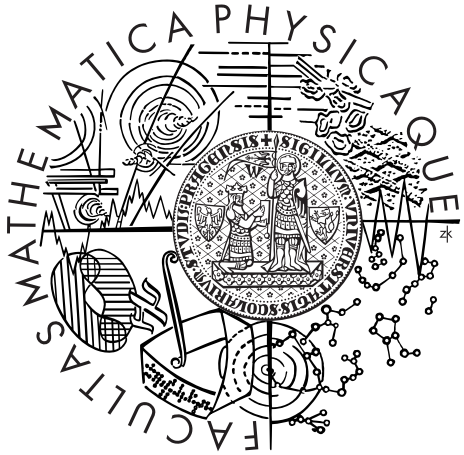
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Ondřej Klejch

Nástroj pro porovnávání a vyhodnocování strojového překladu

Ústav formální a aplikované lingvistiky

Vedoucí bakalářské práce: Mgr. Martin Popel
Studijní program: Informatika
Studijní obor: Obecná informatika

Praha 2013

Chtěl bych poděkovat vedoucímu své bakalářské práce Mgr. Martinu Popelovi za skvělé vedení mé práce, trpělivost a za čas, který mi věnoval při častých konzultacích, jakož i při sepisování této práce. Také bych na tomto místě velmi rád poděkoval rodičům za to, že mne po celou dobu studia podporovali.

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

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

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

Název práce: Nástroj pro porovnávání a vyhodnocování strojových překladů

Autor: Ondřej Klejch

Katedra: Ústav formální a aplikované lingvistiky

Vedoucí bakalářské práce: Mgr. Martin Popel

Abstrakt: Tato bakalářská práce se zabývá vývojem nástroje pro porovnávání a vyhodnocování strojových překladů nazvaného MT-ComparEval. V tomto nástroji je možné porovnávat překlady na základě několika kritérií. Mezi ně patří automatické metriky kvality strojových překladů počítaných pro celé dokumenty nebo jednotlivé věty, porovnání kvality překladů jednotlivých vět pomocí zvýraznění potvrzených, zlepšujících a zhoršujících n-gramů nebo podle souhrnu nejvíce zlepšujících a zhoršujících n-gramů v celém dokumentu. Při porovnávání dvou různých překladů nástroj MT-ComparEval také vykresluje graf s absolutními rozdíly metrik počítaných pro jednotlivé věty a graf s hodnotami z párového bootstrap resamplingu.

Klíčová slova: porovnání strojových překladů, vyhodnocování strojových překladačů, metriky strojových překladů

Title: Tool for comparison and evaluation of machine translation

Author: Ondřej Klejch

Department: Institute of Formal and Applied Linguistics

Supervisor: Mgr. Martin Popel

Abstract: This bachelor thesis is about development of a tool for comparison and evaluation of machine translation called MT-ComparEval. With this tool it is possible to compare translations according to several criteria, such as automatic metrics of machine translation quality computed on whole documents or single sentences, quality comparison of single sentence translation with highlighting confirmed, improving and worsening n-grams or summaries of the most improving and worsening n-grams for the whole document. When comparing two translations, MT-ComparEval also plots a chart with absolute differences of metrics computed on single sentences and a chart with values obtained from paired bootstrap resampling.

Keywords: machine translation comparison, machine translation evaluation, machine translation metrics

Obsah

Úvod

Současný vývoj softwarových projektů probíhá v krátkých iteracích, aby mohl pružně reagovat na změny v zadání. Stejně tak i vývoj strojových překladačů probíhá v krátkých iteracích, aby bylo možné pravidelně ověřovat zlepšení nebo zhoršení kvality překladů. Z toho plyne potřeba vývojářů testovat změny, které provedli v rámci iterace, která mohla trvat 30 sekund, hodinu, den nebo týden. K testování funkčnosti softwarových projektů lze použít různé sady testů – od unit testů po akceptační testy. Je možné stejným způsobem testovat i kvalitu strojových překladačů?

Je samozřejmé, že vývojáři strojových překladačů testují svůj kód stejně jako ostatní vývojáři. Pomocí těchto testů mohou zajistit, aby překladač fungoval, ale nemohou takto jednoduše zajistit kontrolu kvality překladů. Kontrolu překladů může provádět sám vývojář, ale ztratí tím spoustu času a energie, kterou by mohl lépe využít při dalším vývoji.

Žádná kontrola kvality, kterou není možné automatizovat, vývojářům nepomůže zrychlit vývojový cyklus. Proto je třeba, aby existoval způsob jak rychle a opakovaně kontrolovat kvalitu překladů.

Z tohoto důvodu byly vynalezeny metriky strojového překladu, které umožňují rychle a opakovaně testovat překlady. Tyto metriky se snaží hodnotit překlady tak, aby jejich výsledky co nejvíce odpovídaly lidskému hodnocení.

Avšak výsledek většiny metrik je pouhé jedno číslo. Je tedy třeba, aby vývojaři měli k dispozici nástroje, pomocí kterých budou moci efektivně vyhodnocovat své překlady. Pomocí takovýchto nástrojů si pak vývojáři mohou vyhodnotit, co způsobily změny, které provedli.

Vývojář by měl mít možnost překlady nejen vyhodnocovat, ale i porovnávat. Porovnání dvou různých překladů může totiž vývojáři objasnit jevy, o kterých nemusel vůbec vědět. A na základě takto získaných poznatků pak může vyvinout novou lepší verzi překladače. A po mnoha iteracích vývoje snad vyvine překladač, s kterým bude moci být spokojený.

Na trhu existuje několik nástrojů, které umožňují vyhodnocovat překlady pomocí různých metrik nebo porovnávat překlady na základě různých jiných kritérií. Důležité je, aby mohl vývojář použít přístup, který mu vyhovuje.

Nástrojem, který umožňuje vývojářům efektivně porovnávat a vyhodnocovat překlady, se snaží být MT-ComparEval, o němž pojednává tato bakalářská práce.

1. Motivace

V této kapitole budou v krátkosti představeny všechny důvody, proč a jak byl vyvinut nástroj MT-ComparEval. Zvolená řešení pak budou více rozebrána v následujících kapitolách.

1.1 Experimenty a tasky

Strojové překlady jsou při automatickém vyhodnocování porovnávány s referenčním překladem (dále jen reference). Vyhodnocování strojových překladačů může být zaměřeno na různé domény textu (např. novinové články, beletrie apod.) různých délek nebo na různé jazykové páry (např. z angličtiny do češtiny, z němčiny do češtiny apod.). Aby uživatelé nástroje MT-ComparEval mohli snadno vyhodnocovat své strojové překladače na různých doménách nebo pro různé jazykové páry, mohou si vytvořit různé **„experimenty“**, v rámci kterých mohou porovnávat své překlady s příslušnými referencemi. Obrázek 1.1 ukazuje příklad vytvořených experimentů z různých domén.

V každém experimentu pak uživatel může vytvářet **„tasky“**, které později může vyhodnocovat a porovnávat (viz Obrázek 1.2).

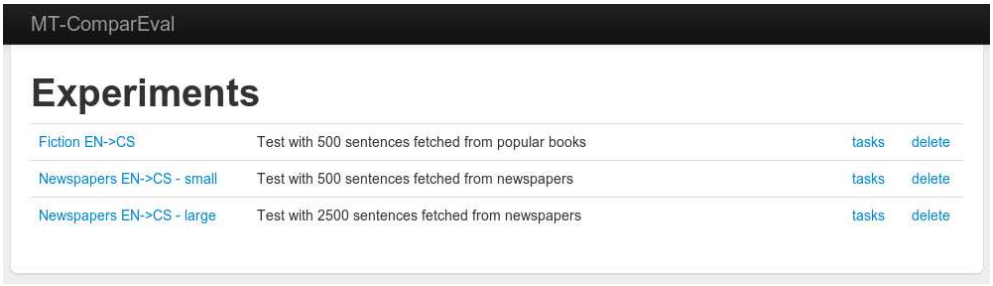
Experimenty, tasky a jiné pojmy podrobněji rozebírá 2. kapitola.

1.2 Metriky strojového překladu

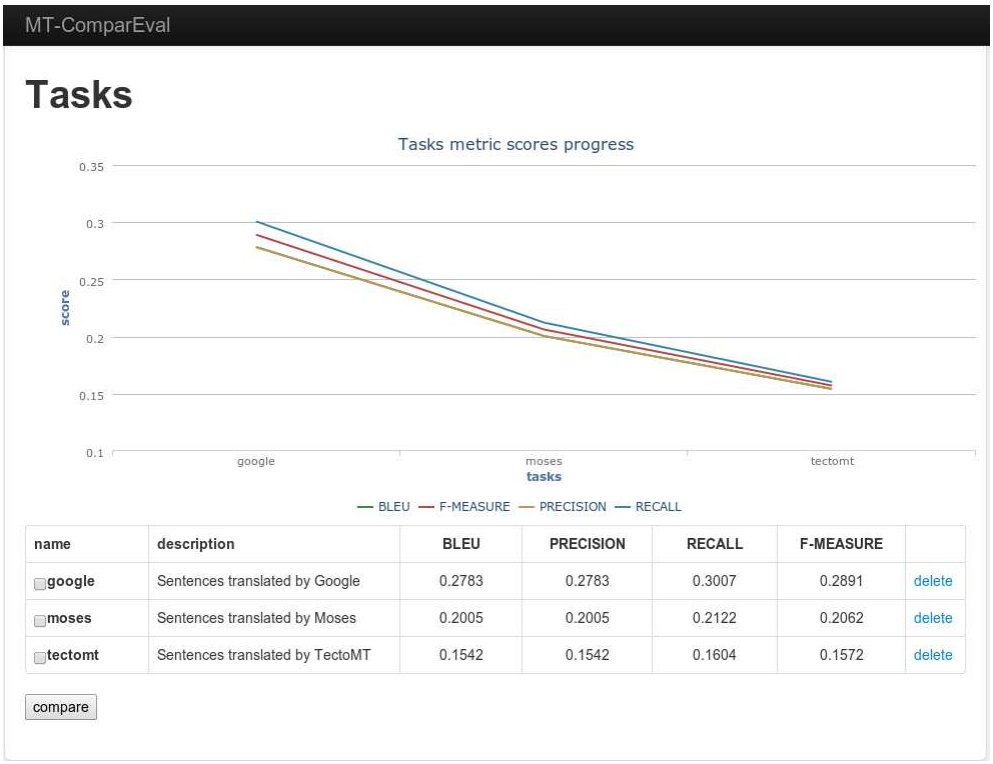
K porovnávání překladů mohou být použity metriky strojového překladu. V nástroji MT-ComparEval jsou metriky počítány na úrovni celých tasků (viz Obrázek 1.3) i na úrovni jednotlivých vět (viz Obrázek 1.4).

Metriky vypočtené na úrovni jednotlivých vět slouží i k řazení vět ve webovém rozhraní. Věty jsou řazeny podle absolutního rozdílu hodnoty dané metriky u jednotlivých překladů. Čím vyšší je tento rozdíl, tím došlo k většímu zlepšení při překladu dané věty. Proto lze věty řadit podle míry zlepšení, ke které došlo během překladu dané věty.

Z vypočítaných metrik lze vytvořit grafy, které lépe ilustrují rozdíl v kvalitě jednotlivých překladů. Nástroj MT-ComparEval obsahuje graf rozložení absolutního rozdílu výsledků metrik v jednotlivých větách (viz Obrázek 1.5) a graf zobrazující rozdíly metrik celých dokumentů získaných pomocí párového bootstrap resamplingu (viz Obrázek 1.6).



Obrázek 1.1: Přehled vytvořených experimentů v nástroji MT-ComparEval

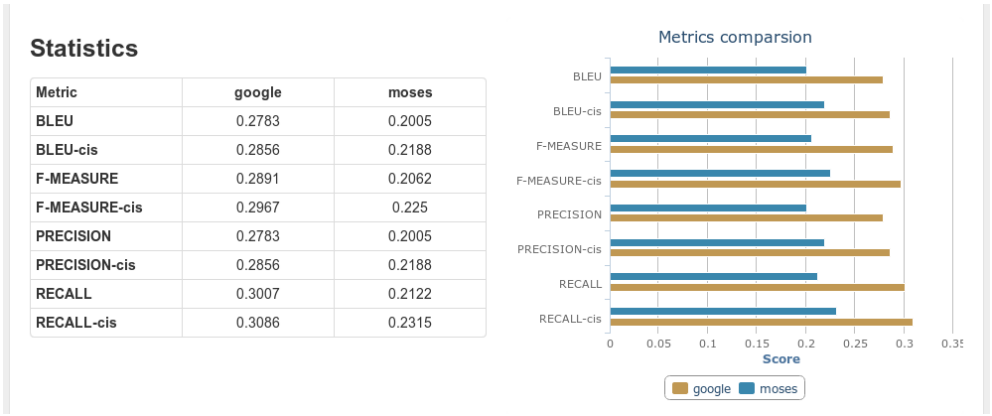


Obrázek 1.2: Přehled vytvořených tasků v jednom z experimentů v nástroji MT-ComparEval

Metrikami strojových překladů se více zabývá 3. kapitola. V té budou představeny všechny metriky, které jsou použity v nástroji MT-ComparEval.

1.3 Porovnávání dvou strojových překladů

Kvalitu strojových překladačů je možné vyhodnotit i porovnáním jednotlivých vět. Nástroj MT-ComparEval umožňuje procházet věty seřazené podle metriky strojových překladů a hledat v těchto větách rozdíly, které ovlivnily výsledné metriky. Aby uživatelé mohli snadněji vyhledat rozdíly mezi větami, je možné zobrazit potvrzené n-gramy (n-gramy, které se nachází v referenci i strojovém

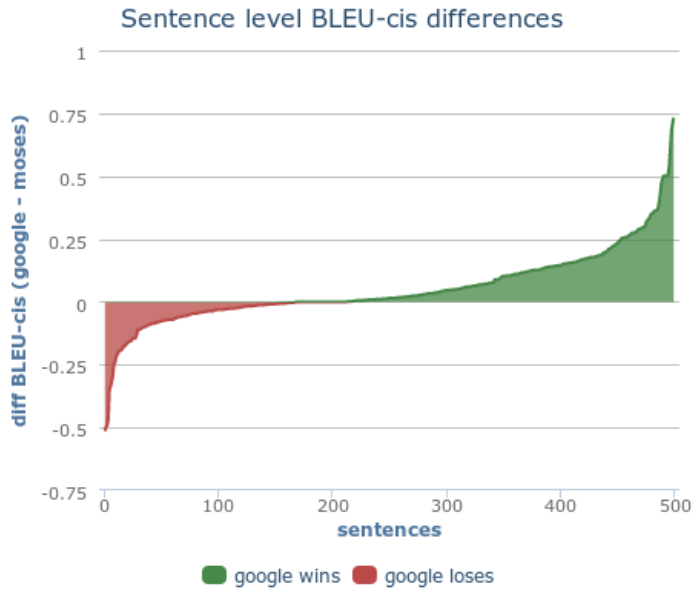


Obrázek 1.3: Porovnání metrik dvou tasků v nástroji MT-ComparEval.

