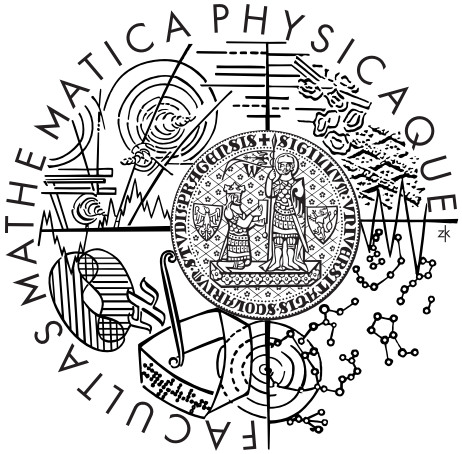
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Dušan Variš

Japonsko-český strojový překlad

Ústav formální a aplikované lingvistiky

Vedoucí bakalářské práce: RNDr. Ondřej Bojar, Ph.D.
Studijní program: Informatika
Studijní obor: Programování

Praha 2014

Rád bych poděkoval vedoucímu bakalářské práce RNDr. Ondřejovi Bojarovi Ph.D. za cenné rady, vstřícnost a trpělivost při psaní této práce.

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 Praze dne 22. května 2014 Podpis autora

Název práce: Japonsko-český strojový překlad

Autor: Dušan Variš

Ústav: Ústav formální a aplikované lingvistiky

Vedoucí bakalářské práce: RNDr. Ondřej Bojar Ph.D., Ústav formální a aplikované lingvistiky

Abstrakt: Strojový překlad s použitím hloubkového větného rozboru není v současné době ve srovnání s jinými metodami tolik rozšířen, věříme však, že některé jeho aspekty jsou schopny přispět k zlepšení kvality strojového překladu. Je přitom důležité vyzkoušet danou metodu pro různé jazykové páry, v našem případě se jednalo o dvojici japonština-čeština. Nedílnou součástí tohoto úkolu je i získání a zpracování potřebných paralelních dat. Kvůli malému množství těchto dat jsme se snažili vyzkoušet různé postupy, které by nám pomohly potřebná data nahradit. Náš systém je založen na stejném principu jako anglicko-český překladač TectoMT, v rámci této práce jsme jej implementovali do stejného prostředí. Snažili jsme se přitom zachytit alespoň základní jazykové jevy charakteristické pro japonštinu. Při zkoumání našeho systému jsme jej porovnávali s jednoduchým frázovým překladačem.

Klíčová slova: strojový překlad, tektogramatická rovina, japonština-čeština, zpracování přirozeného jazyka

Title: Japanese-Czech Machine Translation

Author: Dušan Variš

Department: Institute of Formal and Applied Linguistics

Supervisor: RNDr. Ondřej Bojar Ph.D., Institute of Formal and Applied Linguistics

Abstract: Machine translation (MT) using deep sentence analysis is not as widespread as other MT methods, however we believe that some of its aspects can contribute to the overall translation quality. It is also important to try out deep MT methods with various language pairs. In our case, we experiment with the language pair Japanese-Czech. As a part of this task, we also had to collect and process necessary parallel data. Due to a very small amount of such data being available, we were forced to devise aproaches tackling this problem. Our system is based on the same principles as the TectoMT translation system, therefore it was implemented within the same platform. In the process, we tried to capture at least some basic linguistic phenomena characteristic for Japanese. As a part of our research, we also compared our system with a simple phrase-based baseline.

Keywords: machine translation, tectogrammatical layer, Japanese-Czech, natural language processing

Obsah

1. Úvod

Tato práce se zabývá strojovým překladem z japonštiny do češtiny. Hlavním zaměřením je přitom překlad s využitím hloubkového větného rozboru a jeho porovnání s dalšími používanými metodami. Cílem práce je jednak pro danou dvojici jazyků vytvořit základní překladový systém, který by bylo možno v budoucnu dále rozvíjet, a jednak shromáždit dostatečné množství paralelních dat, které budou sloužit k jeho natrénování.

1.1 Motivace

Strojový překlad do češtiny a dalších morfologicky podobně bohatých jazyků je obecně obtížný úkol. V případě anglicko-českého překladu bylo dosaženo dobrých výsledků za pomoci systému, který využívá reprezentace vět na tektogramatické rovině [14]. V současné době sice tento systém, je-li použit samostatně, nedosahuje tak dobrých výsledků jako systémy využívající n-gramové překladové modely, je zde ale stále mnoho prostoru pro zlepšení. V kombinaci s n-gramovým (frázovým) systémem je navíc jeho příspěvek velmi hodnotný [3].

S rozvojem této metody překladu souvisí i snaha vyzkoušet ji i na dalších jazykových párech, proto jsme se rozhodli ji aplikovat pro dvojici japonština-čeština. Ta sice nepatří k nejvýznamnějším z hlediska praktického využití, vezmeme-li ale v potaz dostupnost teorie, dat a nástrojů pro zpracování češtiny, a pak hlavně kontrast s jazykovými rysy japonštiny, může být japonsko-český pár zajímavý pro výzkum strojového překladu.

1.2 Srovnání jazyků

Hlavním úskalím japonsko-českého překladu je výrazná odlišnost těchto dvou jazyků, která je dána jejich příslušností do rozdílných jazykových rodin. Hlavní rozdíly, kterými se japonština od češtiny liší, jsou:

- Struktura japonské věty je podmět-předmět-sloveso.
- Japonština nemá tvar pro vyjádření množného čísla.
- Slovesa časováním nevyjadřují osobu ani číslo, pouze čas, způsob a rod. Navíc jsou tvary přítomného a budoucího času společné, v případě potřeby se rozlišují příslovečným určením.
- Vztahy mezi větnými členy jsou určovány pomocí částic, nikoli pomocí pádů a předložek.
- Vyplývají-li z kontextu, mohou být jednotlivé prvky věty vynechány. K tomu často dochází zejména v praktické mluvě.

