrů bylo taktéž potřebné rozšířit tuto sadu i o větné páry z části `trainset`, aby celkový počet ručně anotovaných řádků, které byly filtrem označeny jako špatné, dosáhl 200.

Filtry byly spuštěny v základním nastavení, to jest:

- filtr pravděpodobnosti překladu byl spuštěn na lemmatizovaných segmentech, se spodním limitem skóre -9.5;

- slovníkový filtr také zpracovával lemmata a použil oba slovníky, tj. slovník z `http://slovnik.zcu.cz` a slovník vygenerovaný ze slovního zarovnání části `news` korpusu CzEng 0.9; spodní práh pokrytí slov je jedna čtvrtina;
- číselný filtr vyžadoval pokrytí všech číselných výrazů v anglické větě;
- n-gramový filtr posuzoval pouze anglické segmenty delší než 35 znaků a spodní limit skóre byl -1.5;
- filtr porovnávající použití znaků mimo tabulku ASCII vyžadoval pokrytí všech znaků mimo tabulku ASCII z anglického segmentu v segmentu českém.

Parametry filtrů byly nastaveny s pomocí sady `devset`. Výsledné hodnocení se nachází v tabulce 4.2.

Z výsledků je vidět, že s výjimkou filtru počítajícího statistiky n-gramů mají všechny nové filtry přesnost okolo 80%. Nejvyšší recall mají slovníkový filtr a filtr sledující pravděpodobnost překladu. Ostatní tři filtry mají recall okolo pěti procent.

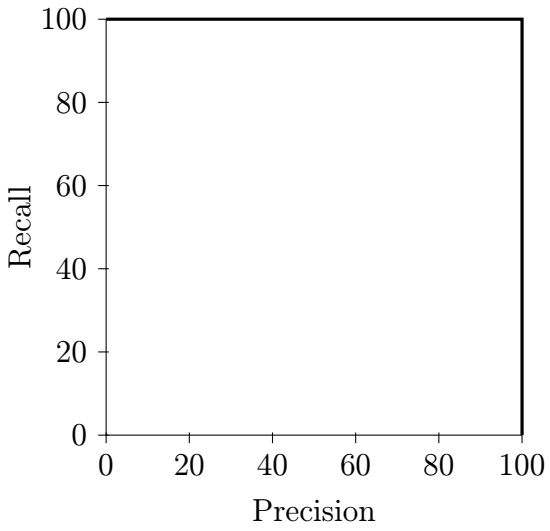
4.2.3 Vyhodnocení filtrů pomocí ROC křivek

Tři filtry z této kolekce - konkrétně filtr pravděpodobnost překladu, slovníkový a n-gramový filtr - jsou závislé na parametru, jež představuje nejnižší možné skóre větného páru, který je ještě považován za správný. Z toho důvodu je možné na testovacích datech vyčíslit, jaké hodnoty přesnosti a pokrytí dané filtry při konkrétních hodnotách limitů dávají, a díky této informaci stanovit nejvhodnější práh.

Křivka vzájemné závislosti přesnosti a pokrytí pro ideální filtr je zobrazena na obrázku 4.1.

Vzhledem k orientaci na přesnost (i při nižším pokrytí) lze jako vhodný tvar považovat i křivku na obrázku 4.2.

ROC křivky pro filtr pravděpodobnosti překladu, slovníkový filtr a n-gramový filtr jsou zobrazeny na obrázcích 4.3, 4.4 a 4.5. Tato data byla získána na sadě `testset` až po vyhodnocení nově implementovaných filtrů a neovlivnila tedy stanovení limitů pro jednotlivé filtry.