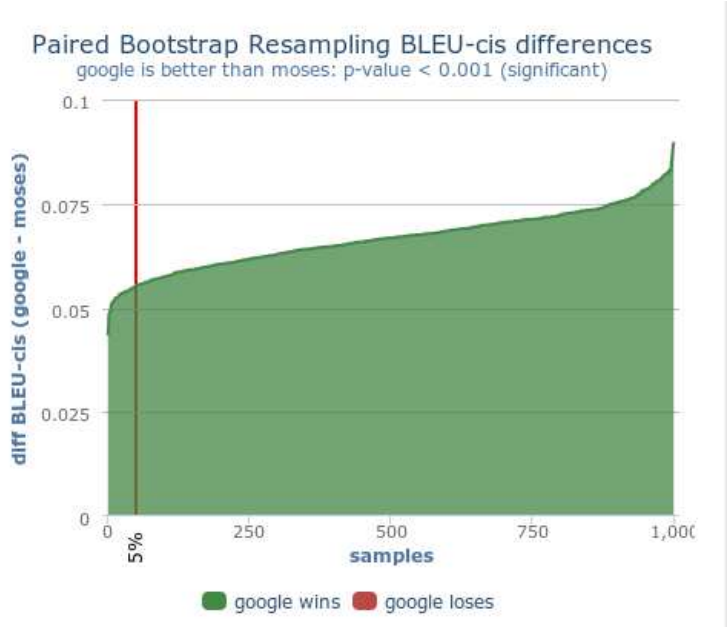
Source	Bělohlávek considers the Czech national song to be one of the most beautiful national anthems .
Reference	Bělohlávek považuje českou národní píseň za jednu z nejkrásnějších hymen .
google	Bělohlávek považuje českou národní píseň za jednu z nejkrásnějších národní hymny .
moses	Bělohlávek za českou národní píseň , která je jedním z nejkrásnějších národní hymny .

	BLEU	BLEU-cis	F-MEASURE	F-MEASURE-cis	PRECISION	PRECISION-cis	RECALL	RECALL-cis
google	0.751	0.751	0.7862	0.7862	0.751	0.751	0.8248	0.8248
moses	0.2139	0.2139	0.2414	0.2414	0.2139	0.2139	0.2769	0.2769
Diff	0.5371	0.5371	0.5448	0.5448	0.5371	0.5371	0.5479	0.5479

Obrázek 1.4: Porovnání metrik dvou překladů v nástroji MT-ComparEval.

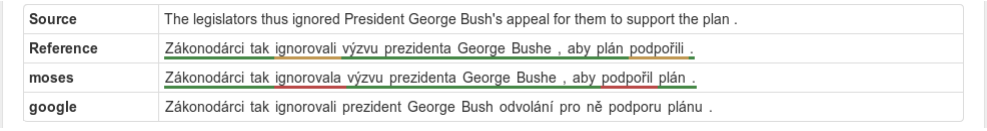


Obrázek 1.5: Graf zobrazující rozdělení absolutních rozdílů metrik spočtených na jednotlivých větách.



Obrázek 1.6: Graf zobrazující rozdíly metrik získaných pomocí párového bootstrap resamplingu.

Obrázek 1.7: Porovnání dvou překladů v nástroji MT-ComparEval. V jednotlivých překladech jsou pastelovými odstíny žluté a modré barvy zvýrazněny potvrzené n-gramy, sytými odstíny žluté a modré barvy jsou zvýrazněny zlepšující n-gramy a červenou barvou jsou zvýrazněny zhoršující n-gramy. V referenci jsou zelenou barvou zvýrazněny potvrzené n-gramy, které se nacházejí v obou překladech, pastelovým odstínem modré barvy jsou zvýrazněny zlepšující n-gramy ze systému ***google*** *a pastelovým odstínem žluté barvy jsou zvýrazněny zlepšující n-gramy ze systému* ***moses****.*



Obrázek 1.8: Porovnání překladu a reference se zvýrazněným diffem v nástroji MT-ComparEval. Zeleně jsou podtržena slova, která se nachází v nejdelší společné podposloupnosti překladu a reference. Slova, která se v této podposloupnosti nenachází, jsou podtržena v překladu červenou barvou a v referenci žlutou barvou.

překladu), zlepšující n-gramy (potvrzené n-gramy, které se nachází pouze v jednom z porovnávaných překladů) nebo zhoršující n-gramy (nepotvrzené n-gramy, které se nachází pouze v jednom z porovnávaných překladů). Na Obrázku 1.7 je vidět porovnání dvou překladů se zvýrazněnými n-gramy. K vyhledání rozdílů mezi referencí a překladem slouží zobrazení jejich diffu (viz Obrázek 1.8).

Strojové překlady nemusí být porovnávány pouze na základě strojových metrik. Další informací, díky které je možné si vytvořit lepší představu o vlastnostech strojového překladače, jsou přehledy nejvíce zlepšujících a zhoršujících n-gramů v jednotlivých překladech (viz Obrázek 1.9).

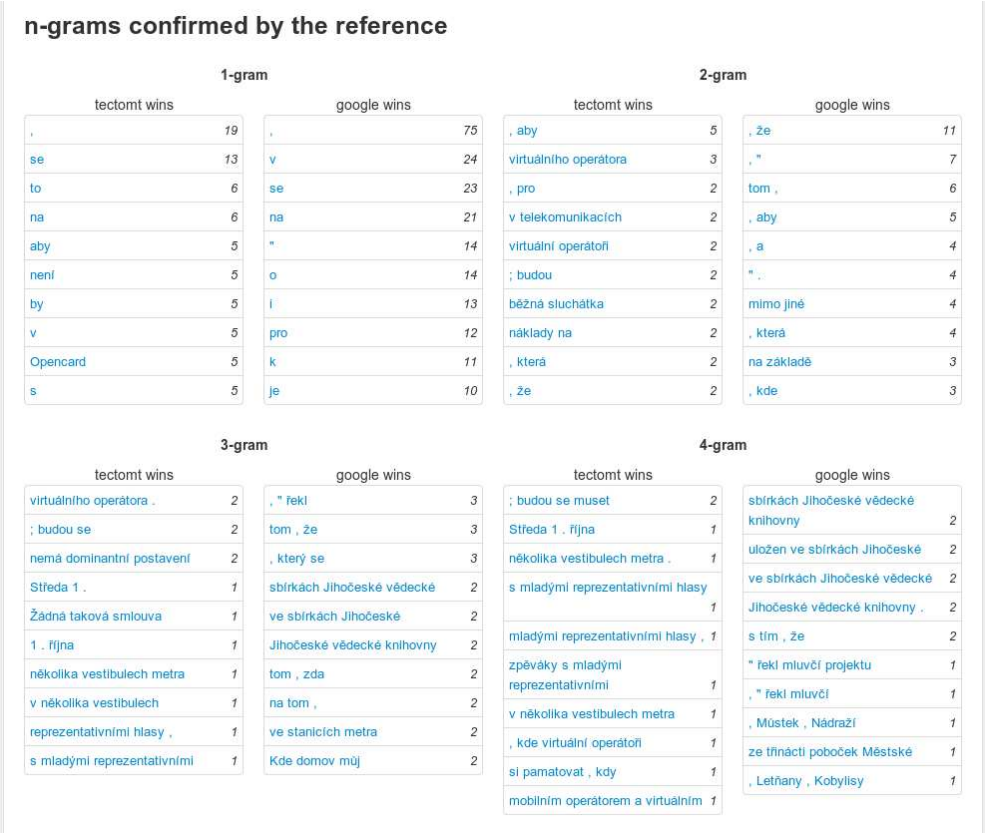
Přehledy zlepšujících i zhoršujících n-gramů mohou být použity k filtrování vět, aby si uživatel mohl snadno prohlédnout věty, ve kterých se dané n-gramy nachází. Na Obrázku 1.10 je vidět výpis vět obsahujících zlepšující n-gram **„mimo jiné“**.

O algoritmech, které byly použity při porovnávání dvou překladů, a hledání pozic potvrzených n-gramů pojednává 4. kapitola.

1.4 Automatizace a snadné použití

Nástroj MT-ComparEval byl navržen tak, aby bylo možné jeho použití co nejvíce automatizovat. Uživatelé proto nemusí ručně vytvářet každý task, ale mohou napsat jednoduché skripty, pomocí nichž jsou schopni vytvářet tasky v nástroji MT-ComparEval automaticky při každé změně strojového překladače.

V 5. kapitole je popsáno, jak se používá nástroj MT-ComparEval a jak je možné nasadit ho do vývojového procesu.



n-grams confirmed by the reference

1-gram

tectomt wins		google wins	
,	19	,	75
se	13	v	24
to	6	se	23
na	6	na	21
aby	5	"	14
není	5	o	14
by	5	i	13
v	5	pro	12
Opencard	5	k	11
s	5	je	10

2-gram

tectomt wins		google wins	
, aby	5	, že	11
virtuálního operátora	3	, "	7
, pro	2	tom ,	6
v telekomunikacích	2	, aby	5
virtuální operátoři	2	, a	4
; budou	2	" .	4
běžná sluchátka	2	mimo jiné	4
náklady na	2	, která	4
, která	2	na základě	3
, že	2	, kde	3

3-gram

tectomt wins		google wins	
virtuálního operátora .	2	, " řekl	3
; budou se	2	tom , že	3
nemá dominantní postavení	2	, který se	3
Středa 1 .	1	sbírkách Jihočeské vědecké	2
Žádná taková smlouva	1	ve sbírkách Jihočeské	2
1 . října	1	Jihočeské vědecké knihovny	2
několika vestibulech metra	1	tom , zda	2
v několika vestibulech	1	na tom ,	2
reprezentativními hlasy ,	1	ve stanicích metra	2
s mladými reprezentativními	1	Kde domov můj	2

4-gram

tectomt wins		google wins	
; budou se muset	2	sbírkách Jihočeské vědecké knihovny	2
Středa 1 . října	1	uložen ve sbírkách Jihočeské	2
několika vestibulech metra .	1	ve sbírkách Jihočeské vědecké	2
s mladými reprezentativními hlasy	1	Jihočeské vědecké knihovny .	2
mladými reprezentativními hlasy ,	1	s tím , že	2
zpěváky s mladými reprezentativními	1	" řekl mluvčí projektu	1
v několika vestibulech metra	1	, " řekl mluvčí	1
, kde virtuální operátoři	1	, Můstek , Nádraží	1
si pamatovat , kdy	1	ze třinácti poboček Městské	1
mobilním operátorem a virtuálním	1	, Letňany , Kobylisy	1

Obrázek 1.9: Přehled nejvíce zlepšujících n-gramů v jednotlivých překladech.

You are displaying sentences with improving n-gram **mimo jiné**. Show all senteces

Source	The advantage of the music players , on the other hand , is , among other reasons , that they are easy to operate .
Reference	Výhodou přehrávačů zase je mimo jiné snadná obsluha .
tectomt	Výhoda hudeb hráčů na druhé straně je mezi jinými důvody , že jsou snadní fungovat .
google	Výhodou přehrávačů hudby , na druhé straně , je **mimo jiné** z důvodů , že jsou snadno ovladatelná .

Source	At the same time , the financial group Penta , which owns , among others , U : fon , stands behind Mobilking .
Reference	Přitom za Mobilkingem stojí finanční skupina Penta , která v Česku vlastní mimo jiné U : fona .
tectomt	Na stejné době finanční skupina Penta , která má mezi ostatními , U : fon stojí za Mobilking .
google	Ve stejné době , finanční skupina Penta , která vlastní **mimo jiné** , U : fon , stojí za MOBILKING .

Obrázek 1.10: Výpis vět, ve kterých se nachází zlepšující n-gram ***„mimo jiné“***

1.5 Rozšiřitelnost

Nástroj MT-ComparEval si neklade za cíl implementovat co nejvíce metrik, proto jsou v něm předprogramovány pouze některé. Zároveň je však umožněno doprogramovat si vlastní metriky.

Informace o tom, jak si uživatel může doprogramovat vlastní metriky nebo jak byl nástroj MT-ComparEval vyvinut, se nachází v 6. kapitole.

2. Názvosloví MT-ComparEval

Pojmy **experiment**, **task** a **n-gram** (potvrzený, zlepšující, zhoršující) jsou důležité k pochopení fungování celého nástroje MT-ComparEval. Proto v této kapitole bude podrobněji vysvětleno, co tyto pojmy znamenají a jaké jsou mezi nimi vztahy.

2.1 Experiment

Uživatelé mohou chtít testovat své překladače na různých textových doménách různých délek nebo testovat své překladače pro různé jazykové páry. Nástroj MT-ComparEval proto umožňuje vytvořit různé **experimenty**.

Každý experiment obsahuje vlastní **zdrojové věty** (věty, které mají být strojovým překladačem přeloženy) a **referenční věty** (člověkem přeložené věty, které budou později použity k vyhodnocování jednotlivých strojových překladačů). Pomocí různých dvojic zdrojů a referencí mohou uživatelé vyhodnocovat strojové překladače na různých testovacích sadách, což pro ně může být výhodné v případě, kdy se překladač na různých zdrojích chová různým způsobem.

2.2 Task

V rámci experimetu mohou uživatelé porovnávat různé překlady zdrojových vět – ať už se jedná o různé verze jednoho strojového překladače, nebo různé strojové překladače. Aby mohl uživatel vyhodnocovat různé překlady, umožňuje nástroj MT-ComparEval nahrávat do experimentů tzv. **tasky**. Každý task reprezentuje jednu verzi překladu zdrojových vět, která později může být porovnána s jinou verzí překladu. V rámci této bakalářské práce budou strojové překlady nazývány zkráceně slovem **překlady**.

2.3 N-gramy – potvrzené, zlepšující a zhoršující

Při vysvětlování počítání strojových metrik nebo popisu, jak jsou porovnávány dva překlady jedné věty, jsou často použity termíny spojené se slovem n-gram. Proto budou v této části všechny tyto termíny vysvětleny.

Posloupnost n po sobě jdoucích slov[1] se nazývá **n-gram**. N-gramy, které se nacházejí i v referenci i ve strojovém překladu, se v rámci nástroje MT-ComparEval nazývají **kandidáti potvrzených n-gramů**. Z těchto kandidátů je možné vybrat **potvrzené n-gramy**. Ty jsou vybrány z kandidátů potvrzených n-gramů tak, aby počet výskytů daného potvrzeného n-gramu nebyl větší než počet výskytů daného n-gramu v referenci.

Potvrzené n-gramy, které se při porovnávání dvou překladů nacházejí pouze v jednom z nich, jsou nazývány **zlepšující n-gramy**. N-gramy, které nejsou

[1] V této práci se za slova považují i interpunkční znaménka, a **slovo** tedy znamená totéž co **token**.

Obrázek 2.1: Porovnání dvou překladů v nástroji MT-ComparEval. V jednotlivých překladech jsou pastelovými odstíny žluté a modré barvy zvýrazněny potvrzené n-gramy, sytými odstíny žluté a modré barvy jsou zvýrazněny zlepšující n-gramy a červenou barvou jsou zvýrazněny zhoršující n-gramy. V referenci jsou zelenou barvou zvýrazněny potvrzené n-gramy, které se nacházejí v obou překladech, pastelovým odstínem modré barvy jsou zvýrazněny zlepšující n-gramy ze systému ***google*** *a pastelovým odstínem žluté barvy jsou zvýrazněny zlepšující n-gramy ze systému* ***moses****.*

potvrzené referencí a při porovnávání dvou překladů se nacházejí pouze v jednom z nich, jsou nazývany **zhoršující n-gramy**. O významu jednotlivých typů n-gramů je možné si udělat lepší představu na Obrázku 2.1.

V nástroji MT-ComparEval se nachází i přehled **nejvíce zlepšujících n-gramů**. Nejvíce zlepšující n-gramy jsou ty n-gramy, které byly nejčastěji prohlášeny za zlepšující. Stejně tak je možné nalézt i přehled **nejvíce zhoršujících n-gramů**, pro které platí totéž jako pro nejvíce zlepšující n-gramy.

To, že byl některý n-gram prohlášen za zlepšující, neznamená nutně, že zlepšil skutečnou kvalitu překladu. Zlepšení je posuzováno pouze vzhledem k metrice BLEU (tedy vzhledem k podobnosti s referencí). Stejně tak i zhoršující n-gram nemusel nutně zhoršit kvalitu překladu.

2.4 Diff dvou vět

Nástroj MT-ComparEval umožňuje zobrazit **diff** dvou překladů anebo překladu a reference. Pojmem diff dvou vět se rozumí rozdíl těchto vět. Rozdíl vět je možné pro každou z těchto vět definovat jako slova, která se nachází v dané větě a zároveň se nenachází v nejdelší společné podposloupnosti slov porovnávaných vět.