Určitě by se daly najít další příklady, výše uvedené charakteristiky japonštiny by ale měly mít na překlad největší vliv.

Japonština není češtině vzdálená pouze po gramatické stránce, což se projevuje také například při sběru paralelních dat. Je obecně známo, že v oblasti strojového překladu bývá často problém zajistit vhodné jazykové nástroje a data. V současné době neexistují téměř žádné dostatečně velké japonsko-české korpusy ani žádné strojově čitelné slovníky. Proto jsme nuceni obstarat potřebná data jinými způsoby.

1.3 Související práce

Strojový překlad je v současné době velmi široký pojem, což je každoročně patrné i z množství konferencí a workshopů, které se mu věnují. Za zmínku stojí například ACL Workshop on Statistical Machine Translation[1], Worskshop on Example-Based Machine Translation[2], či European Machine Translation Conference[3].

Tradičně v rámci strojového překladu obecně rozlišujeme dvě základní paradigmata: statistické překladové systémy a systémy založené pravidlech. Strojový překlad řízený pravidly je závislý na rozsahu dostupných lingvistických znalostí, kdežto statistický překlad naopak potřebuje ručně přeložené paralelní texty, z kterých si posléze extrahuje potřebné informace. Jako zástupce první skupiny můžeme jmenovat například systémy APAČ [9] a RUSLAN [5]. Z druhé skupiny dnes nejvíce vyčnívají systémy využívající frázový překlad [10], [11].

Je samozřejmě možné výše zmíněné přístupy vzájemně kombinovat a vytvářet hybridní překladové systémy. Příkladem takového systému je anglicko-český překladač TectoMT [14]. Jedná se o systém, který bývá označován jako transfer-based, neboť se nejprve provede analýza vstupního textu na požadovanou úroveň abstrakce, poté se analyzovaný text přeloží, a nakonec se na straně cílového jazyka provede syntéza přeložených vět. Data určená k transferu jsou v tomto případě obvykle reprezentována syntaktickými stromy.

Náš systém využívá během překladu stejných principů jako TectoMT, z tohoto důvodu je také implementován do stejného rozhraní. Zvolenou úrovní abstrakce je v případě anglicko-českého překladu tektogramatická rovina, protože právě na této úrovni jsou zachyceny hloubkové sémantické vztahy mezi uzly stromu, kterými jsou v tomto případě pouze plnovýznamová slova. Stejnou úroveň abstrakce volíme i my pro japonštinu-češtinu, což nám nabízí i možnost použít během syntézy stejnou kaskádu nástrojů pro vygenerování českých vět.

1.4 Členění práce

V kapitole 2 je blíže popsán princip hloubkového překladu spolu s výhodami a nevýhodami jeho užití. V kapitole 3 popíšeme veškeré nástroje, které jsme při překladu použili. Kapitola 4 se věnuje rozboru dostupných paralelních dat a našemu výběru z vyjmenovaných možností. Zpracování získaných dat je dále popsáno v kapitole 5. V kapitole 6 podrobněji popisujeme průběh celého překladu. Pozornost je věnována zejména fázi analýzy a transferu. V kapitole 7 jsou čtenáři blíže

[1] `http://www.statmt.org/`
[2] `http://computing.dcu.ie/`
[3] `http://www.eamt.org/`

představeny formémy a jejich role v tektogramatickém překladu. Výsledná evaluace našeho překladače a jeho porovnání s frázovým překladem je prezentována v kapitole 8.

V příloze A je pospán obsah přiloženého CD, v příloze B uvádíme použitý překladový scénář, v příloze C jsou pak stručně popsány knihovny, které byly v rámci této práce implementovány do rozhraní Treex.

2. Tektogramatický překlad

V úvodu jsme uvedli, že můžeme současné strojové překladače obecně rozdělit na dva druhy (statistické a pravidlové). Překladové systémy ovšem můžeme klasifikovat i podle úrovně porozumění danému textu, jak je vidět na obrázku 2.1. Uvedené schéma reprezentuje různé přístupy k překladu. Na spodku pomyslné pyramidy jsou metody, které se vstupním textem pracují jako s posloupností slov bez dalšího rozboru (v tomto případě se jedná o tzv. *přímý překlad*), na vrcholku naopak stojí překlad přes *interlingvu*, která jakožto univerzální jazyk reprezentuje význam věty bez ohledu na to, v jakém jazyce byla původně napsána. Uprostřed se nachází metody, které provádějí překlad ve třech krocích: analýza, transfer a syntéza.

Jako příklad přímého překladu můžeme uvést například frázový překlad, se kterým byl náš systém porovnáván[1] (viz kapitola 8). Náš systém naopak provádí překlad ve výše uvedených třech krocích.

Fáze transferu je různě obtížná podle předem zvolené úrovně analýzy, což mimo jiné schematicky znázorňuje i úsečka na obrázku 2.1. Na druhou stranu, čím větší úroveň abstrakce zvolíme, tím složitější kaskádu nástrojů pro analýzu textu je potřeba použít. Tyto nástroje nám ale mohou do překladu vnést nové chyby. Obecně však platí, že hlubší úroveň analýzy nám dává větší naději zachovat gramatickou správnost a zachytit některé složitější jevy.

Mohlo by se zdát, že je interlingva z hlediska zjednodušení transferu pro překlad nejvýhodnější. Pomineme-li výše uvedenou možnost vzniku chyb během analýzy, nebylo dosud dokázáno, jestli je interlingva v praxi vůbec dosažitelná. V praxi jsme tedy nuceni hledat při volbě vhodné úrovně abstrakce kompromisy.