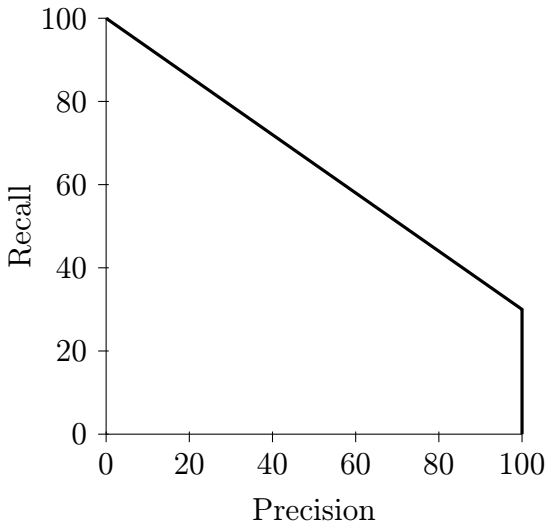


Obrázek 4.1: ROC křivka pro ideální filtr

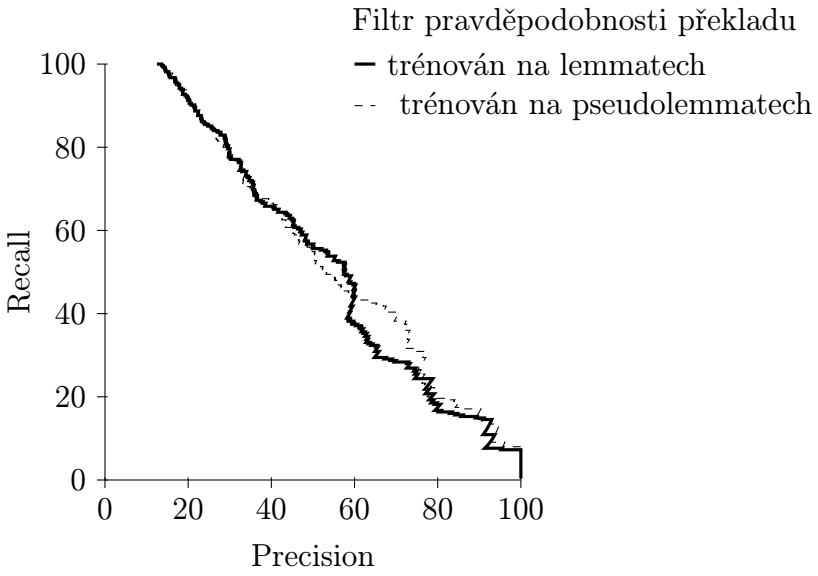
Filtr pravděpodobnosti překladu byl navíc implementován i ve variantě, kdy pracuje s tzv. pseudolemmaty (a nikoliv s lemmaty). Pseudolemmatizace je jednoduchý proces, při kterém se slova obou segmentů zkrátí na maximálně pět písmen (kratší slova jsou ponechána beze změny). Takto je možné se u velké části slov zbavit morfologických koncovek a získaná pseudolemmata jsou jistou aproximací lemmat. Křivka precision/recall pro takto upravený filtr je vynesena na obrázku 4.3. Zde si lze všimnout, že rozdíl mezi použitím pseudolemmat a lemmat není velký, ovšem v oblasti kolem 60% přesnosti vykazují pseudolemmata větší pokrytí než lemmata.

N-gramový filtr naopak má úzkou oblast, ve které dosahuje vyšší než 60% přesnosti při hodnotě recall mezi 5 a 10%, mimo niž hodnota přesnosti rychle klesá.