3. Metriky strojového překladu

Při porovnávání dvou strojových překladů je třeba, snadno a rychle určit, jak jsou jednotlivé překlady dobré. Posuzování kvality překladů člověkem je časově náročné a není možné ho automatizovat. Proto byly vytvořeny metriky, jejichž výsledky jsou podobné výsledkům získaným od lidí, s tím rozdílem, že je možné je rychle spočítat, aplikovat na různé jazyky a podle potřeby opakovat. Mezi tyto metriky patří i metrika BLEU [1], která byla zvolena jako výchozí metrika v nástroji MT-ComparEval.

Při počítání metrik strojových překladů se zjišťuje, jak moc se strojový překlad podobá překladu vytvořenému člověkem – tzv. referenčnímu překladu. Čím více se podobá strojový překlad referenci, tím je považován za lepší. To ovšem neznamená, že překlad s nižší hodnotou metriky je špatný. Překlad může být totiž proveden několika různými způsoby (např. může být změněn slovosled, mohou být použita synonyma, ...), z nichž se pouze jeden bude shodovat s referenčním překladem. Tento problém sice může zmírnit použití více referenčních překladů, ale úplně ho odstranit nelze. Některé překladatelské systémy jsou optimalizovány pro určité metriky, avšak jejich výsledný překlad nemusí být ideální. Proto je třeba, aby se při vyhodnocování strojového překladu nespoléhalo pouze na jednu metriku.

3.1 BLEU

Nejvíce používanou metrikou pro vyhodnocování překladů je metrika BLEU. Hlavní důvod spočívá v tom, že koreluje s lidským hodnocením překladu [2] a je možné ji rychle spočítat. Rychlost výpočtu metrik je pro vývojáře překladových systémů velmi důležitá, protože vývojáři potřebují každou změnu v systému otestovat, aby si mohli být jisti, že nezhoršili kvalitu překladového systému.

Při počítání metriky BLEU počítá **modified precision** n-gramů, která říká, kolik procent n-gramů z vyhodnocovaného překladu se nachází i v referenčním překladu. Výpočet modified precision n-gramů se liší od běžného výpočtu precision tím, že je místo počtu kandidátů potvrzených n-gramů použit počet potvrzených n-gramů (viz definice v 1. Kapitole).

$$\mathrm{PREC}_n = \frac{|\{\text{potvrzené n-gramy}\}|}{|\{\text{ n-gramy ze strojového překladu }\}|}$$

Tímto způsobem je možné spočítat modified precision pro libovolné délky n-gramů. Avšak při výpočtu metriky BLEU se počítá pouze s 1-gramy, 2-gramy, 3-gramy a 4-gramy. Modified precision pro jednotlivé délky n-gramů jsou pak kombinovány dohromady pomocí geometrického průměru.

$$\mathrm{PREC}_{1-4} = \sqrt[4]{\prod_{n=1}^{4} \mathrm{PREC}_n}$$

Aby bylo zaručeno, že překlad bude mít podobnou délku jako reference, používá se **brevity penalty**, která zhoršuje skóre větám, které jsou kratší než refe-

rence. Věty, které jsou delší než reference, nemusí být takto postihovány, protože postih za rozdíl v délkach vět je již zahrnut ve výpočtu modified precision.

$$\mathrm{BP} = \begin{cases} 1 & \mathrm{r} \leq \mathrm{t} \quad \text{r je počet slov v referenci} \\ e^{(1-r/t)} & \mathrm{r} > \mathrm{t} \quad \text{t je počet slov v překladu} \end{cases}$$

Metrika BLEU se pak spočítá pomocí následujícího vzorečku:

$$\mathrm{BLEU} = \mathrm{BP} \cdot \mathrm{PREC}_{1-4}$$

3.2 BLEU$_S$

Aby bylo možné počítat metriku BLEU i pro jednotlivé věty, musí být vyřešena situace, kdy se v některých větách nevyskytují žádné potvrzené 1-gramy, 2-gramy, 3-gramy nebo 4-gramy. V této situaci by výsledná metrika těchto vět byla rovna nule, i kdyby se v nich nějaký potvrzený n-gram nacházel. Tento problém řeší vylepšení metriky BLEU – metrika BLEU$_S$ [3]. Ta připočítává +1 ke všem n-gramům delším než jedna. Tím pádem může být metrika spočítána i pro věty, které nemají potvrzené n-gramy libovolné délky. Zároveň je možné hodnotit úplně špatný překlad nulovým skóre, protože 1-gramy vyhlazovány nejsou.

Skript mteval-13a.pl,[1] vytvořený americkým Národním institutem pro standardizaci a technologie NIST,[2] používá jiný způsob vyhlazování. Místo přičítání +1 ke všem n-gramům má speciální vzorec pro počítání precision u chybějících potvrzených n-gramů. Nástroj mteval-13a.pl nalezne nejmenší číslo k takové, že věta, pro kterou je počítáno BLEU, neobsahuje žádný potvrzený k-gram (n-gram délky k). Pak je modified precision počítána dvěma způsoby. U n-gramů délky menší než k je použit běžný vzorec pro výpočet modified precision a u n-gramů délky větší nebo rovné k je pro výpočet modified precision použit následující vzorec:

$$\mathrm{PREC}_n = \frac{1}{2^{n-k+1} \cdot |\{ \text{ n-gramy ze strojového překladu } \}|}$$

Rozdíl obou metrik je možné demonstrovat na větě obsahující 4 slova, v níž se nachází jeden potvrzený 2-gram (z čehož plyne, že obsahuje i dva potvrzené 1-gramy).

[1] ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v13a.pl
[2] http://www.nist.gov

délka	počet potvrzených	celkový počet	PRECISION	
			BLEU_S	mteval
1-gram	2	4	$\frac{2}{4} = 0.50$	$\frac{2}{4} = 0.50$
2-gram	1	3	$\frac{1+1}{3+1} = 0.50$	$\frac{1}{3} \approx 0.33$
3-gram	0	2	$\frac{0+1}{2+1} \approx 0.33$	$\frac{1}{2^{1} \cdot 2} = 0.25$
4-gram	0	1	$\frac{0+1}{1+1} = 0.50$	$\frac{1}{2^{2} \cdot 1} = 0.25$
		BLEU	$\sqrt[4]{\frac{1}{2^3 \cdot 3}}$	$\sqrt[4]{\frac{1}{2^5 \cdot 3}}$

To, že se způsoby vyhlazování liší, ničemu nevadí, protože obě metriky ve většině případů dokáží rozpoznat lepší překlad od horšího.

V nástroji MT-ComparEval bylo pro výpočet metriky BLEU u jednotlivých vět použito vyhlazování BLEU_S, u výpočtu metriky BLEU pro celé dokumenty se v nástroji MT-ComparEval vyhlazování nepoužívá. Ale pro zjednodušení uživatelského rozhraní jsou obě dvě metriky nazvány BLEU.

3.3 Recall

Při výpočtu BLEU se počítá modified precision n-gramů ze strojového překladu. Metrika (modified) **recall** se spočítá tak, že se místo n-gramů ze strojového překladu použijí n-gramy z referenčního překladu. Tato metrika určuje, s jakou pravděpodobností se nachází n-gram z referenčního překladu i v překladu strojovém.

$$\text{REC}_n = \frac{|\text{potvrzené n-gramy}|}{|\text{referenční n-gramy}|}$$

$$\text{REC}_{1-4} = \sqrt[4]{\prod_{n=1}^{4} \text{REC}_n}$$

Nicméně i recall má své problémy. Strojové překlady, které jsou výrazně delší než překlady referenční, by mohly mít vysokou hodnotu recall, ale i tak by se jednalo o špatné překlady, protože by obsahovaly mnoho nadbytečných slov. Tento problém řeší metrika F-Measure, která bude vysvětlena v následující části.

3.4 F-Measure

Metrika F-Measure kombinuje recall i precision, čímž zajišťuje, že zbytečně dlouhé překlady budou mít horší výsledky. F-Measure se počítá pomocí následujícího vzorce:

$$\text{F-MEASURE} = 2 \cdot \frac{\text{PREC}_{1-4} \cdot \text{REC}_{1-4}}{\text{PREC}_{1-4} + \text{REC}_{1-4}}$$

3.5 Distribuce rozdílů hodnot metrik ve větách

Ve webovém rozhraní nástroje MT-ComparEval se nachází graf, který znázorňuje distribuci (rozdělení) rozdílů hodnot metrik ve větách porovnávaných systémů. Z tohoto grafu je patrné, kolik překladů vět bylo zlepšeno a jak moc byl daný překlad zlepšen (označeno zelenou barvou v Obrázku 1.5). To samé lze z tohoto grafu zjistit pro věty, jejichž překlad byl zhoršen (označeno červeně).

3.6 Nepárový bootstrap resampling

Metoda **Bootstrap resampling** [4] slouží k výpočtu 95% intervalu spolehlivosti pro danou metriku. Při této metodě jsou vytvořeny „nové“ vzorky náhodným výběrem vět (s opakováním) z překladu. Pro každý takto získaný vzorek je pak spočtena metrika a z těchto výsledků je vybrán 95% interval spolehlivosti.

3.7 Párový bootstrap resampling

Pokud se při porovnávání hodnoty metrik spočtených na celých dokumentech příliš neliší, nelze tvrdit, že překlad s vyšším skóre je automaticky lepší. Aby bylo možné spolehlivě prohlásit, že překlad s vyšším skóre je lepší než překlad s nižším skóre, je nutné zjistit, jestli je rozdíl mezi těmito hodnotami signifikantní.

K tomuto účelu se používá metoda **Paired bootstrap resampling** [4]. Tato metoda generuje vzorky stejným způsobem jako bootstrap resampling. Místo hodnot metrik porovnávaných překladů se používá rozdíl těchto metrik. Z intervalu spolehlivosti rozdílu lze následně určit, jestli je první z porovnávaných systémů signifikantně lepší, signifikantně horší nebo není signifikantně ani lepší ani horší.

V nástroji MT-ComparEval se nachází grafy, které zobrazují 95% interval spolehlivosti metrik jednotlivých překladů, a graf, který zobrazuje výsledky rozdílů metrik získaných během paired bootstrap resamplingu.

4. Porovnávání dvou překladů

Při zobrazování rozdílů dvou překladů mohou být zvýrazněna slova či slovní spojení, která byla přeložena správně (potvrzené n-gramy), nebo která zlepšila či zhoršila překlad.

Druhou možností pro porovnávání překladů je zobrazení rozdílů porovnávaného překladu s referencí anebo dvou překladů. Pomocí této volby může být odhalena věta, která je složená ze samých potvrzených n-gramů, a přesto se liší od reference, protože n-gramy se nacházejí na špatných pozicích ve větě, což může být způsobeno špatným slovosledem v překladu.

Při hledání potvrzených, zlepšujících n-gramů, zhoršujících n-gramů nebo zobrazování diffu může nastat několik problémů, které budou vysvětleny v následující kapitole.

4.1 Hledání potvrzených n-gramů

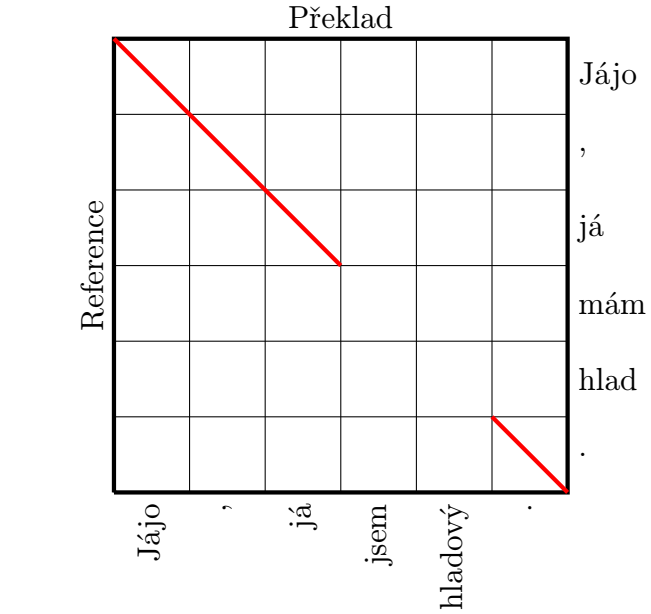
Jelikož se ve větách mohou slova či tokeny libovolně opakovat, může nastat situace, kdy se v dané větě bude nacházet více kandidátů na potvrzený n-gram (viz Obrázek 4.2). Člověk sice snadno rozezná, který n-gram je potvrzený, ale v případě dlouhých vět by to mohlo být časově náročné. Jedním z cílů této bakalářské práce bylo najít algoritmus, který by ze všech kandidátů na potvrzený n-gram našel ty, které jsou s největší pravděpodobností potvrzené n-gramy.

Pro znázornění hledání potvrzených n-gramů je použit graf reprezentující porovnání dvou vět. Tento graf se používá i při výpočtu diffu nebo LCS (nejdelší společná podposloupnost), které budou později použity. Každá hrana v tomto grafu odpovídá jednomu slovu z porovnávaných vět. Horizontální čáry v grafu reprezentují slova, která jsou v překladu navíc oproti referenci. Pro vertikální čáry to platí naopak, t.j. reprezentují slova, která jsou v referenci navíc oproti překladu. Diagonální čáry reprezentují shodu mezi referencí a překladem. Kandidáti na potvrzený n-gram mohou být reprezentováni jako diagonální hrany v grafu, který představuje porovnání dvou vět (viz Obrázek 4.1).

V případě, že počet kandidátů na potvrzený n-gram je stejný jako počet potvrzených n-gramů, je řešení našeho problému jednoduché. V grafu to odpovídá situaci, kdy se v řádku a sloupci nachází stejný počet diagonálních čar. V případě, že se některý z potvrzených n-gramů vyskytuje vícekrát, je vždy použit první ještě nepoužitý kandidát. Všichni kandidáti na potvrzený n-gram jsou potvrzené n-gramy, a tak mohou být patřičně zvýrazněny. Ve všech obrázcích budou potvrzené n-gramy zvýrazněny červenou barvou (viz Obrázek 4.1).

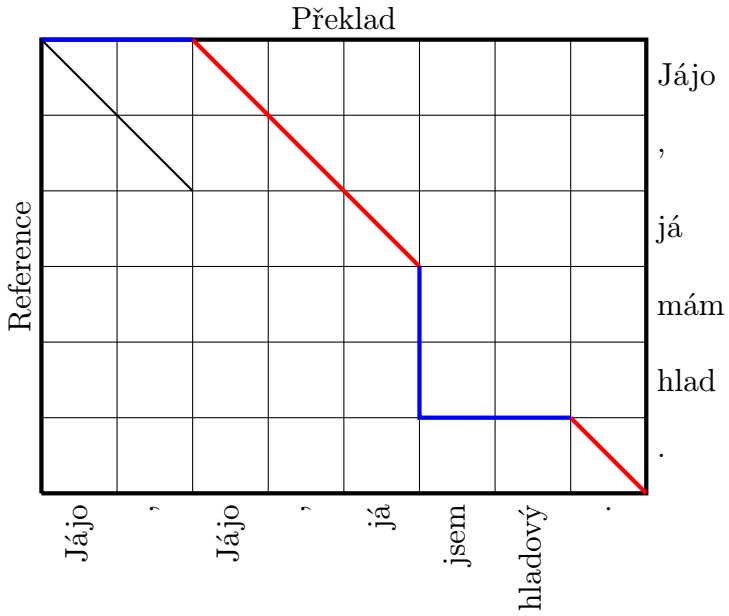
Těžší situace nastane, když je kandidátů na potvzený n-gram více než výskytů daného n-gramů v referenci (viz Obrázek 4.2).

Tento jev se objevuje např. u často se vyskytujících slov, předložek, spojek nebo interpunkce. V takovéto situaci by měli být vybráni takoví kandidáti, kteří se nacházejí v nejdelších možných potvrzených n-gramech. K hledaní kandidátů nacházejících se v nejdelších potvrzených n-gramech je možné použít algoritmus pro hledání nejdelší společné podposloupnosti (LCS). Nejdelší společná podposloupnost leží na nejkratší monotónní cestě v grafu, která vede z levého horního do pravého dolního rohu. V obrázku bude tato cesta znázorněna zelenou barvou.