Systém TectoMT, který je předlohou našemu překladači, se při popisu analyzovaného textu opírá o schéma anotace Pražského závislostního korpusu 2.0 [6] (zkráceně PDT). Anotace použitá v PDT přitom vychází z teorie Funkčního generativního popisu (FGP) vyvíjeného Petrem Sgallem a jeho spolupracovníky od 60. let 20. století [18], [19].

V první sekci této kapitoly jsou popsány jednotlivé roviny abstrakce použité v rámci PDT, v následující sekci pak očekávané výhody a nevýhody překladu v případě, že si jako úroveň transferu zvolíme tektogramatickou rovinu.

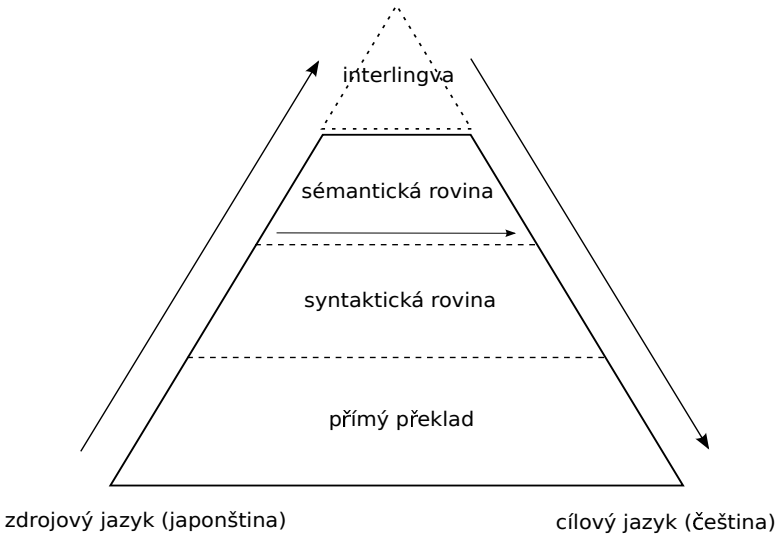
2.1 Roviny jazykové reprezentace Pražského závislostního korpusu

Důležitým aspektem FGP je dělení popisu jazyka na roviny podle úrovně abstrakce. PDT používá k popisu tři úrovně abstrakce: morfologickou rovinu (m-rovinu), analytickou rovinu (a-rovinu) a tektogramatickou rovinu (t-rovinu)[2].

V rámci morfologické roviny je každá věta tokenizována, každému tokenu je pak přiděleno lemma (základní tvar slova) a morfologická značka.

[1] Frázový překlad může být samozřejmě různými rozšířeními povýšen na překlad s transferem, v našem případě ale pro porovnání použijeme jeho základní podobu.

[2] Ve zbytku této práce budou předpony m-, a-, t- používány k rozlišení, ke které úrovni abstrakce dané prvky přísluší.



Obrázek 2.1: Diagram popisující různou hloubku větného rozboru během překladu. Vodorovné úsečky znázorňují zmenšující se obtížnost transferu s rostoucí hloubkovou analýzou.

Na analytické rovině jsou věty převedeny do povrchově-syntaktických závislostních stromů. Každý token ve větě je reprezentován právě jedním a-uzlem. Každému a-uzlu je přidělena analytická funkce podle jeho závislosti na řídícím uzlu. V případě našeho systému užíváme analytických funkcí zatím pouze k rozlišení, které uzly mají a které nemají být přítomny na t-rovině.

Tektogramatická rovina reprezentuje nejvyšší úroveň abstrakce teorie FGP, její struktura je tedy ze všech tří rovin nejsložitější. Každá věta je reprezentována hloubkově-syntaktickým závislostním stromem, kde jsou uzly až na výjimky tvořeny pouze plnovýznamovými slovy. Každému uzlu je přiřazeno tektogramatické lemma a obyčejně také funktor zachycující jeho vztah vůči řídícímu uzlu.

Kromě t-lemmat a funktorů, které tvoří jádro struktury t-stromů, jsou t-uzlům často přiřazeny i další atributy. V našem případě jsou využity zejména formémy (viz kapitola 7) a v malé míře gramatémy. Gramatémy slouží k zachycení vlastností vyjádřených morfologií (jedná se například o čas u sloves, číslo u podstatných jmen, negaci apod.).

Kromě lingvistických informací obsahuje každá rovina také odkazy, které danou rovinu pojí s rovinami „nižšími“.

2.2 Výhody a nevýhody tektogramatického překladu

V následujících dvou oddílech jsou prezentovány očekávané výhody a nevýhody transferu skrze tektogramatickou rovinu.

2.2.1 Výhody

Z našeho pohledu jsou hlavní výhody tektogramatického překladu následující:

- Přestože tektogramatika není zcela jazykově nezávislá, neboť vždy vychází z vlastností daného jazyka, reprezentuje nelexikální atributy poměrně jednotným způsobem. Například vezmeme-li atribut slovesného času u českého slovesa na t-rovině, bude budoucí čas vždy reprezentován stejným způsobem nehledě na to, zdali byl budoucí čas vyjádřen předponou (*pojedu*) nebo pomocným slovesem (*budu jezdit*). Díky tomu máme lepší možnost stejným způsobem reprezentovat větu v případě dvou typologicky různých jazyků.

- Umožňuje nám „zahodit" gramatické informace uzlů, které můžeme odvodit od uzlů řídících. Například česká adjektiva nacházející se v pozici shodného atributu musí mít stejné mluvnické kategorie (rod, čísl, pád) jako podstatná jména, která modifikují. Proto není nutné tuto informaci u přídavných jmen na t-rovině ukládat.