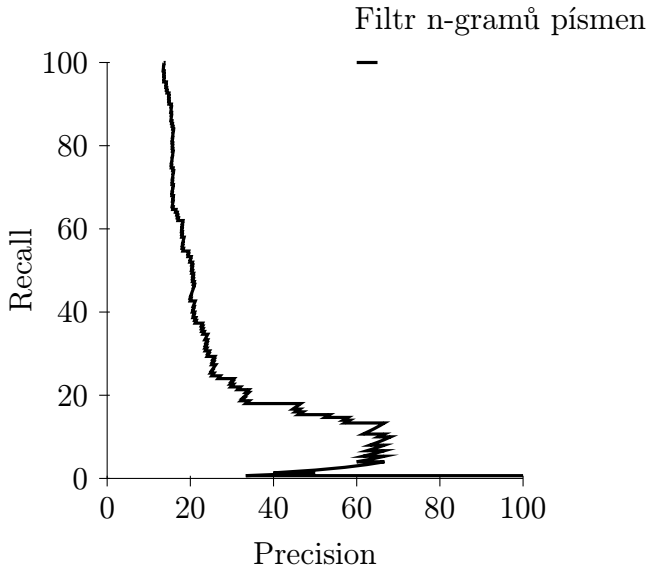
Data pro křivky slovníkového filtru byla získána při třech různých konfiguracích slovníkového filtru. První využívala oba dostupné slovníky, druhá pouze slovník ZČU a třetí pouze slovník generovaný ze slovního zarovnání. Zde lze rozpoznat, že nejlepší výsledky má konfigurace kombinující oba slovníky, naopak slovník získaný z části `news` korpusu CzEng 0.9 není dostačující. Zde se projevuje specifičnost jeho zdroje, neobsahující mnoho slov používaných v jiných zdrojích.



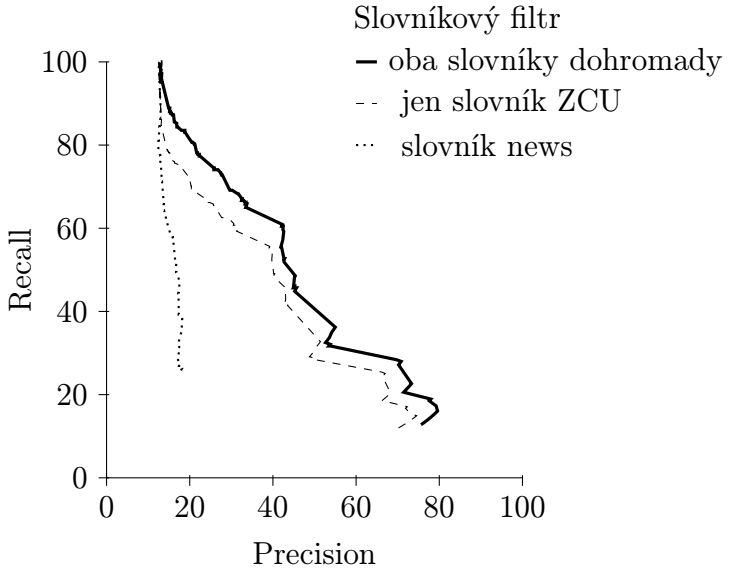
Obrázek 4.2: ROC křivka pro vhodný filtr



Obrázek 4.3: Hodnoty metrik precision a recall filtru pravděpodobnosti překladu pro různé hodnoty limitu.



Obrázek 4.4: Hodnoty metrik precision a recall n-gramového filtru pro různé hodnoty limitu. Uvažovány jsou pouze segmenty delší než 35 znaků.



Obrázek 4.5: Hodnoty metrik precision a recall slovníkového filtru pro různé hodnoty limitu. Uvažovány jsou pouze páry mající alespoň 2 slova na české straně.

Kapitola 5

Závěr

V této práci jsem navrhl, implementoval a manuálně vyhodnotil pět různých postupů čištění paralelních dat. Pouze postup založený na kontrole znaků mimo znakovou sadu ASCII je jazykově závislý, ostatní lze bez jakýchkoliv úprav použít i na jiné jazykové páry než angličtina-čeština. Dále jsem vyhodnotil filtry implementované dříve, použité při přípravě velkého česko-anglického korpusu CzEng 0.9.

Nově implementované filtry mohou výrazně zlepšit proces filtrace česko-anglických segmentů při práci na novém vydání korpusu CzEng. S celkovou hodnotou recall téměř 50% odstraňují polovinu chybných větných párů obsažených v současné verzi korpusu. S celkovou přesností přes 70% je zajištěno, že během tohoto procesu nebude zbytečně odstraňováno velké množství správných větných párů.