Obrázek 4.1: Ukázka grafu porovnání dvou vět se zvýrazněnými potvrzenými n-gramy.
***Reference**: Jájo, já mám hlad.*
***Překlad**: Jájo, já jsem hladový.*

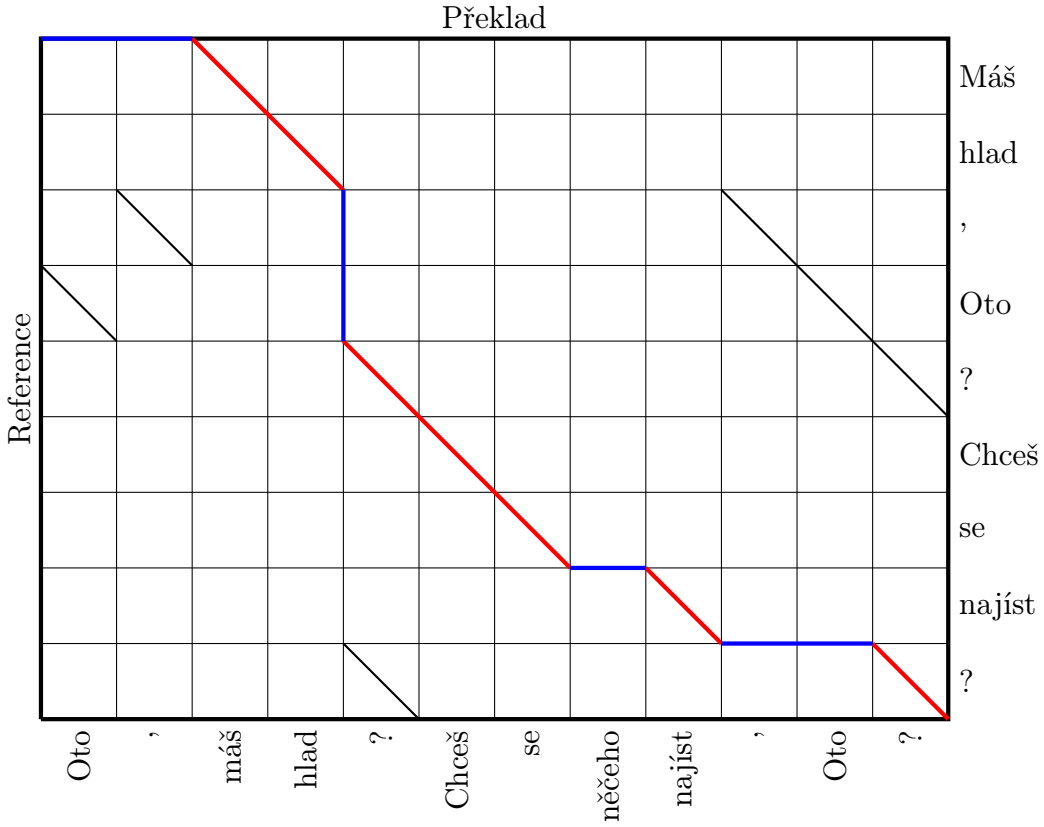
Všichni kandidáti, kteří se nacházejí na této cestě, vždy patří mezi potvrzené n-gramy (viz Obrázek 4.2).



Obrázek 4.2: Ukázka grafu s referencí a překladem, ve kterém se vyskytuje více kandidátů pro n-gram „Jájo,“. Zde je už zvýrazněna nejdelší společná podposloupnost, pomocí které je možné určit potvrzené n-gramy.
***Reference**: Jájo, já mám hlad.*
***Překlad**: Jájo, Jájo, já jsem hladový.*

Tímto postupem je možné nalézt pozice všech potvrzených n-gramů, které se

nacházejí v nejdelší společné podposloupnosti. Ale ne všechny potvrzené n-gramy se zde musí nacházet. Například při změně pořadí slov v překladu se potvrzený n-gram nemusí nacházet v nejdelší společné podposloupnosti. Tuto situaci ilustruje Obrázek 4.3.

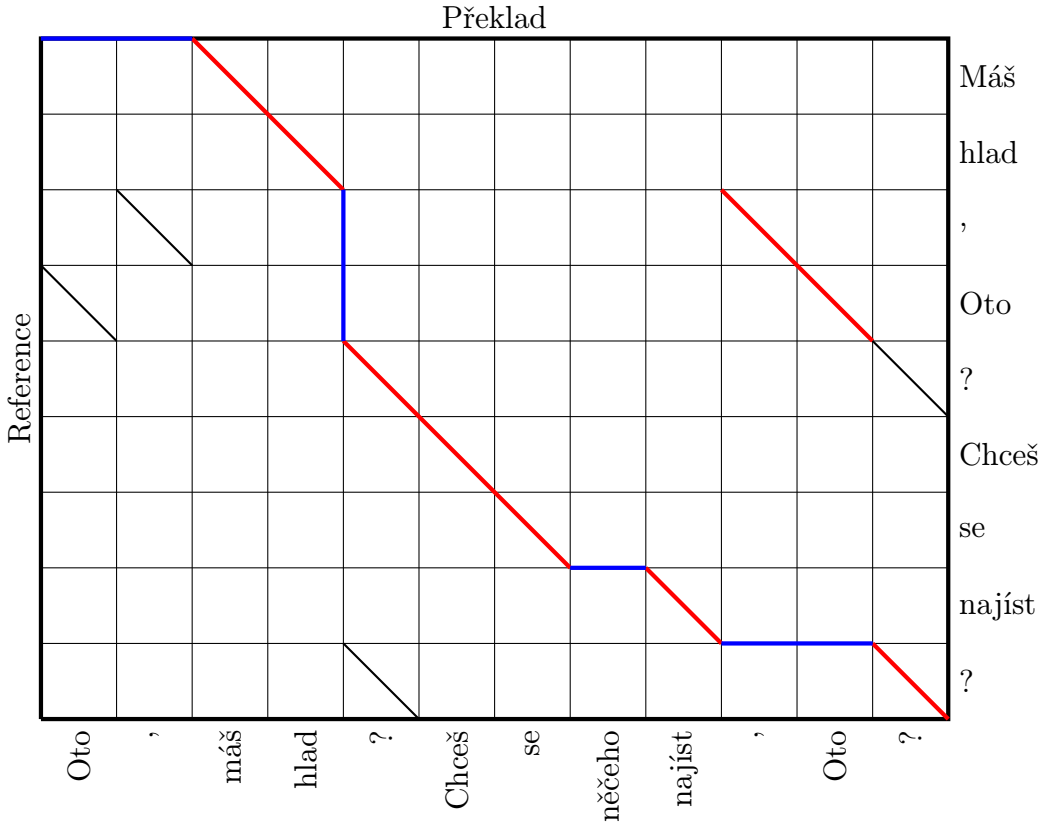


Obrázek 4.3: Ukázka grafu s referencí a překladem, ve kterém se nachází více kandidátů pro n-gramy „Oto“ a „,“.
***Reference**: Máš hlad, Oto? Chceš se najíst?*
***Překlad**: Oto, máš hlad? Chceš se něčeho najíst, Oto?*

I tato situace může nastat poměrně často, a proto musel být nalezen způsob, jak z kandidátů potvrzených n-gramů, kteří se nacházejí mimo nejdelší společnou podposloupnost, budou vybrány potvrzené n-gramy.

Při hledání algoritmu pro řešení této situace byl použit stejný předpoklad jako v minulém případě. To znamená, že z kadidátů potvrzených n-gramů budou vybráni vždy ti kandidáti, kteří se nacházejí uvnitř nejdelších potvrzených n-gramů. Každý token v překladu může být ohodnocen pomocí skóre, které určuje, v kolika kandidátech potvrzených n-gramů a již potvrzených n-gramech se daný token nachází. Z kadidátů na potvrzený n-gram pak jsou vybráni ti, kteří mají největší skóre celého n-gramu, které se spočítá jako součet skóre tokenů v n-gramu. Aby dále nebyli zvýhodňováni kandidáti, kteří byli součástí delších nezvolených kandidátů, musí být po každé volbě potvrzených n-gramů upraveno skóre u tokenů patřících do kandidátů potvrzených, kteří nebyli zvoleni za potvrzené n-gramy. Výsledek použití tohoto algoritmu je zachycen na Obrázku 4.4.

Aby oba přístupy (hledání pomocí nejdelší společné podposloupnosti a hledání pomocí počítání skóre jednotlivých slov) mohly být kombinovány do jednoho



Obrázek 4.4: Ukázka grafu s referencí a překladem, ve kterém se nachází více kandidátů pro n-gramy „Oto“ a „,“. Pozice těchto n-gramů byla nalezena pomocí metody s počítáním skóre. I když nalezené pozice n-gramů „Oto“ a „,“ neodpovídají lidskému úsudku, jsou tyto pozice nalezené správně vzhledem k metrice BLEU.
***Reference**: Máš hlad, Oto? Chceš se najíst?*
***Překlad**: Oto, máš hlad? Chceš se něčeho najíst, Oto?*

algoritmu (viz Algoritmus 1) a jednotlivé případy nemusely být řešeny odděleně, byla přidána bonifikace pro tokeny, které se nacházejí v nejdelší společné podposloupnosti reference a překladu. Tyto tokeny budou mít vždy vyšší skóre než jejich konkurenti, a proto nemůže dojít k situaci, že by token ležící v nejdelší společné podposloupnosti nebyl částí potvrzeného n-gramu. Nalezení potvrzených n-gramů[1] pak může být provedeno pouze pomocí počítání skóre.

Avšak i kombinace těchto dvou přístupů nemusí vždy fungovat. Jelikož tento algoritmus počítá pouze s n-gramy do délky čtyř tokenů, potvrzené n-gramy, které se nenacházejí v nejdelší společné podposloupnosti a jsou delší než sedm tokenů, nemusí být správně označeny za potvrzené. U takovýchto n-gramů záleží pouze na pořadí, v jakém se ve větě vyskytují. Část n-gramu, který se ve větě vyskytne dříve, bude prohlášen za potvrzený n-gram.

V praxi se ale tyto případy moc nevyskytují - nestává se často, že by se ve

[1] Ve skutečnosti tento algoritmus hledá pouze pozice potvrzených 1-gramů. Hledání pozic potvrzených n-gramů různých délek je složitější úkol, protože potvrzené n-gramy se mohou překrývat (viz Obrázek 4.4), a proto se v nástroji MT-ComparEval nachází pouze aproximace hledání pozic potvrzených n-gramů pomocí hledání pozic potvrzených 1-gramů.

```
CANDIDATES := find candidate n-grams for REFERENCE and TRANSLATION
LCS := find longest common subsequence of REFERENCE and TRANSLATION

set all CANDIDATES score to 0

foreach CANDIDATE in LCS do
    foreach WORD in CANDIDATE do
        increase WORD's score by 1
    done
done


for LENGTH in 4..1 do
    foreach CANDIDATE in CANDIDATES with LENGTH do
        foreach WORD in CANDIDATE do
            increase WORD's score by 1
        done
    done

    foreach CANDIDATE in CANDIDATES with LENGTH do
        N = count CANDIDATE's occurences in REFERENCE
        CONFIRMED = get top N CANDIDATE by their WORD score sum
        confirm CONFIRMED candidates

        foreach CANDIDATE that hasn't been confirmed do
            foreach WORD in CANDIDATE do
                decrease WORD's score by 1
            done
        done
    done
done

return all CONFIRMED CANDIDATES
```

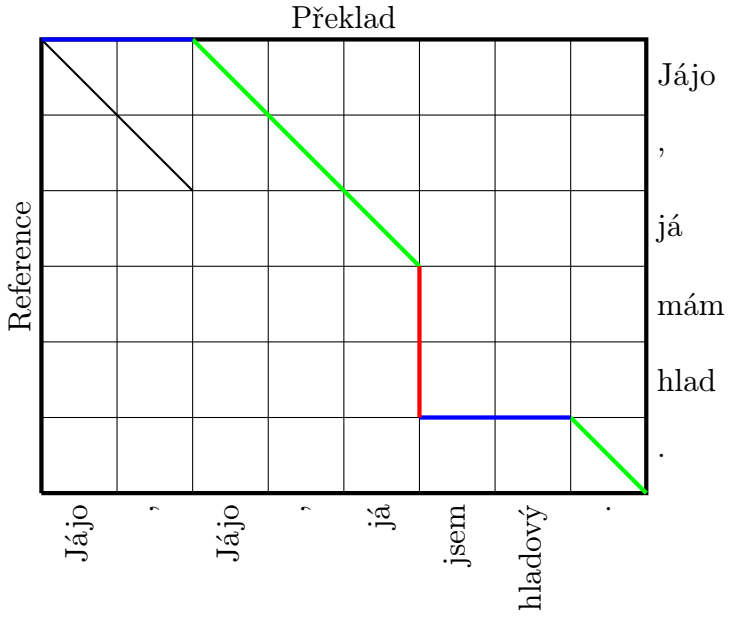
Algoritmus 1: Algoritmus ukazující kombinaci hledání potvrzených n-gramů pomocí nejdelší společné podposloupnosti a hledání potvrzených n-gramů pomocí počítání skóre pro jednotlivá slova.

větě nacházelo více kandidátů pro n-gram délky vyšší než sedm tokenů. Proto by tento algoritmus měl ve většině případů fungovat dobře.

4.2 Počítání diffu

Pro počítání diffu mezi překladem a referencí může být použit stejný algoritmus, který byl použit k hledání nejdelší společné podposloupnosti překladu a reference. Diff je stejně jako nejdelší společná podposloupnost reprezentován nejkratší monotónní cestou vedoucí z levého horního do pravého dolního rohu v grafu porovnání dvou překladů. Změna je pouze v tom, že horizontální hrany ležící na této cestě odpovídají slovům, které byly navíc vloženy do překladu. V grafu budou označeny modrou barvou. Vertikální hrany ležící na cestě reprezentující nejdelší společnou podposloupnost odpovídají tokenům, které byly navíc vloženy do reference. V grafu budou označeny červenou barvou. Diagonální hrany ležící na výše definované cestě reprezentují tokeny, které se nacházejí i v referenci i v překladu. V grafu budou označeny zelenou barvou. Výsledek počítání diffu dvou vět ukazuje Obrázek 4.5. Z takto obarvené cesty může být zjištěn rozdíl mezi porovnávanými větami a na jeho základě může být zobrazen rozdíl ve webovém prostředí (viz Obrázek 1.8).

Stejný algoritmus je možné použít i pro hledaní diffu mezi dvěma překlady.



Obrázek 4.5: Ukázka diffu reference s překladem.
***Reference**: Jájo, já mám hlad.*
***Překlad**: Jájo, Jájo, já jsem hladový.*

5. Uživatelská dokumentace

5.1 Systémové požadavky

Nástroj MT-ComparEval byl vyvinut jako webová aplikace, která poběží na serveru s Linuxem.[1] Dále musí být na uživatelském počítači nainstalované PHP 5.4 a databáze SQLite 3.

5.2 Instalace systémových požadavků

Před popisem samotné instalace nástroje MT-ComparEval bude popsán způsob, jakým je možné nainstalovat PHP 5.4 a databázi SQLite3 na operačním systému Ubuntu 12.10. Potřebné závislosti lze nainstalovat následujícím postupem:

```
sudo apt-get install sqlite3
sudo apt-get install php5-cli php5-sqlite
```

5.3 Instalace a spuštění programu

Instalace není vůbec náročná, stačí spustit script **./bin/install.sh**, který připraví vše potřebné pro běh programu. Jelikož chceme, aby uživatel nemusel používat webserver Apache nebo jiné webservery, musí si uživatel před každým spuštěním aplikace zapnout lokální server pomocí skriptu **./bin/server.sh**, který spustí aplikaci na adrese **http://localhost:8080**. Poslední krok spočívá ve spuštění programu, který kontroluje nově přidané experimenty a tasky. Ten se spustí pomocí skriptu **./bin/watcher.sh**.