- V případě tektogramatického překladu máme možnost rozložit fázi transferu na lexikální a nelexikální část. V povrchové reprezentaci věty jsou tyto dvě komponenty promíchány, na t-rovině jsou naopak téměř ortogonální. Například lexikální hodnota slovesa (uložena v atributu *t_lemma*) je názorně oddělena od jeho slovesného času (uloženého v atributu *gram/tense*).

- Předpokládáme, že lokální *stromový* kontext t-stromu (ve smyslu potomků a především rodiče daného t-uzlu) nese větší množství informací než lokální *lineární* kontext povrchové reprezentace.

- V praxi se ukázalo, že slovní zarovnání dosahuje mnohem lepších výsledků na linearizovaných t-stromech než na pouhých povrchových reprezentacích vět. Díky tomu jsme například byly schopni z nepříliš velkých paralelních dat automaticky extrahovat dostatečně spolehlivé unigramové překladové slovníky.

Přestože výše zmíněné vlastnosti tektogramatického překladu přinesly příznivé výsledky zejména při překladu z angličtiny, věříme, že japonština, která je češtině ještě vzdálenější, by mohla svými vlastnostmi (role slov ve větě pevně určené pomocí částic) z tektogramatiky také těžit (například při tvorbě překladových modelů).

2.2.2 Nevýhody

Navzdory slibným vlastnostem tektogramatického překladu je třeba poznamenat, že ve srovnání se současnými frázovými překladovými modely má i několik praktických nedostatků:

- Kvůli rozsáhlé struktuře potřebují tektogramatická data mnohem větší paměťovou reprezentaci a komplexnější formáty souborů, což snižuje rychlost zpracování.

- Dále je tu fakt, že v současné době existuje několik různých technik pro lineární data (např. Skryté Markovovy modely), pro stromové struktury nejsou podobné techniky (např. Skryté Markovovy stromové modely) natolik rozšířené, stále se ale pracuje na jejich vývoji a aplikaci [12].

- V případě tektogramatické teorie zůstává stále otevřeno několik otázek. Například není zcela jasné, které další lingvistické informace na t-rovině reprezentovat. V případě japonštiny by se mohlo jednat například o stupně zdvořilosti, kterými je tento jazyk známý. V rámci PDT totiž tato problematika doposud nebyla relevantní.

- V neposlední řadě není tektogramatický překlad příliš oblíben také proto, že k jeho vývoji je nutná alespoň základní znalost tektogramatiky (a ostatních rovin PDT a jejich vzájemné vztahy). V současné době je ale již k dispozici vhodná literatura [1], díky které se potenciální nováčci mohou s danou problematikou snadno seznámit. Lze tedy doufat, že s rostoucí komunitou dojde i k většímu rozvoji tohoto přístupu ke strojovému překladu.

3. Použité nástroje

Při strojovém překladu skrze tektogramatickou rovinu je kromě samotného transferu stromové reprezentace věty důležitá i její důkladná analýza na straně zdrojového jazyka a správná syntéza na straně cíle. Tím pádem se úloha překladu rozpadá na řadu podproblémů, které musíme zvlášť vyřešit. Totéž platí i v případě přípravy paralelních dat. Pro řešení těchto lingvistických podúloh jsme se snažili využít co nejvíce již existujících nástrojů. Jako základ nám posloužilo rozhraní Treex, které většinu potřebných nástrojů již obsahuje. Chybějící nástroje jsme pak pro účely této bakalářské práce do Treexu integrovali pomocí samostatných bloků. Jednalo se zejména o nástroje pro povrchovou analýzu japonských vět.

3.1 Treex

Systém pro zpracování přirozených jazyků Treex [14][1], dříve známý pod názvem TectoMT, vznikl původně za účelem anglicko-českého strojového překladu. V dnešní době je ovšem využíván i při vývoji řešení pro další samostatné úlohy zpracování přirozeného jazyka. Jeho modularita nám umožňuje nejen integrovat různorodé externí nástroje pro zpracování přirozených jazyků, ale i kombinovat statistické a pravidlové metody.

Nejmenší jednotkou kódu Treexu je *blok*. Zpracování dat funguje na principu roury, kdy je kód jednotlivých bloků vykonáván v pořadí, v jakém jsou uvedeny. Sekvenci bloků nazýváme *scénář*. Všechny bloky jsou potomkem třídy `Treex::Block`, nebo jejích potomků. Vnitřní reprezentace dat má během zpracování hierarchickou strukturu. Zpracovávaná data jako celek odpovídají *dokumentu*, ten dále obsahuje jeden, či více *bundle*, z nichž každý odpovídá zpravidla jedné větě. Ty pak obsahují reprezentace věty na jednotlivých úrovních abstrakce. Vzhledem k tomu, že mnohé bloky potřebují často pro správnou funkčnost některé hodnoty dat předem vyplněné (většinou předcházejícími bloky), nelze bloky volat ve zcela libovolném pořadí.

Treex v současné době podporuje několik vstupních a výstupních formátů, přičemž čtení a zápisu každého z nich odpovídá specifický blok. Kromě jednoduchého formátu holých vět podporuje například i formát CoNLLX nebo formát Treex, který má strukturu XML dokumentu a přesně zachycuje vnitřní strukturu zpracovávaných dat.

Scénář japonsko-českého překladu vychází ze vzoru anglicko-českého překladového scénáře používaného v TectoMT (viz Příloha B). Zejména syntéza češtiny je prováděna stejným způsobem.