Dalším výsledkem práce jsou trénovací a evaluační sady, které lze využít při přípravě nových filtrů, a nástroje pro evaluaci filtrů.

Práci lze rozšířit ve dvou oblastech, jež spolu vzájemně souvisí. První oblastí je zvětšení anotovaných sad pro přesnější určení precision a recall stávajících filtrů. Druhou oblastí je vývoj nových filtrů, s využitím dodatečných informací obsažených v CzEng Export Formatu, t.j. kromě lemmat např. i značek. Tím se zlepší nejenom samotný filtrovací proces, ale i kvalita strojového překladu natrénovaného na korpusu CzEng.

Příloha A

Filtry - uživatelský manuál

Filtry se spouští pomocí shellskriptů uložených na přiloženém CD v adresáři `shellscripts`. Všechny skripty vyžadují dva parametry:

- adresář pro dočasné soubory;
- soubor s paralelním textem ve formátu CzEng Export Format.

Dále je nutné nastavit některé parametry přímo v shellskriptu. Proměnná `jar` by měla obsahovat cestu k souboru `jar` obsahující balíček `filters`. U slovníkového filtru je dále nutné nastavit proměnné `dict` a `gizadict`, odkazující na soubory obsahující slovník ve formátu ZČU a slovník extrahovaný z výstupu nástroje GIZA++. U filtru pravděpodobnosti překladu je nutné správně nastavit cestu k nástrojům GIZA++ a MKCLS v proměnných `gizabin` a `mkclsbin`. U n-gramového filtru je třeba nastavit cestu nástroje SRILM v proměnné `srilmbin` a cestu k n-gramovému modelu jazyka v proměnné `model`.

Výstup filtrů je na standardní výstup ve formátu:

segment1 segment2 ErRoR_nazevFiltru

Jednotlivá pole jsou od sebe oddělena tabulátorem.

Příloha B

Filtry - programátorský pohled

Filtry jsou uloženy v balíčku `filter`:

- `ASCIIFilter.java`
- `DictionaryFilter.java`
- `GizaFilter.java`
- `NgramFilter.java`
- `NumberFilter.java`

.

V balíčku `filter.tools` jsou uloženy pomocné třídy:

- `ExportFormatProcessor.java` – zpracování CzEng Export Formatu;
- `ExportMode.java` – módy exportu z CzEng Export Formatu (plain, lemma, pseudolemma, tag);
- `GizaTranslations.java` – zpracování výstupu nástroje GIZA++ do slovníku používaného slovníkovým filtrem;
- `SeperateChars.java` – rozdělení znaků mezerami, nutné pro práci s nástrojem SRILM;
- `SrilmProcessor.java` – zpracování výstupu nástroje SRILM.

Literatura

[1] Och F.J. (2009): Google Faculty Summit 2009: Statistical Machine Translation. http://www.youtube.com/watch?v=y_PzPDRPwlA (2.5. 2010)

[2] Bojar O., Janíček M., Žabokrtský Z., Češka P., Beňá P. (2008): Czeng 0.7: Parallel Corpus with Community-Supplied Translations. In Proceedings of LREC 2008, Marrákeš.

[3] Khadivi S., Ney H. (2005): Automatic Filtering of Bilingual Corpora for Statistical Machine Translation. Natural Language Processing and Information Systems, volume 3513/2005, 263-274

[4] Bojar 0., Liška A., Žabokrtský Z. (2010): Evaluating Utility of Data Sources in a Larga Parallel Czech-English Corpus CzEng 0.9. Language Resource and Evaluation Conference 2010 Proceedings

[5] Bojar O., Žabokrtský Z. (2009): CzEng 0.9: Large Parallel Treebank with Rich Annotation. Prague Bulletin of Mathematical Linguistics, 92, v tisku