V případě, že má uživatel nainstalován nějaký webový server a chce ho použít pro provoz nástroje MT-ComparEval, může ho použít jako v případě ostatních webových stránek, ale nesmí zapomenout spustit skript pro kontrolu nově přidaných experimentů a tasků.

5.4 Import experimentů

Aby uživatel mohl porovnávat překlady, musí nejprve vytvořit experiment, v jehož rámci bude jednotlivé překlady vyhodnocovat. Každý experiment je uložen ve vlastním podadresáři adresáře **./data**, do kterého uživatel musí nahrát zdrojový text a referenční překlad. Výchozí jména souborů jsou **source.txt** pro zdrojový překlad a **reference.txt** pro referenční překlad. U všech souborů s větami se předpokladá, že každá věta či souvětí je na vlastním řádku a je zachováno pořadí vět. To znamená, že první řádek v souboru se zdrojovým textem je přeložen na prvním řádku souboru s referenčním a strojovým překladem, druhý řádek v souboru se zdrojovým textem odpovídá druhému řádku v souboru s referencí atd. Pro správný běh aplikace je nutné, aby soubory se zdrojovým textem a

[1] Pokud je aplikace nainstalována na vzdáleném serveru, může být používána pomocí libovolného moderního webového prohlížeče na libovolném operačním systému.

referenčním překladem měly stejný počet řádků. Pokud nebudou mít stejný počet řádků, experiment nebude importován. Jako jméno experimentu je použito jméno adresáře.

5.4.1 Konfigurace experimentu

Výše zmíněný přístup je trochu omezený a nenabízí uživatelům možnosti konfigurace. Proto je možné, aby si uživatel v konfiguračním souboru **./config.neon** předefinoval výchozí hodnoty. Je tak možné změnit jméno experimentu, popisek experimentu nebo jména souborů, ve kterých se bude nácházet zdrojový text či referenční překlad.

Konfigurace experimentu by mohla vypadat například takto:

```
name: Nakonfigurovaný experiment
description: Ukázka konfigurace experimentu
source: zdrojovy_text.txt
reference: referencni_preklad.txt
```

Pole name odpovídá názvu experimentu, které bude použito místo jména adresáře. Description je jednořádkový popis experimentu, který odpovídá např. hlavičce commit message. Source a reference jsou relativní jména souborů, ze kterých bude načten zdrojový text resp. referenční překlad.

Informace o importu experimentu je možné nalézt jak v lokálním logu každého experimentu v souboru **./data/experiment/import.log**, tak v globálním logu všech importů v souboru **./log/import.log**. Z toho logu uživatel může poznat, jestli byl daný experiment úspěšně importován, případně k jakému problému při importu došlo.

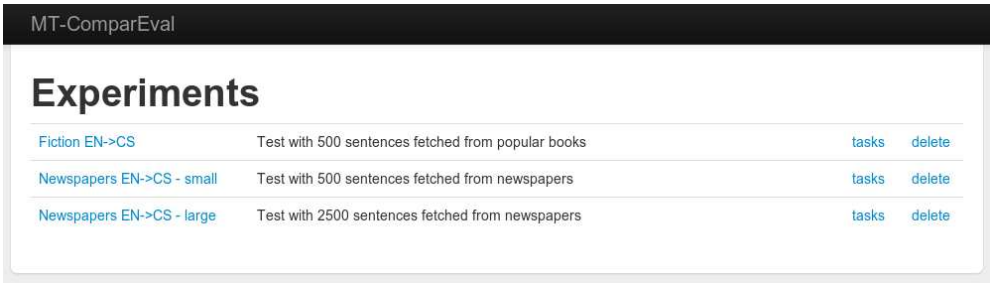
V případě, že se import nezdaří, může uživatel opravit chyby, kvůli kterým byl import přerušen, a znovu experiment nahrát do daného adresáře. Musí však smazat soubor **./data/experiment/.notimported**, který zabraňuje dalším importům u experimentů, při jejichž importu došlo k chybě.

5.5 Import tasků

Tasky jsou nahrávány jako podadresáře v adresáři experimentu, s jehož referenčním překladem má být daný překlad porovnán. Jediný soubor, který musí být do tohoto podadresáře nahrán, je soubor **translation.txt**, ve kterém budou všechny přeložené věty. Opět je nutné, aby soubor s přeloženými větami měl stejný počet řádků jako soubor s referenčním překladem. Výchozí jméno tasku je stejné jako jméno podadresáře, ve kterém se nachází.

5.5.1 Konfigurace tasku

Stejně jako je možné konfigurovat experiment, je možné konfigurovat i task. Task má konfigurovatelné jméno (name), popisek (description), relativní cestu k souboru s překladem (translation) a volbu, zda mají být předpočítány zlepšující a zhoršující n-gramy (precompute_ngrams). U všech tasků, pokud uživatel neřekne



Obrázek 5.1: Přehled vytvořených experimentů v nástroji MT-ComparEval.

jinak, jsou tyto n-gramy předpočítány. Konfigurace tasku může vypadat například takto:

```
name: Nakonfigurovaný task
description: Ukázka konfigurace task
translation: preklad.txt
precompute_ngrams: false
```

Záznam o průběhu importu je uložen v adresáři tasku v souboru **import.log**, pomocí něhož mohou být odhaleny případné problémy při importu a následně mohou být odstraněny stejným způsobem, jako tomu bylo při importu experimentů.

V obou konfiguračních souborech není nutné konfigurovat všechny položky. Pokud bude některá položka vynechána, bude použita její výchozí hodnota.

5.6 Webové prostředí

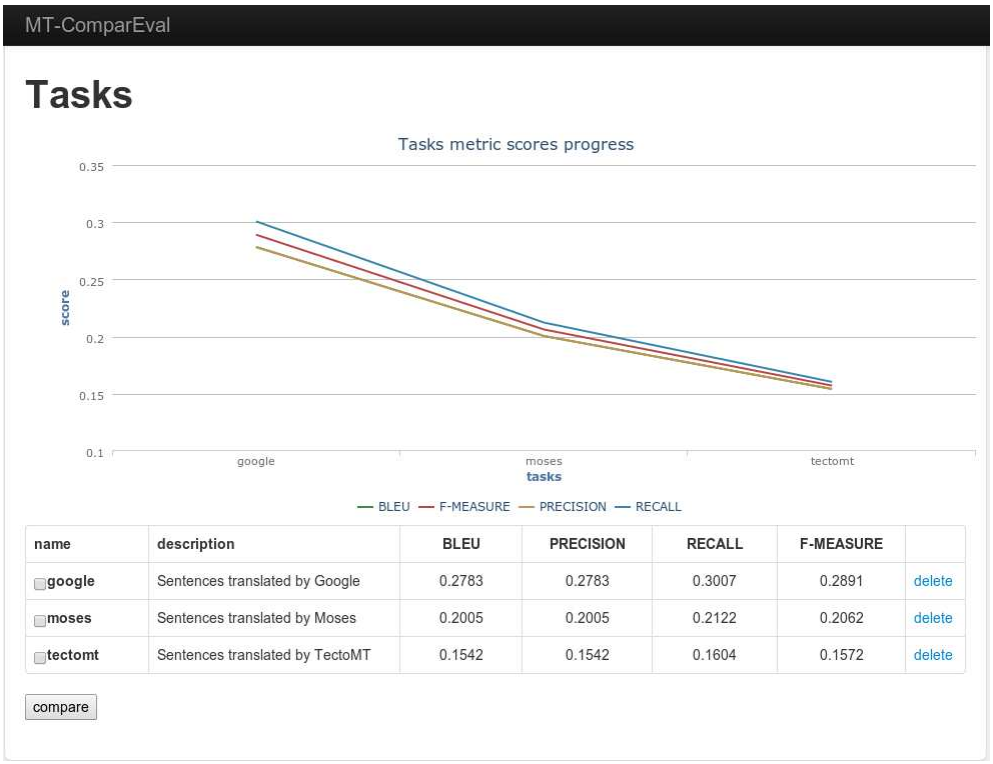
Porovnávání tasků je realizováno pomocí webové aplikace, která při zapnutém lokálním serveru běží na adrese **localhost:8080**. Uživatel si nejprve musí zvolit experiment (viz Obrázek 5.1), jehož tasky chce porovnávat, a následně si z těchto tasků vybere dva k porovnání (viz Obrázek 5.2).

V případě, že uživatel již v budoucnu nebude chtít používat daný experiment nebo task, může jej smazat pomocí příslušného odkazu ve webovém prostředí.

5.6.1 Porovnávání tasků

K porovnání dvou tasků je možné přistoupit z několika pohledů. Každému pohledu odpovídá jedna záložka v horním menu stránky s porovnáním. Kvalitu překladu tasků je možné posuzovat na základě překladu jednotlivých vět, vypočtených metrik nebo nalezených zlepšujících resp. zhoršujících n-gramů. Při porovnávání si uživatel může zvolit metriku, podle níž se bude řídit výpis vět nebo grafů metrik. Změnou metriky se tak mění pořadí vět, grafy s hodnotami z nepárového bootstrap resamplingu, graf s hodnotami z párového bootstrap resamplingu a graf s rozdělením rozdílů hodnot metriky na úrovni jednotlivých vět. Ve výchozím nastavení se věty zobrazují od nejvíce zlepšujících po nejvíce zhoršující, ale toto pořadí lze stejně jako metriku snadno změnit.[2]

[2] Míra zlepšení daného překladu je určena rozdílem aktivní metriky u porovnávaných překladů.



Obrázek 5.2: Přehled vytvořených tasků v jednom z experimentů v nástroji MT-ComparEval.

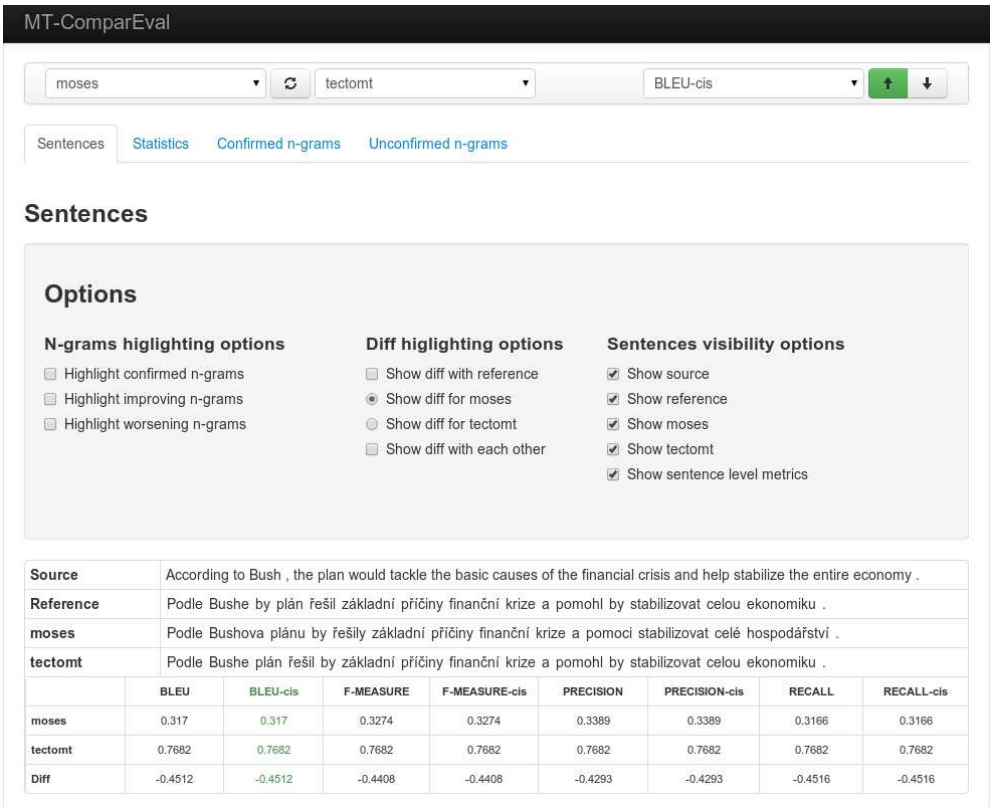
V případě potřeby je možné přímo v porovnání měnit, jaké tasky se mají porovnávat, což může být výhodné v případě, že uživatele zajímá, jak by jeden z porovnávaných tasků obstál v porovnání s jiným.

Věty

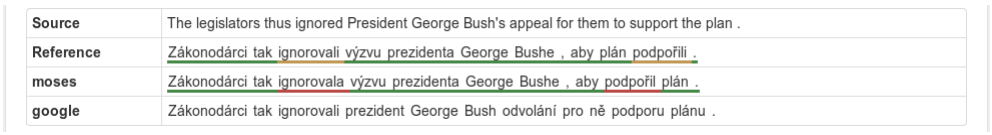
Záložka **Sentences** (viz Obrázek 5.3) slouží k zobrazení všech vět z obou překladů. U každé věty se zobrazuje zdrojový text, referenční překlad, oba porovnávané překlady a metriky jednotlivých překladů. Každou z těchto informací je možné zobrazit/skrýt pomocí panelu **Options**. Jak již bylo zmíněno dříve, věty jsou seřazeny podle rozdílu aktivní metriky, kterou je možné kdykoliv změnit. Věty se po každé změně metriky načítají znovu, aby vždy byly správně seřazeny. Načítání vět probíhá po částech, s tím jak uživatel posouvá stránku, nové věty se načítají až ve chvíli, kdy se dostane na konec stránky.

V panelu **Options** je možné zapnout i zvýraznění potvrzených n-gramů, zlepšujících nebo zhoršujících n-gramů (viz Obrázek 5.4). N-gramy jsou zvýrazněny pomocí barvy pozadí. Všechny potvrzené n-gramy jsou zvýrazněny pastelovou barvou v referenčním i porovnávaném překladu, zlepšující n-gramy jsou zvýrazněny v porovnávaných překladech sytou barvou a zhoršující n-gramy jsou zvýrazněny pastelovou červenou barvou.

Kromě zvýraznění n-gramů lze zobrazit diff mezi referenčním překladem a jedním z porovnávaných překladů (viz Obrázek 5.5). Diff je zvýrazněn pomocí barevného podtržení jednotlivých slov, zelené podtržení znamená, že se podtržené slovo nachazí v referenčním i porovnávaném překladu, a červené podtržení znamená, že se podtržené slovo nachází pouze v jednom z překladů.



*Obrázek 5.3: Záložka **Sentences** v nástroji MT-ComparEval. V horní části lze přepínat mezi jednotlivými tasky (na obrázku jsou vybrané tasky moses a tectomt), metrikou určenou pro řazení vět (na obrázku je vybrána metrika BLEU-cis) a pořadím zobrazovaných vět (na obrázku jsou věty řazeny od nejhorší po nejlepší). V panelu **Options** lze: zobrazit/skrýt zdroj, referenci, jednotlivé překlady i metriky vypočítané pro jednotlivé věty, zapnout/vypnout zvýraznění potvrzených, zlepšujících, zhoršujících n-gramů a zapnout/vypnout zvýraznění diffu.*



Obrázek 5.4: Porovnání překladu a reference se zvýrazněným diffem v nástroji MT-ComparEval. Zeleně jsou podtržena slova, která se nachází v nejdelší společné podposloupnosti překladu a reference. Slova, která se v této podposloupnosti nenachází, jsou podtržena v překladu červenou barvou a v referenci žlutou barvou.