3.2 Externí nástroje

Vzhledem k tomu, že v době tvorby našeho překladového systému Treex neobsahoval žádné nástroje pro práci s japonštinou, bylo nutné potřebné komponenty do rozhraní přidat. V případě některých úloh souvisejících s analýzou japonských

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

Netokenizovaná věta	彼は本を読まない人だ							
Tokenizace (MeCab)	彼	は	本	を	読ま	ない	人	だ
Tokenizace (bunsetsu)	彼は		本を		読まない		人だ	
Překlad bunsetsu	on		kniha		nečíst		člověk	

Obrázek 3.1: Příklad různé tokenizace věty „On je člověk, který nečte knihy“.

textů byly již k dispozici volně dostupné nástroje třetí strany (POS-tagger, závislostní parser). V těchto případech jsme využili jejich existence a pouze provedli potřebnou integraci do Treexu.

Tokenizaci a značkování slovními druhy (POS tagging) japonské věty provádíme v jednom kroku pomocí morfologického analyzéru MeCab [13]. Tagger využívá sadu tagů IPADIC, obsahující téměř 70 morfosyntaktických kategorií, jež mají hierarchickou strukturu (až čtyři úrovně, jedna hlavní a tři podkategorie). Pro řešení této úlohy v současné době samozřejmě existují i jiné nástroje (např. Chasen[2]), MeCab jsme zvolili díky jeho obecné popularitě, snadné dostupnosti a především kompatibilitě s dále použitým parserem.

Závislostní parsing provádí JDEPP [21][3], přesnost parsování se pohybuje kolem 92%. Nejmenšími jednotkami, se kterými JDEPP pracuje, nejsou tokeny jako je tomu v případě tokenizace MeCabem, ale tzv. *bunsetsu*[4]. Samotný parser nám tedy vygeneruje pouze hrubý závislostní strom a závislosti tokenů v rámci jednotlivých bunsetsu dotváříme až v následujících blocích Treexu. Příklad tokenizace na bunsetsu a tokenizace MeCabem je zobrazen na obrázku 3.1.

Pomocí těchto dvou nástrojů jsme schopni získat povrchově syntaktickou reprezentaci japonské věty, která je dále upravena pro potřeby Treexu. Kromě výše uvedeného doplnění zbývajících závislostí mezi tokeny je v současné době například prováděna i romanizace tagů pro snazší práci.

[2] `http://chasen-legacy.sourceforge.jp/`

[3] `http://www.tkl.iis.u-tokyo.ac.jp/~ynaga/jdepp/`

[4] Problém japonské tokenizace je poměrně složitý a stejně jako například v případě čínštiny do jisté míry nejednoznačný, což vysvětluje mimo jiné i existenci více odlišných tagsetů.

4. Použitá data

V současné době přímá japonsko-česká paralelní data, která by byla v praxi použitelná, téměř neexistují. V databázi paralelních korpusů Opus[1] se sice nachází relativně slibné množství textů (kolem 5,4 milionů tokenů v češtině a zhruba 400 tisíc tokenů v japonštině, japonské věty ovšem nejsou tokenizovány), větné zarovnání těchto dat bylo ovšem ve velké míře provedeno automaticky a po bližším zkoumání jsme se rozhodli tato data prozatím nevyužít. Doména, kterou pokrývají, také není zrovna ideální pro naše účely: v menší míře se jedná o dokumentace PHP a KDE4, velkou část pak tvoří převážně filmové titulky. V budoucnu, po vhodné ruční úpravě, bychom je ale mohli využít.

Z těchto důvodů jsme se rozhodli spolehnout se na jiné paralelní korpusy a vhodný prostřední jazyk. Pro tyto účely se nám jako vhodný kandidát nabízí angličtina. Nejenže existuje mnohem více dostupných japonsko-anglických dat, dalším důvodem je i velké množství anglicko-českých dat nacházejících se v korpusu CzEng.

4.1 CzEng 1.0: Anglicko-česká data

CzEng 1.0 [2][2] je paralelní korpus s bohatou automatickou anotací. Obsahuje 15 milionů paralelních vět (233 milionů anglických a 206 milionů českých tokenů) ze sedmi různých druhů zdrojů. Tyto věty jsou automaticky anotovány na povrchové a hloubkové (a- a t-) rovině syntaktické reprezentace.

V současné době z něj využíváme pouze t-lemmata a jejich zarovnání, které extrahujeme z „exportního formátu“ korpusu (viz tabulka 4.1). V budoucnu stojí za zvážení možnost využití povrchové (a-rovina) analýzy vět a porovnání výsledků.

4.2 Japonsko-anglická data

V případě anglicko-japonských dat existuje více veřejně dostupných zdrojů. Těchto dat je ale výrazně méně než v případě CzEngu. Tato část nám tedy z hlediska přípravy slovníků a překladových modelů v současné době poskytuje největší prostor pro zlepšení.

Jako první jsme se rozhodli použít The Japanese-English Bilingual Corpus of Wikipedia's Kyoto Articles[3]. Jedná se o přesný a především rozsáhlý korpus obsahující zhruba 500 tisíc ručně přeložených vět. Bohužel, vzhledem k tomu, že se jedná o články vztahující se ke Kyotu, a dále pak k tradiční japonské kultuře a historii, není doména tohoto korpusu ideální. Výsledný slovník byl také nakonec mnohem menší, než jsme očekávali (pouze kolem 15 tisíc unikátních japonských hesel).

Proto jsme dále použili korpus Tanaka[4], který je v dnešní době připojen do

[1] http://opus.lingfil.uu.se/
[2] http://ufal.mff.cuni.cz/czeng/
[3] http://alaginrc.nict.go.jp/WikiCorpus/
[4] http://www.edrdg.org/wiki/index.php/Tanaka_Corpus

Sloupec	Příklad	Vysvětlení
4	**zachránit**|PRED|1|0|complex| |v:fin|v|-|neg0|ant|ind|decl|-| |cpl|-|-|disp0|-|it0|-|-|res0|-|-|1| |-|- **#PersPron**|ADDR|2|1| |complex|n:3|n.pron.def.pers|sg| |-|-|-|-|-|-|-|-|nr|-|1|basic|-|-|-|-|-| |- ...	Czech t-layer (tectogrammatical tree): **t-lemma**|functor| |index-in-tree|index-of-governor| |nodetype|formeme|semantic-part-of-speech|... and many detailed t-layer attributes.
8	**#PersPron**|ACT|1|2|complex| |n:subj|n.pron.def.pers|sg|-|-| |-|-|-|-|-|-|-|inan|-|3|-|-|-|-|0|-|- **save**|PRED|2|0|complex|v:fin| |v|-|neg0|ant|ind|decl|-|-|-| |-|disp0|-|it0|-|-|res0|-|-|1|-|- **#PersPron**|APP|3|4|complex| |n:poss|n.pron...	English t-layer (tectogrammatical tree): **t-lemma**|functor| |index-in-tree|index-of-governor| |nodetype|formeme|semantic-part-of-speech|... and many detailed t-layer attributes.
15	**0-1 1-2 2-2 3-3 4-4**	**T-alignment** „there“ for cs2en.
16	**0-0 0-1 2-2 3-3 4-4**	**T-alignment** „back“ for cs2en.

Obrázek 4.1: Příklad exportního formátu CzEngu 1.0. Zobrazeny jsou pouze příslušné sloupce, tučně jsou zvýrazněny informace, které extrahujeme pro naše účely. Kromě slov jako je například “*save*” = “*zachránit*”, se zde nacházejí i speciální t-lemmata „#PersPron“, která odpovídají zájmenům a jako taková nejsou pro náš slovník zajímavá.

Zdroj	Počet vět	Počet JA tokenů	Počet EN tokenů
Wikipedia's Kyoto articles	500 000	˜11 000 000	˜9 900 000
Tanaka Corpus	˜150 000	˜1 700 000	˜1 100 000
JENAAD	150 000		
Aligned Reuters Corpora	˜56 000	˜1 900 000	˜1 300 000

Tabulka 4.1: Přehled známých dat. Počty tokenů byly spočteny na námi tokenizovaných větách. V případě JENAAD korpusu nejsou uvedeny počty tokenů, neboť jsme neměli možnost ho blíže prozkoumat.

projektu Tatoeba[5]. Tento korpus obsahuje 150 tisíc větných párů zejména z učebnic, které se v Japonsku užívají při výuce angličtiny. Vzhledem k tomu, že byl vytvářen převážně studenty, může obsahovat drobné chyby v překladu. Přestože je menší než výše zmíněný korpus článků z Wikipedie, domníváme se, že jeho doména nám výrazně pomohla rozšířit velikost výsledného slovníku.

Dále jsme, zejména díky snadné dostupnosti, využili Alignment of Reuters Corpora[6]. Jedná se sice jen o zhruba 56 tisíc vět, ovšem oblast, ze které pocházejí, nám také do určité míry přispěla při tvorbě slovníku.

Jako další možný zdroj dat bychom mohli ještě zmínit třeba Japanese-English News Article Alignment Data (JENAAD)[7], který stejně jako Alignment of Reuters Corpora obsahuje převážně novinové články. Vzhledem k jeho špatné dostupnosti jej ale v tuto chvíli nepoužíváme. Souhrnný přehled všech nám známých zdrojů je zobrazen v tabulce 4.1.

[5] http://tatoeba.org/eng

[6] http://www2.nict.go.jp/univ-com/multi_trans/member/mutiyama/jea/reuters/index.html

[7] http://www2.nict.go.jp/univ-com/multi_trans/member/mutiyama/jea/index.html

5. Příprava dat

Jak již bylo řečeno, překlad na t-rovině probíhá faktorově. V našem případě dochází pouze k překladu t-lemmat a formémů. Volbu vhodných protějšků zvolených atributů v cílovém jazyce zajišťují pravděpodobnostní unigramové překladové modely. K jejich tréninku používáme japonsko-české slovníky obsahující frekvenci výskytu jednotlivých dvojic unigramů (t-lemmat či formémů). Následující kapitola popisuje extrakci těchto slovníků z nám dostupných paralelních dat.

V současné době jako zdrojová data používáme paralelní korpusy s větným zarovnáním. Japonsko-anglická data jsou zpracována nezávisle na anglicko-českých datech. Při tvorbě japonsko-českých unigramových slovníků, které posléze sloužily k natrénování překladových modelů, jsme se rozhodli vyzkoušet dva postupy:

- Vytvoření dílčích slovníků (japonsko-anglického a anglicko-českého) z příslušných paralelních korpusů a jejich následné spojení skrze shodující se anglická hesla.
- Strojový překlad anglické části japonsko-anglických dat do češtiny a přímá extrakce slovníku z těchto umělých japonsko-českých dat.

Oba postupy si v mnoha ohledech jsou velmi podobné. Při přímé extrakci je nutné nejdříve přeložit anglické věty do češtiny. Toho jsme dosáhli skrze frázový překlad.

V obou případech je pak provedena hloubková analýza vstupních vět. V případě anglicko-českých dat byl tento krok proveden již v CzEngu a my jen přebíráme hotové anotace. Postup analýzy na t-rovinu je pro jednotlivé jazyky popsán v následujících sekcích.

Po analýze následuje výpočet slovního zarovnání pro jednotlivé jazykové páry a extrakce samotných slovníků. Tyto kroky jsou také detailněji popsány dále v této kapitole.