Obrázek 5.5: Porovnání dvou překladů v nástroji MT-ComparEval. V jednotlivých překladech jsou pastelovými odstíny žluté a modré barvy zvýrazněny potvrzené n-gramy, sytými odstíny žluté a modré barvy jsou zvýrazněny zlepšující n-gramy a červenou barvou jsou zvýrazněny zhoršující n-gramy. V referenci jsou zelenou barvou zvýrazněny potvrzené n-gramy, které se nacházejí v obou překladech, pastelovým odstínem modré barvy jsou zvýrazněny zlepšující n-gramy ze systému ***google*** *a pastelovým odstínem žluté barvy jsou zvýrazněny zlepšující n-gramy ze systému* ***moses****.*

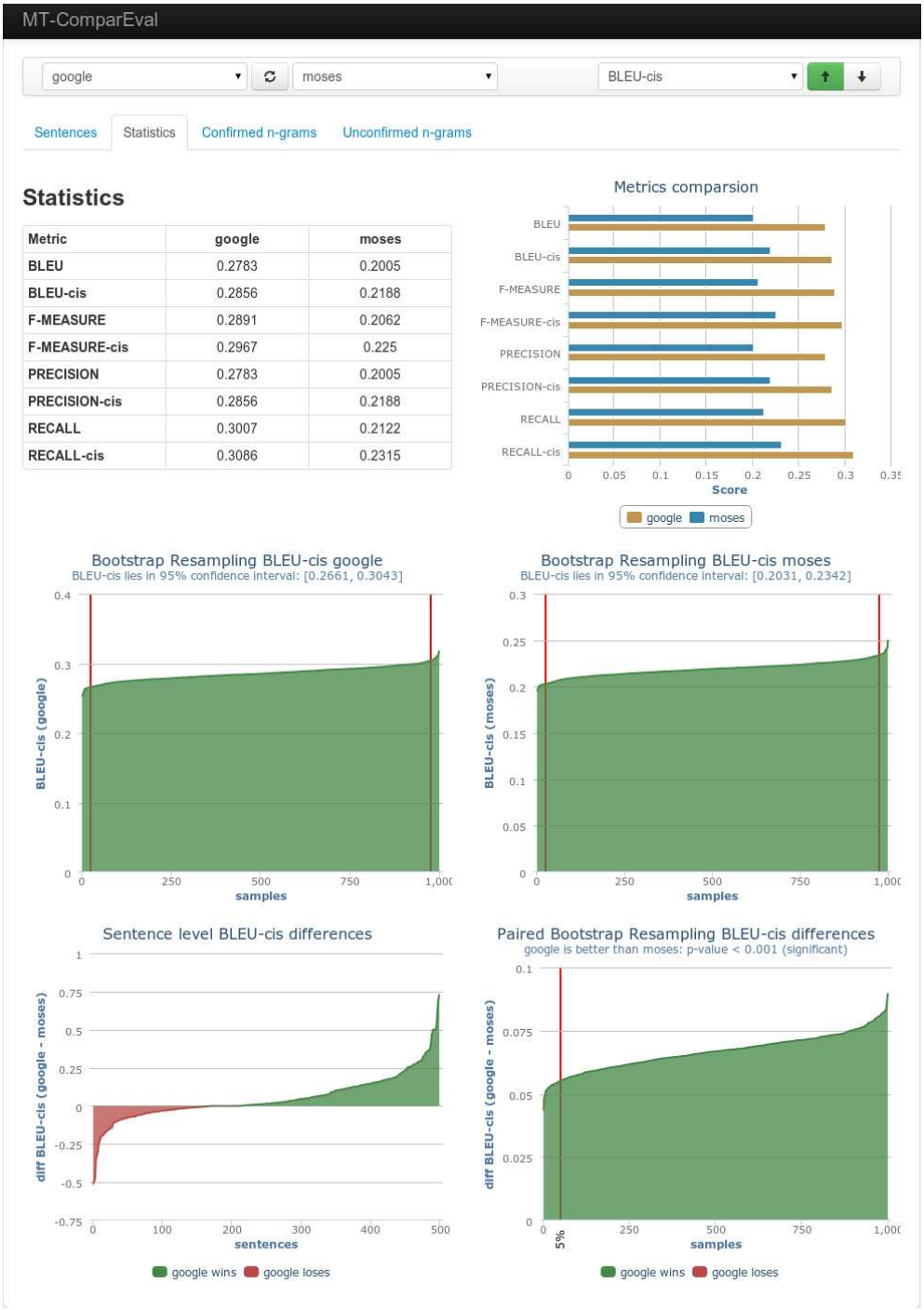
Statistiky

V záložce **Statistics** (viz Obrázek 5.6) může uživatel nalézt porovnání všech spočtených metrik pro oba tasky. Zároveň s tímto porovnáním se zde nachází i graf s rozdíly hodnot metrik na úrovni jednotlivých vět, který určuje rozdělení těchto rozdílů. Druhý graf, který se nachází na této záložce, je graf hodnot z bootstrap resamplingu, na jehož základě je možné určit, který z tasků je signifikantě lepší.

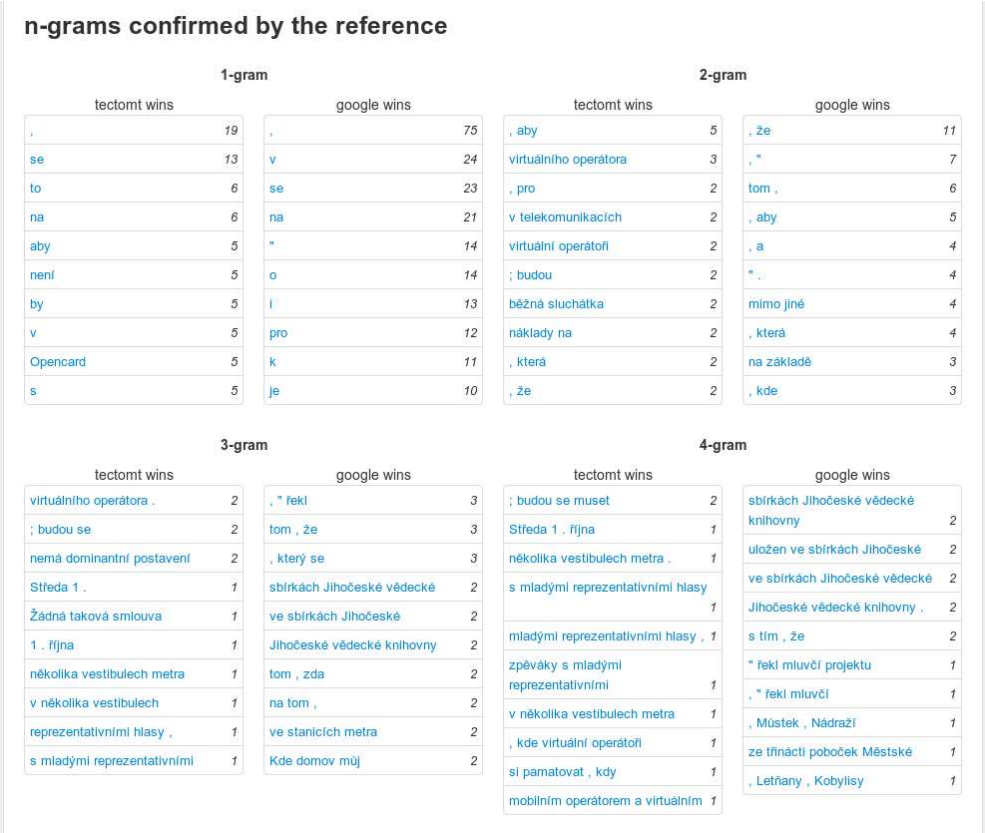
Zlepšující a zhoršující n-gramy

Na záložkách **Confirmed n-grams** a **Unconfirmed n-grams** jsou vypsány tabulky s nejvíce zlepšujícími resp. zhoršujícími n-gramy (viz Obrázek 5.7). Kliknutím na některý z n-gramů jsou zobrazeny věty, ve kterých se daný zlepšující resp. zhoršující n-gram nachází.

Tento n-gram je ve větách zvýrazněn černým rámečkemi (viz Obrázek 5.8), aby bylo na první pohled patrné, kde se nachází. Věty se zlepšujícími resp. zhoršujícími n-gramy nejsou řazeny podle vybrané metriky, ale podle počtu výskytů daného n-gramu ve větě. Uživatel se v případě potřeby může snadno vrátit k výpisu všech vět, pomocí příslušného odkazu.



Obrázek 5.6: V záložce ***Statistics*** *se nachází tabulka a graf s porovnáním metrik pro jednotlivé tasky, grafy s hodnotami z nepárového bootstrap resamplingu pro zjištění 95% intervalu spolehlivosti vybrané metriky, graf s distribucí rozdílů hodnot metrik spočtených pro jednotlivé věty a graf s hodnotami z párového bootstrap resamplingu.*



n-grams confirmed by the reference

1-gram

tectomt wins		google wins	
,	19	,	75
se	13	v	24
to	6	se	23
na	6	na	21
aby	5	"	14
není	5	o	14
by	5	i	13
v	5	pro	12
Opencard	5	k	11
s	5	je	10

2-gram

tectomt wins		google wins	
, aby	5	, že	11
virtuálního operátora	3	, "	7
, pro	2	tom ,	6
v telekomunikacích	2	, aby	5
virtuální operátoři	2	, a	4
; budou	2	" .	4
běžná sluchátka	2	mimo jiné	4
náklady na	2	, která	4
, která	2	na základě	3
, že	2	, kde	3

3-gram

tectomt wins		google wins	
virtuálního operátora .	2	, " řekl	3
; budou se	2	tom , že	3
nemá dominantní postavení	2	, který se	3
Středa 1 .	1	sbírkách Jihočeské vědecké	2
Žádná taková smlouva	1	ve sbírkách Jihočeské	2
1 . října	1	Jihočeské vědecké knihovny	2
několika vestibulech metra	1	tom , zda	2
v několika vestibulech	1	na tom ,	2
reprezentativními hlasy ,	1	ve stanicích metra	2
s mladými reprezentativními	1	Kde domov můj	2

4-gram

tectomt wins		google wins	
; budou se muset	2	sbírkách Jihočeské vědecké knihovny	2
Středa 1 . října	1	uložen ve sbírkách Jihočeské	2
několika vestibulech metra .	1	ve sbírkách Jihočeské vědecké	2
s mladými reprezentativními hlasy	1	Jihočeské vědecké knihovny .	2
mladými reprezentativními hlasy ,	1	s tím , že	2
zpěváky s mladými reprezentativními	1	" řekl mluvčí projektu	1
v několika vestibulech metra	1	, " řekl mluvčí	1
, kde virtuální operátoři	1	, Můstek , Nádraží	1
si pamatovat , kdy	1	ze třinácti poboček Městské	1
mobilním operátorem a virtuálním	1	, Letňany , Kobylisy	1

Obrázek 5.7: Přehled nejvíce zlepšujících n-gramů v jednotlivých překladech.

You are displaying sentences with improving n-gram **mimo jiné**. Show all senteces

Source	The advantage of the music players , on the other hand , is , among other reasons , that they are easy to operate .
Reference	Výhodou přehrávačů zase je mimo jiné snadná obsluha .
tectomt	Výhoda hudeb hráčů na druhé straně je mezi jinými důvody , že jsou snadní fungovat .
google	Výhodou přehrávačů hudby , na druhé straně , je **mimo jiné** z důvodů , že jsou snadno ovladatelná .

Source	At the same time , the financial group Penta , which owns , among others , U : fon , stands behind Mobilking .
Reference	Přitom za Mobilkingem stojí finanční skupina Penta , která v Česku vlastní mimo jiné U : fona .
tectomt	Na stejné době finanční skupina Penta , která má mezi ostatními , U : fon stojí za Mobilking .
google	Ve stejné době , finanční skupina Penta , která vlastní **mimo jiné** , U : fon , stojí za MOBILKING .

*Obrázek 5.8: Výpis vět, ve kterých se nachází zlepšující n-gram **„mimo jiné“**.*

6. Programátorská dokumentace

MT-ComparEval slouží k porovnávání a vyhodnocování strojových překladů. Hlavním požadavkem při vývoji aplikace bylo jednoduché spuštění aplikace na vývojářově počítači. Proto byly použity technologie, které jsou běžně dostupné na většině linuxových distribucí. Jako hlavní programovací jazyk byl použit jazyk PHP ve verzi 5.4 s použitím frameworku Nette[1] a jako databáze byla použita SQLite3.[2] Pokud by měla aplikace zpracovávat velké množství překladů, lze použít i jiné databáze, ale je třeba přizpůsobit konfiguraci aplikace. Jazyk PHP ve verzi 5.4 má v sobě obsažen jednoduchý webový server, tudíž je možné aplikaci bez větších obtíží na vývojářově počítači spustit. Místo serveru, který je obsažen v PHP 5.4, lze použít libovolný jiný webserver (např. Apache HTTP Server[3]). Na frontendu byl použit javascript s frameworkem AngularJS.[4]

MT-ComparEval je webová aplikace skládající se ze tří částí:

- serverové části pro import experimentů a tasků,
- serverové části pro vykreslování šablon a REST API,
- frontendové části pro interaktivní porovnávání překladů.

6.1 Import experimentů

Všechny experimenty, které jsou určeny k importu do nástroje MT-ComparEval, jsou ukládány do adresáře **./data** v kořenu aplikace. Pokud je do tohoto adresáře vložen nový experiment, proces (**./bin/watcher.sh**), který hlídá změny v tomto adresáři, spustí další proces, který tento experiment importuje.

Při importu experimentu jsou nahrány všechny zdrojové věty a referenční překlady [5] do databáze, aby je bylo možné později použít při importu tasků nebo zobrazování výsledků.

Po úspěšném importu je do adresáře s experimentem přidán soubor `.imported`, který označuje úspěšně importované experimenty. Tento soubor je důležitý proto, aby nebyly importovány tasky v ještě neimportovaných experimentech. Zároveň je tento soubor využit k tomu, aby některý experiment nebyl importován vícekrát.

Při importu experimentů může dojít k různým chybám, které mohl způsobit uživatel (např. různý počet řádků v souborech se zdrojovými a referenčními větami) i nástroj MT-ComparEval (např. málo paměti). V případě, že dojde k nějaké chybě, je do adresáře s experimentem přidán soubor `.notimported`. Tento soubor označuje neúspěšně importované experimenty, které nemají být znovu importovány. Když uživatel opraví chybu,[6] která způsobila neúspěch importu, může tento

[1] http://www.nette.org
[2] http://www.sqlite.org
[3] http://httpd.apache.org
[4] http://www.angularjs.com
[5] V současné době je možné mít pouze jeden referenční překlad v experimentu, protože ve veřejně dostupných datech ze soutěží WMT je k dispozici pouze jeden [5, 6, 7].
[6] Informace o chybě, která způsobila, že experiment nebyl importován, se nacházejí v logu(viz Kapitola 6.2.6).

soubor odstranit a nástroj MT-ComparEval se pokusí daný experiment znovu importovat.

6.2 Import tasků

Tasky, které jsou určeny k importu do nástroje MT-ComparEval, jsou uloženy do adresářů experimentů, ke kterým daný task patří. Změny v těchto adresářích hlídá stejný proces (**./bin/watcher.sh**), který hledá nové experimenty. Pokud nalezne nový task, spustí další proces, který daný task importuje.

Při importu tasku jsou všechny přeložené věty nahrány do databáze. Pro každou větu jsou předpočítány hodnoty, které jsou později zobrazeny na frontendu. Mezi tyto hodnoty patří:

- metriky pro celý překlad i jednotlivé věty,
- hodnoty pro párový bootstrap resampling a
- nejvíce zlepšující/zhoršující n-gramy.

Import jednoho tasku probíhá ve třech krocích:

- načtení a předzpracování vět,
- zpracování jednotlivých vět a
- uložení vypočtených hodnot do databáze.