5.1 Zpracování angličtiny

Při analýze anglických vět z japonsko-anglických korpusů byla použita stejná kaskáda nástrojů TectoMT jako při zpracování CzEngu, neboť je použitá pipeline stabilní a od roku 2010 téměř nezměněná. Věty byly tokenizovány pomocí taggeru Morče [20]. Povrchový parsing provedl MST parser [15]. Zbylé kroky zahrnovaly konstrukci t-roviny v závislosti na povrchovém parsingu. Během těchto kroků byla vytvořena i t-lemmata, která byla později použita při slovním zarovnání a samotné stavbě slovníku.

5.2 Zpracování češtiny

Analýza českých vět, které vznikly strojovým překladem anglických vět v našich japonsko-anglických paralelních datech probíhala podobně jako v případě zpracování angličtiny. Opět jsme použili nástroje, které byly použity při zpracování

CzEngu. Tagging ovšem tentokrát provedl tagger Featurama[1], povrchový parsing pak opět MST parser. Konstrukce t-roviny spolu s tvorbou t-lemmat jednotlivých uzlů byla provedena podobným způsobem jako u angličtiny.

5.3 Zpracování japonštiny

Zpracování japonských vět jsme také prováděli v rámci platformy Treex. Tokenizaci a tagování měl na starosti tagger MeCab, závislostní parsing pak JDEPP. Z povrchové reprezentace (a-stromu) pak byla vytvořena hloubková reprezentace vět (t-strom).

Převod do t-roviny byl dosažen prostřednictvím několika bloků s ručně psanými pravidly. Všechny uzly, které nebyly taggerem označeny jako částice, spojky či pomocná slovesa, automaticky považujeme za plnovýznamová slova. Kromě nich jsme na t-rovině ponechali adverbiální částice (副助詞 - *FukuJoshi*), které je potřeba překládat jako příslovce, dále pak japonské spony (např. です - *„desu“*), které jsou taggerem označovány jako *Jodoshi* neboli pomocná slovesa. V jejich případě se sice nejedná o slova nesoucí význam, věříme ale, že jejich přítomnost na t-rovině může přinést lepší výsledky jak při stavbě slovníku, tak při samotném překladu. Dá se očekávat, že v budoucnu ještě dojde k drobným změnám při tvorbě japonské t-roviny, současná podoba nám ale prozatím připadá dostačující.

5.3.1 Japonská tokenizace

Problém japonské tokenizace je stejně jako například v případě čínštiny poměrně složitou úlohou. Jednotlivá slova v japonské větě totiž nejsou oddělena mezerami jako tomu bývá v případě evropských jazyků. Rozdílné tokenizace s sebou navíc přinášejí i rozdílně sady morfologických tagů (viz Kawata [7]).

Při tvorbě našeho překladového systému jsme se mohli setkat s různými způsoby tokenizace (tokenizace MeCabem a tokenizace na bunsetsu). Zde se ale případné odlišnosti daly napravit několika snadnými pravidly (rozvěšení uzlů po hrubém parsingu pouze na bunsetsu).

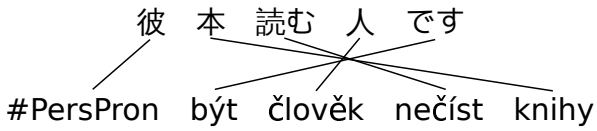
5.4 Zarovnání slov

Pro získání dvojic slov, která by si měla vzájemně v daných jazycích významově odpovídat, jsme použili program GIZA++ [16][2]. Spustili jsme jej na linearizované t-stromy, ve kterých každý uzel odpovídá jednomu plnovýznamovému slovu. V následujících odstavcích, nebude-li uvedeno jinak, budeme místo „uzlů t-stromů“ používat termín „slovo“.

Tvorbou zarovnání na slovech reprezentovaných t-lemmaty se mimo jiné snažíme vyhnout možnému problému řídkosti dat, který bývá často způsoben bohatou morfologií českého jazyka.

GIZA++ je spouštěn dvakrát, jednou v směru zdroj-cíl, podruhé ve směru opačném. Pro větší přesnost pak sloučíme obě zarovnání tím, že provedeme jejich průnik. Příklad zarovnání na t-lemmatech je uveden na obrázku 5.1.

[1] `http://sourceforge.net/projects/featurama/`
[2] `http://code.google.com/p/giza-pp/`



Obrázek 5.1: Příklad slovního zarovnání t-lemmat věty „On je člověk, který nečte knihy“. Z obrázku je vidět, že výskyt spony na t-rovině, může v některých případech přispět nejen k lepšímu překladu věty, ale i ke kvalitnějšímu zarovnání.

Výše popsaný postup provádíme pouze pro japonsko-anglická a námi vytvořená umělá japonsko-česká data. V případě páru angličtina-čeština jsou zarovnání, která jsou ovšem získána stejnými postupy, dostupná v CzEngu[3].

5.5 Stavba slovníku

Jelikož v našich datech nedochází k téměř žádnému překrytí mezi anglicko-českými a japonsko-anglickými větami, provádíme extrakci japonsko-českého slovníku spojením dílčích slovníků.

Přímá extrakce z japonsko-českých dat probíhá stejným způsobem bez nutnosti spojování slovníků.

5.5.1 Od slovního zarovnání k slovníku

S hotovým slovním zarovnáním, jsme schopni provést extrakci slovních párů z linearizovaných t-stromů pomocí jednoduchých skriptů. Takto vzniklé japonsko-anglické a anglicko-české slovníky rovnou obsahují i počty výskytů jednotlivých překladových dvojic.