6.2.1 Načtení a předzpracování vět

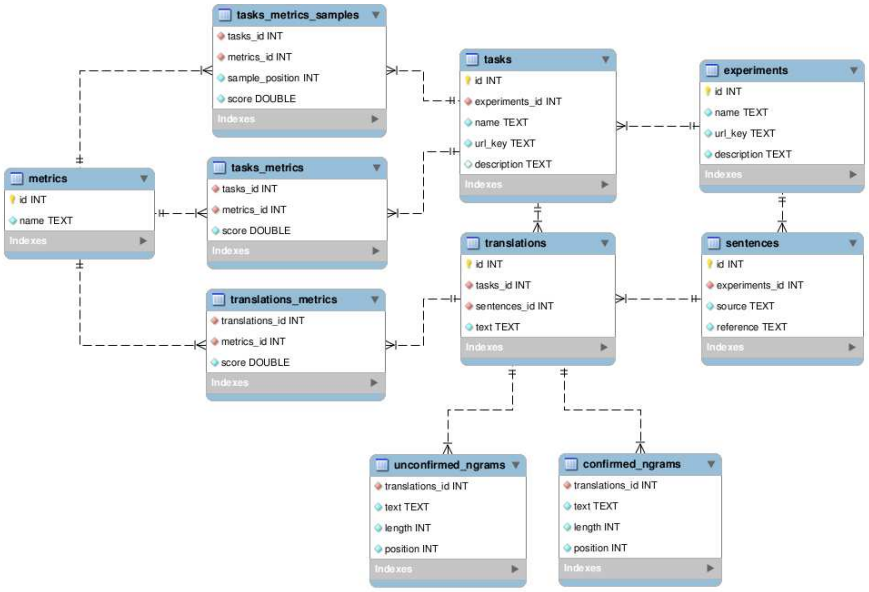
Aby při importu tasků nemusely být do paměti nahrávány celé soubory s překladem, používá nástroj MT-ComparEval iterátory, pomocí kterých je možné načítat a upravovat jednotlivé věty ze souboru. Pomocí iterátorů je také implementováno předzpracování vět.[7] Během předzpracování vět jsou všechny věty tokenizovány[8] (všechna slova jsou oddělena mezerou) a také vypočteny všechny potvrzené a nepotvrzené n-gramy, které budou dále použity při zpracování vět.

6.2.2 Zpracování vět

Hlavním cílem při zpracování vět je spočítat metriky jak pro jednotlivé věty, tak pro celé překlady. Metriky jsou automaticky počítány v páru – case-sensitive (s ohledem na velikost písmen) a case-insensitive (reference i překlad jsou převedeny na malá písmena). Tradičně se uvádí hlavně case-insensitive metrika, protože chyba v překladu velkých písmen není tak závažná. Porovnáním výsledků case-sensitive a case-insensitive metrik lze tuto chybu odhalit. Nástroj MT-ComparEval také umožňuje implementaci vlastních metrik. O tom, jak je možné implementovat vlastní metriku, pojednává Kapitola 6.5.

[7] Je použita obdoba funkce `map` z funkcionálních jazyků.

[8] Způsob tokenizace je stejný jako v nástroji **mteval-11b.pl** – ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v11b.pl.



Obrázek 6.1: Schéma databáze, kterou používá nástroj MT-ComparEval.

Před samotným uložením vypočítaných metrik do databáze jsou dopočítány poslední informace, které také budou použity na frontendu. Mezi tyto informace patří hodnoty vzorků z párového bootstrap resamplingu (viz Kapitola 6.2.4) a nejvíce zlepšující a zhoršující n-gramy (viz Kapitola 6.2.5).

6.2.3 Uložení vypočtených hodnot do databáze

Hodnoty vypočtené při vyhodnocování tasku jsou na závěr uloženy do databáze. Celý proces ukládání probíhá v jedné transakci, což zabezpečí, že v případě, kdy dojde k chybě při importu, nebudou neúspěšně importované experimenty a tasky zobrazeny na frontendu. Celý proces ukládání do databáze je také díky zabalení do transakce mnohem rychlejší, než kdyby se data ukládala mimo transakci. Schéma databáze, do které jsou předpočítaná data uložena, je zobrazeno na Obrázku 6.1.

6.2.4 Párový bootstrap resampling

Aby na frontendu bylo možné porovnávat dva překlady pomocí párového bootstrap resamplingu, musí být předgenerovány náhodné vzorky. Pro porovnání dvou překladů pomocí této metody je nutné, aby pro porovnávané překlady byly vzorky vygenerovány se stejnými větami. Proto jsou pro každý experiment náhodně vygenerována čísla vět pro každý vzorek. Tato čísla jsou pak zafixována a použita pro všechny tasky daného experimentu.[9] Pro takto vygenerované vzorky vět jsou spočteny všechny metriky a z jejich výsledků je na frontendu vykreslen graf.

[9] Ve skutečnosti není potřeba zafixovávat čísla vět, při generování vzorků stačí nastavit vždy stejné jadérko pro generátor náhodných čísel.

6.2.5 Hledání nejvíce zlepšujících a zhoršujících n-gramů

Jelikož výpočet nejvíce zlepšujících n-gramů trvá relativně dlouho ($\approx$ 1 minutu pro 1000 vět) a na frontendu se zobrazuje pouze 10 nejlepších, je třeba, aby nejvíce zlepšující n-gramy byly předpočítány. Při importu tasku jsou proto předpočítány nejvíce zlepšující n-gramy pro porovnání daného tasku se všemi ostatními tasky, které se nacházejí ve stejném experimentu. Předpočítání lze pro jednotlivé tasky vypnout. Algoritmus pro hledání nejvíce zlepšujících n-gramů je vysvětlen v části Algoritmus 2. Pomocí obdobného postupu se také hledají nejvíce zhoršující n-gramy.

Protože předpočítání nejvíce zlepšujících/zhoršujících n-gramů pro všechny tasky může být časově náročné, je možné toto předpočítání v konfiguraci jednotlivých tasků vypnout. V porovnání dvou tasků pak nebudou předpočítané n-gramy ihned k dispozici a budou se muset při prvním požadavku dopočítat.

6.2.6 Logování importů

Všechny důležité operace spojené s importem tasků či experimentů (načítání konfiguračního souboru, načítání vět ze souborů, počítání metrik, ukládání do databáze atd.), jsou logovány do souboru, z kterého je možné vyčíst, proč nebyl některý task či experiment úspěšně importován.

6.3 REST API

REST API je implementované v jazyce PHP s použitím frameworku Nette. Toto API slouží k předávání předpočítaných informací frontendu, který si je dle potřeby získává za použití AJAXu. API vždy vrací data ve formátu JSON, který lze snadno použít v Javascriptu. Pomocí API navíc mohou být získávána pouze data, která v danou chvíli frontend potřebuje. To znamená, že se nemusí načítat všechny věty naráz, ale mohou být načítány postupně, s tím, jak si je uživatel prohlíží. Stejně tak nemusí být stahována všechna data pro grafy, ale stačí stáhnout data pouze pro právě vybranou metriku.

Pomocí API je možné získat většinu dat pro porovnání dvou překladů, ať už se jedná o dostupné metriky, data pro vykreslení grafů právě vybrané metriky, nejvíce zlepšující a zhoršující n-gramy nebo věty, které mohou být řazeny podle libovolné metriky.

6.4 Frontend

Na frontendu byl použit CSS framework Bootstrap[10] od firmy Twitter, javascriptový framework AngularJS od firmy Google a knihovna Highcharts[11] pro vykreslování grafů.

Na frontendu se nacházejí tři typy stránek:

- seznam všech experimentů,

[10] http://twitter.github.io/bootstrap/
[11] http://www.highcharts.com

```
foreach SENTENCE in EXPERIMENT do
    CONFIRMED_NGRAMS_A := confirmed n-grams for TASK_A in SENTENCE
    CONFIRMED_NGRAMS_B := confirmed n-grams for TASK_B in SENTENCE

    sort CONFIRMED_NGRAMS_A lexicographically
    sort CONFIRMED_NGRAMS_B lexicographically

    while CONFIRMED_NGRAMS_A and CONFIRMED_NGRAMS_B are not empty do
        CONFIRMED_A := first n-gram in CONFIRMED_NGRAMS_A
        CONFIRMED_B := first n-gram in CONFIRMED_NGRAMS_B

        if CONFIRMED_A > CONFIRMED_B then
            set CONFIRMED_B as improving in TASK_B
            pop first n-gram from CONFIRMED_NGRAMS_B
        else if CONFIRMED_A < CONFIRMED_B then
            set CONFIRMED_A as improving in TASK_A
            pop first n-gram from CONFIRMED_NGRAMS_A
        else then
            pop first n-gram from CONFIRMED_NGRAMS_A
            pop first n-gram from CONFIRMED_NGRAMS_B
        done
    done

    while CONFIRMED_NGRAMS_A is not empty do
        CONFIRMED_A := first n-gram in CONFIRMED_NGRAMS_A
        set CONFIRMED_A as improving in TASK_A

        pop first n-gram from CONFIRMED_NGRAMS_A
    done

    while CONFIRMED_NGRAMS_B is not empty do
        CONFIRMED_B := first n-gram in CONFIRMED_NGRAMS_A
        set CONFIRMED_B as improving in TASK_B

        pop first n-gram from CONFIRMED_NGRAMS_B
    done
done

choose top 10 improving n-grams for TASK_A by occurences
choose top 10 improving n-grams for TASK_B by occurences
```

Algoritmus 2: Algoritmus pro nalezení 10 nejvíce zlepšujících n-gramů v porovnávaných tascích.

- seznam všech tasků v daném experimentu a
- porovnání dvou tasků.

Na frontendu dochází k výpisu dat, která byla předpočítána během importu jednotlivých tasků. Pouze pozice potvrzených n-gramů a diff se počítají až při zobrazení jednotlivých vět. Algoritmus, pomocí kterého se hledají pozice potvrzených n-gramů, byl vysvětlen v Kapitole 4. Způsob, jakým jsou potvrzené n-gramy zobrazeny, je popsán v následující části.

6.4.1 Zobrazení potvrzených n-gramů a diffu pomocí CSS

Pro správné zobrazení potvrzených n-gramů a diffu je třeba, aby bylo možné u každého tokenu určit, zda se nachází v nějakém potvrzeném n-gramu, nebo se jedná o nepotvrzený 1-gram. Také je důležité, aby jednotlivá zvýraznění bylo možné libovolně kombinovat. Technicky byl tento problém vyřešen pomocí CSS tříd, kdy každému tokenu byly přiřazeny třídy v závislosti na informacích, které byly vypočítány při importu tasků. Jednotlivé kombinace lze zapnout přiřazením příslušné třídy kořenovému elementu, ve kterém se nacházejí všechny věty.

6.5 Implementace vlastních metrik

Nástroj MT-ComparEval umožňuje doprogramování vlastních metrik strojového překladu. Stačí pouze doprogramovat implementaci rozhraní **IMetrics** a zaregistrovat ji jako novou metriku.

Jednotlivé kroky budou v následující části podrobně rozebrány.

6.5.1 Rozhraní IMetrics

```
interface IMetrics {
    public function init();
    public function addSentence( $reference, $translation, $meta );
    public function getScore();
}
```

Rozhraní `IMetrics` představuje zapouzdření výpočtu jednotlivých strojových metrik. Toto rozhraní umožňuje vypočítat metriku na úrovni celého dokumentu nebo jednotlivých vět.

- `init()`
  Metoda `init` slouží k inicializaci metriky před začátkem výpočtu. Jelikož jsou instance jednotlivých metrik znovu používány, je třeba, aby byly před každým výpočtem znovu inicializovány, aby nedocházelo k vzájemnému ovlivňování jednotlivých výpočtů.

- `addSentence($reference, $translation, $meta)`
  Metoda `addSentence` slouží k předání informací o právě zpracovávané větě. Zároveň ihned s tím i vrací hodnotu metriky pro danou větu. Parametr `$reference` obsahuje tokenizovaný text reference, parametr `$translation`

obsahuje tokenizovaný text překladu a parametr `$meta` obsahuje metainformace o větě, které byly předpočítány během importu vět. Mezi tyto metainformace patří např. výčet potvrzených a nepotvrzených n-gramů, počet potvrzených n-gramů i počet n-gramů v překladu. Tyto metainformace pak jsou použity např. při výpočtu BLEU, aby nemusely být vždy počítány znovu při každém výpočtu metriky.

- `getScore()` Metoda `getScore` vrací výslednou hodnotu metriky pro celý dokument (tj. pro všechny věty přidané pomocí metody addSentence od posledního zavolání metody `init()`).

Implementaci rozhraní `IMetrics` je třeba uložit do adresáře `./libs`, odkud si ji aplikace automaticky načte.

6.5.2 Registrace nové metriky

Registrace nové metriky se skládá ze dvou kroků.

- **uložení jména nové metriky do tabulky `metrics`**
  Nejprve je nutné vymyslet jméno pro novou metriku a toto jméno uložit do tabulky `metrics`. Jelikož jsou všechny metriky počítány case-sensitive i case-insensitive, je třeba do tabulky `metrics` vložit i jméno pro case-insensitive hodnotu. Ta je určena příponou `-cis`.

  Vložení metriky BLEU do databáze lze provést následujícím způsobem:

```
$ sqlite3 app/database
sqlite> INSERT INTO metrics (name) VALUES ('BLEU');
sqlite> INSERT INTO metrics (name) VALUES ('BLEU-cis');
```

- **zaregistrování nové metriky v konfiguraci aplikace**
  Framework Nette používá ke konfiguraci aplikací konfigurační soubory ve formátu neon.[12] Nejprve je potřeba novou metriku zaregistrovat jako službu a poté přidat tuto službu konfiguraci `TasksImporteru`. Registraci nové metriky lze ukázat na příkladu registrace metriky BLEU a RECALL:

```
common:
    service:
        ...
        bleu: Bleu
        recall: Recall
        ...

        tasksImporters:
            class: TasksImporter(
                ...
                [
```

[12] Pro pochopení konfigurace je třeba přečíst http://doc.nette.org/cs/configuring#toc-vlastni-sluzby.

```
            BLEU: @bleu,
            RECALL: @recall
        ]
    )
```

Pomocí řádku `bleu: Bleu` je zaregistrována instance třídy `Bleu` jako služba se jménem `bleu`. Na tuto službu se lze v konfiguraci odkázat pomocí výrazu `@bleu`. Řádek `BLEU: @bleu` pak předá službě `tasksImporters` službu `@bleu` pod jménem `BLEU`. Toto jméno by mělo být stejné jako jméno, které bylo vloženo do tabulky `metrics`. Službě `tasksImporter` se nemusí předávat metrika se jménem `BLEU-cis`, protože převedení na malá písmena provede a uložení metriky s příslušným jménem provede `tasksImporter` automaticky.

Po přidání nové metriky je třeba, aby byly všechny tasky importovány znovu, aby pro ně mohla být tato metrika dopočítána.

6.6 Problémy při řešení

Při vývoji nástroje MT-ComparEval jsem narazil na mnoho problémů. Ať už se jednalo o volbu jazyka, frameworku či databáze.

Aby bylo možné snadno nainstalovat nástroj MT-ComparEval na vybraném počítači, byla jako databáze zvolena databáze SQLite 3. Ta vyhovuje většině požadavků. Bohužel bylo hodně času věnováno odhalování chyb, způsobených tím, že SQLite 3 zamyká zvláštním způsobem datový soubor, a proto není možné používat databázi v dalších procesech. Při normálním použití na webu se tato chyba nevyskytuje, protože procesy databázi používají pouze během zpracování HTTP požadavku. Avšak při použití dlouho trvajících procesů pro import experimentů a tasků vždy docházelo ke kolizím při přístupu k databázi, které vyústily v pád importu. Proto musel být návrh importu zcela přepracován.

V duchu TDD[13] jsem se snažil před vlastní implementací napsat testy pomocí nástroje Behat.[14] Ten slouží k BDD.[15] Pro všechny kroky importů byly napsány specifikace chování, které byly nástrojem Behat otestovány. Ovšem později se ukázalo, že tento přístup k testování nebyl úplně vhodný, a testy pomocí tohoto nástroje jsem přestal psát. Vše bylo způsobeno tím, že byla porušena pyramida testů,[16] jelikož všechny testy odpovídaly pouze funkčním testům. Lepším řešením by bylo vytvořit pro všechny funkční požadavky unit testy, jejichž spuštění by trvalo kratší dobu.

[13] Test-Driven-Development
[14] http://www.behat.org
[15] Behaviour-Driven-Development
[16] http://martinfowler.com/bliki/TestPyramid.html

7. Podobné aplikace

Vývoj nástroje MT-ComparEval byl v určitých oblastech inspirován nástroji, které už jsou volně dostupné. Z těchto nástrojů byly vybrány nejdůležitější vlastnosti, které byly sloučeny do jednoho funkčního celku. V následující části budou tyto nástroje podrobněji představeny.

7.1 mteval-11b.pl

Je skript napsaný v jazyce Perl, který umožňuje počítat metriky BLEU a NIST. Tento skript je celosvětově používaný, a proto byl použit pro kontrolu, že je metrika BLEU v MT-ComparEval počítána správně.

V současné době už existuje verze mteval-13a.pl, ve které mimo jiné přibyla možnost počítat metriky pro jednotlivé segmenty v překladu (viz Kapitola 3.2).

7.2 iBLEU

Nástroj iBLEU [11] umožňuje vyhodnocovat a porovnávat strojové překlady. Pokud uživatel nemá žadný jiný strojový překlad, se kterým by chtěl svůj překlad porovnat, může si větu nechat přeložit překladačem Google Translate[1] nebo Bing Translator[2] a porovnat svůj překlad s překladem z těchto nástrojů.

Pomocí tohoto nástroje může být metrika BLEU počítána pro celé dokumenty nebo i jednotlivé segmenty. Na základě těchto výsledků je možné si jednotlivé segmenty prohléhnout.

Pokud uživatel porovnává svůj překlad pouze s referencí, je zvýrazněn jejich diff. V případě, že uživatel porovnává dva strojové překlady, je zobrazen diff těchto překladů.

Tento nástroj je možné používat lokálně jako webovou aplikaci bez použití webového serveru, protože je celý napsán v HTML 5, CSS a javascriptu. Stejně jako MT-ComparEval i iBLEU používá pro počítání BLEU jako referenční implementaci mteval-13a.pl.

Na Obrázku 7.1 je možné vidět porování dvou překladů v nástroji iBLEU.

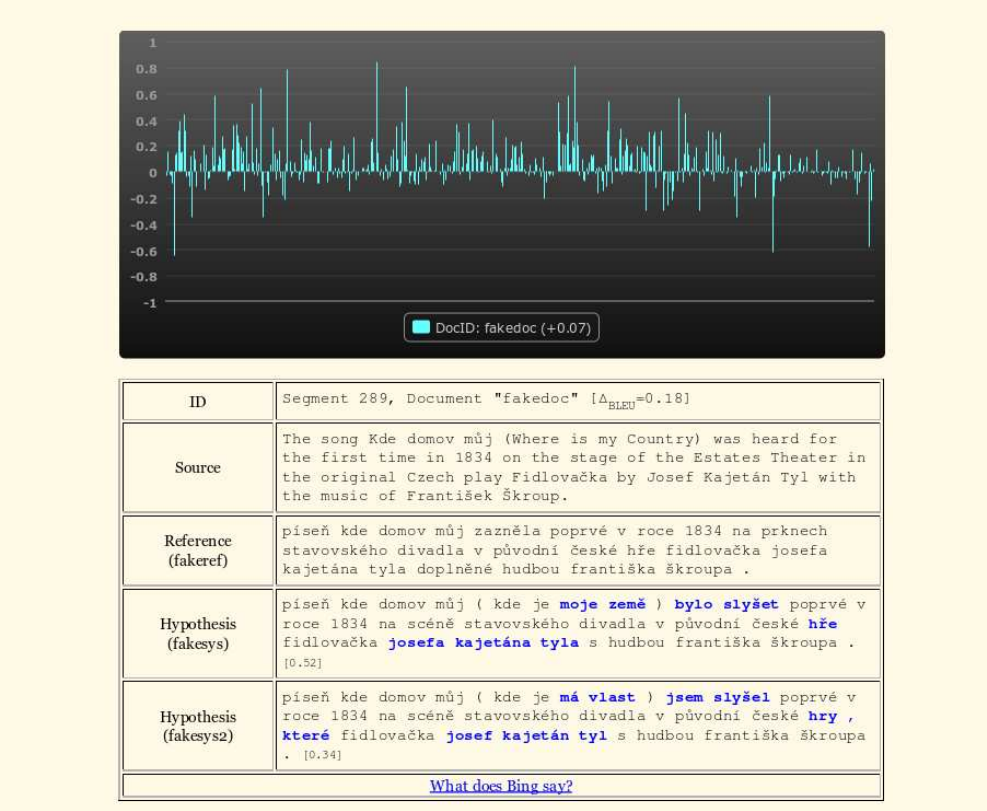
7.3 EMS – An Experimental Management System

Nástroj EMS [12], který je součástí strojového překladače Moses, obsahuje webovou aplikaci, díky níž je možné porovnávat překlady. V porovnávaných větách barevně zvýrazňuje slova na základě délky nejdelšího n-gramu, do kterého zvýrazňované slovo patří. Obrázek 7.2 ukazuje takto zvýrazněnou větu. Pro n-gramy také počítá precision a recall.

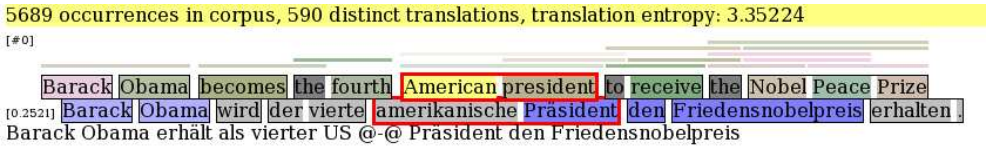
Ve webovém prostředí je také možné vyhledávat věty, ve kterých byl použit zvolený n-gram. Jednotlivá užití n-gramů jsou pak rozřazena do skupin podle

[1] Bohužel Google Translate už nenabízí bezplatné API.

[2] Bing Translator nabízí bezplatné api do limitu 2 miliónů přeložených znaků za den.



Obrázek 7.1: Porovnání dvou překladů v nástroji iBLEU.

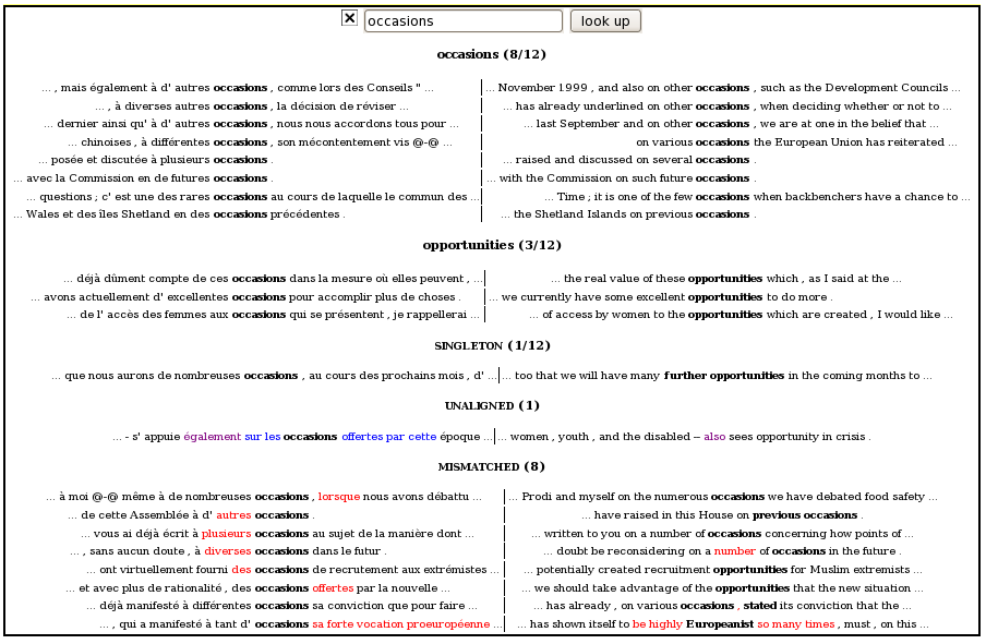


Obrázek 7.2: Zvýrazněné n-gramy podle délky v nástroji EMS. Zdroj: http://www.statmt.org/moses/img/ems-annotation.png

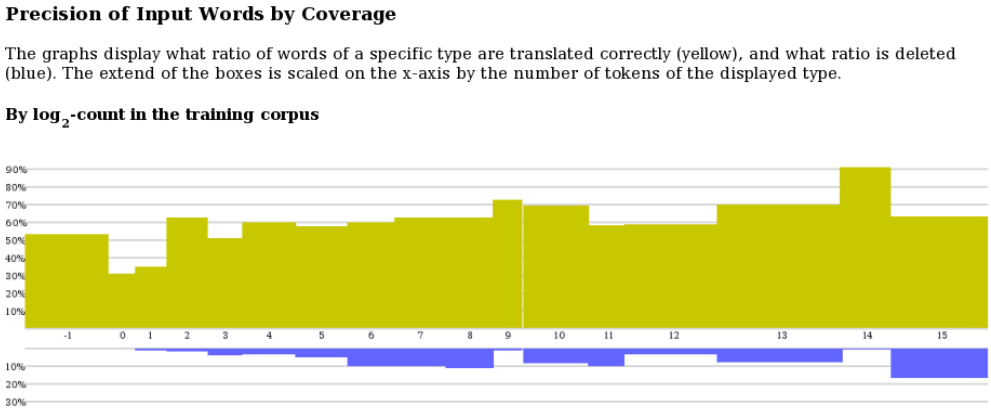
správnosti překladu (viz Obrázek 7.3), ve kterých bylo použito slovo **„occasions“**.

Další informace, které mohou být použity při porovnávání dvou překladů, jsou grafy správně přeložených vs. špatně přeložených slov. Takové grafy je možné vidět na Obrázku 7.4.

Webové prostředí nástroje EMS nabízí i další možnosti porovnání překladů, o kterých je možné se více dozvědět v dokumentaci tohoto nástroje.



*Obrázek 7.3: Výpis vět, ve kterých se vyskytuje slovo **„occasions“**, v nástroji EMS. Zdroj: http://www.statmt.org/moses/img/ems-biconcor.png*



Obrázek 7.4: Grafy správně přeložených vs. špatně přeložených slov v nástroji EMS. Zdroj: http://www.statmt.org/moses/img/ems-precision-by-coverage.png

8. Závěr

8.1 Současný stav aplikace

Nástroj MT-ComparEval je v současné době plně funkční. Umožňuje spravovat experimenty a jejich tasky. Tasky je možné porovnávat na základě různých kritérií - hodnot metrik strojového překladu, porovnání jednotlivých vět či přehledu nejvíce zlepšujících a zhoršujících n-gramů. Při porovnání překladů jedné věty je možné si nechat zvýraznit potvrzené n-gramy, zlepšující n-gramy, zhoršující n-gramy, diff mezi překlady či jeden z nejvíce zlepšujících nebo zhoršujících n-gramů.

8.2 Nápady na vylepšení

I přes to, že nástroj MT-ComparEval umožňuje již v současné verzi důkladně porovnávat dvě různé verze překladů, existují další způsoby, jak tento nástroj vylepšit.

8.2.1 Podpora více referencí

Při vyhodnocování strojových překladů a počítání metrik strojových překladů používá nástroj MT-ComparEval pouze porovnání s jednou referencí. Protože většina vět může být přeložena různými způsoby, může nastat situace, kdy daný překlad neodpovídá zvolené referenci. Aby bylo možné takové překlady lépe ohodnotit, mohly by být překlady porovnávány s několika různými referencemi. Z těchto referencí by pak byla vybrána ta, pro kterou by daný překlad dosáhl nejlepšího skóre (a pro výpočet BLEU by se použily všechny reference). Z tohoto důvodu by bylo dalším možným rozšířením přidání podpory více referencí v experimentu.

8.2.2 Více metrik

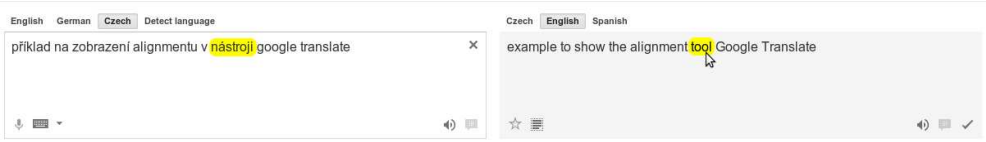
I přes to, že BLEU je nejčastěji používaná automatická metrika kvality překladu a na úrovni dokumentů koreluje dobře s lidským hodnocením [1, 2], nemusí být vždy považována za nejvhodnější pro jejich vyhodnocování [5, 6, 10]. Některé strojové překladače jsou speciálně optimalizovány na metriku BLEU, takže může být vhodné je vyhodnocovat pomocí jiné metriky. Proto by bylo dalším vhodným rozšířením vyhodnocování strojových překladačů na základě několika metrik. Mezi metriky, které by bylo možné a vhodné do nástroje MT-ComparEval doprogramovat (viz Kapitola 6.5, kde je popsáno, jak je možné do nástroje MT-ComparEval přidat novou metriku), patří např. NIST [8], METEOR [9], PORT [10], ...

8.2.3 Zobrazení alignmentu

Při porovnávání dvou překladů jedné věty je velmi užitečná funkce pro zvýraznění slov, která odpovídají danému slovu v referenci či dalším strojovém překladu –

word alignment (zarovnání slov). Díky této funkci je pak snadnější analyzovat chování strojových překladačů.

Na Obrázku 8.1 je ukázáno, jak zvýrazňuje alignment překladač Google Translate.



Obrázek 8.1: Ukázka zobrazení alignmentu v překladači Google Translate.

Seznam použité literatury

[1] Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu. *BLEU: a method for automatic evaluation of machine translation.* In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 311–318. Association for Computational Linguistics, 2002.

[2] Nicolas Stroppa, Karolina Owczarzak, Andy Way. *A cluster-based representation for multi-system MT evaluation.* In: *Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation, Skövde [Sweden]*, pp. 221–230, 2007.

[3] Chin-Yew Lin., Franz Josef Och. *ORANGE: a method for evaluating automatic evaluation metrics for machine translation.* In *Proceedings of the 20th International Conference on Computational Linguistics*, pp. 501. Association for Computational Linguistics, 2004.

[4] Philipp Koehn. *Statistical Significance Tests for Machine Translation Evaluation.* In *Proceedings of EMNLP*, vol. 4, pp. 388–395. 2004.

[5] Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, Omar F. Zaidan. *Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation.* In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pp. 17–53. Association for Computational Linguistics, 2010.

[6] Chris Callison-Burch, Philipp Koehn, Christof Monz, Omar F. Zaidan. *Findings of the 2011 Workshop on Statistical Machine Translation.* In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pp. 22–64. Association for Computational Linguistics, 2011.

[7] Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, Lucia Specia. *Findings of the 2012 workshop on statistical machine translation* In *Proceedings of the 7th Workshop on Statistical Machine Translation*, pp. 10–51. Association for Computional Linguistics, 2012.

[8] George Doddington. *Automatic evaluation of machine translation quality using n-gram co-occurrence statistics.* In *Proceedings of the Second International Conference on Human Language Technology Research*, pp. 138–145. Morgan Kaufmann Publishers Inc., 2002.

[9] Alon Lavie, Abhaya Agarwal. *Meteor: an automatic metric for MT evaluation with high levels of correlation with human judgments.* In *Proceedings of the Second Workshop on Statistical Machine Translation*, pp. 228–231. Association for Computional Linguistics, 2007.

[10] Boxing Chen, Roland Kuhn, Samuel Larkin *PORT: a precision-order-recall MT evaluation metric for tuning.* In *Proceedings of the 50th*

Annual Meeting of the Association for Computational Linguistics, pp. 930–939. Association for Computional Linguistics, 2012.

[11] Nitin Madnani *iBLEU: Interactively Debugging and Scoring Statistical Machine Translation Systems.* In *Proceedings of the 2011 IEEE Fifth International Conference on Semantic Computing*, pp. 213–214. IEEE Computer Society, 2011.

[12] Philipp Koehn *An Experimental Management System.* In *The Prague Bulletin of Mathematical Linguistics No. 94*, pp. 87–96. PBML, 2010.