Dříve, než tyto slovníky spojíme dohromady, jsou vyloučeny nevhodné páry (např. páry s velmi nízkým počtem výskytů, páry obsahující obecná t-lemmata #PersPron apod.). Jelikož japonsko-český slovník má v našem případě mnohem menší velikost, soustředíme se na filtrování nevhodných párů především z anglicko-českého slovníku.

5.5.2 Spojování dílčích slovníků

Spojení slovníků je prováděno na základě shodných anglických hesel (viz tabulka 5.1). Poté jsou opět přepočítány počty výskytů jednotlivých slovních párů jako součet počtů výskytů dvojic, které daný pár vytvořily (anglicko-české straně je přidělena nižší váha). Nakonec jsou zahozeny páry, které se vyskytovaly pouze zřídka (v tabulce 5.2 jsou porovnány jednotlivé slovníky před a po filtraci). Takovýto slovník je poté připraven pro natrénování statického překladového modelu.

Jednou z nevýhod takto vzniklých slovníků je malé pokrytí víceslovných výrazů. Jak totiž bylo zmíněno výše, prováděna je pouze extrakce t-lemmat zarovnaných 1:1. V některých případech ovšem t-lemmata zachycují alespoň nejčastěji se vyskytující složeniny. V případě češtiny se jedná zejména o zvratné zájmeno “*se*”,

[3]Kvalita slovního zarovnání závisí na množství paralelních dat. Zarovnání v CzEngu mají vysokou kvalitu, neboť všech 15 milionů vět bylo zarovnáno najdednou.

ja	en	počet
水	water	1 058
外国	abroad	47
外国	foreigner	362
着る	dress	2
着る	wear	83
通信	communication	65
通信	agency	36

en	cs	počet
courage	odvaha	2 124
foreigner	cizinec	1 713
pace	rázovat	90
reach	dojít	1 705
wear	nosit	34
communication	komunikace	7 512
agency	agentura	42 396

ja	cs	„počet“
外国	cizinec	363,713
着る	nosit	83,034
通信	komunikace	72,512
通信	agentura	78,396

Tabulka 5.1: Příklad japonsko-anglického (tabulka vlevo) a anglicko-českého (uprostřed) dílčího slovníku. Červeně jsou vyznačeny dvojice, které budou přes společné anglické heslo spojeny a umístěny do konečného japonsko-českého slovníku (vpravo). Spodní část tabulky znázorňuje vznik špatného překladového páru. Nesprávný překlad na „agentura“ získal díky vysoké frekvenci výskytu v en-cs datech vyšší skóre než správný překlad na „komunikace“.

	Počet překladových dvojic		Počet japonských hesel	
	Před filtrací	Po filtraci	Před filtrací	Po filtraci
ja-en	397 404	319 712	92 125	79 073
en-cs	2 702 557	2 009 764	-	-
ja-(en)-cs	21 170 050	7 722 742	56 238	31 797
ja-cs	429 117	98 809	91 595	39 077

Tabulka 5.2: Statistika počtu překladových dvojic v jednotlivých slovnících před a po filtraci. U ja-cs a ja-en slovníků jsou uvedeny i počty japonských hesel.

které je nutnou součástí některých sloves (“*smát_se*”), u angličtiny je pro změnu prováděna analýza frázových sloves (např. “*take_off*”, “*settle_down*”). Slova spojená podtržítkem jsou také reprezentována pouze jedním tokenem. V případě japonštiny jsou víceslovné výrazy téměř bez výjimky ignorovány.

5.5.3 Nevýhody prostředního jazyka

Ať už jde o přímou extrakci, nebo spojování dílčích slovníků, v obou případech dochází kvůli spojujícímu jazyku ke vzniku dodatečných chyb.

Vážným problémem při konstrukci je skutečnost, že angličtina obsahuje mnoho slov majících vícero významů (stejný problém by ale přinášel jakýkoli prostřední jazyk). Velmi často se jedná například o slovesa, která tvoří základ frázových sloves (“*go*”→“*go_on*”).

Tato mnohoznačnost způsobuje, že se ve výsledném japonsko-českém slovníku objevují nekorektní páry, které ovšem díky častému souvýskytu v japonsko-anglických či anglicko-českých datech obdržely velký výsledný počet výskytů a jsou tedy při překladu preferovány. Problém jsme do jisté míry vyřešili přidělením menší váhy frekvenční tabulce anglicko-českého slovníku.

Problému by se také dalo vyhnout například přidáním jednoho či více příznaků k anglickým heslům v obou dílčích slovnících. Jako vhodní kandidáti pro tuto roli nám připadají POS tagy. Za zvážení by stálo i použití vhodných nástrojů pro zjednoznačnění významu (Word-Sense Disambiguation, WSD), kterými by se také daly potřebné příznaky získat.

Dalším problémem je ztráta překladů některých japonských hesel. V japonsko-

anglických datech se například mohou vyskytovat překlady pouze na taková anglická hesla, která se v našich anglicko-českých datech vůbec nevyskytují. V těchto případech se potom ve výsledném japonsko-českém slovníku daná japonská hesla neobjeví. Tento problém nastává především u japonských místních jmen a u méně používaných japonských slov.

Při přímé extrakci se mnohoznačnost angličtiny projevovala o něco méně. Bylo to pravděpodobně díky tomu, že při frázovém překladu anglických vět byl brán v potaz alespoň lokální kontext jednotlivých slov. Překlad místních jmen se tentokrát ve výsledném slovníku objevil, ale ne vždy byl správný. Výsledný slovník byl celkově podstatně menší, neboť obsahoval méně špatných slovních párů.

6. Průběh překladu

V následujících odstavcích jsou popsány kroky aplikované v jednotlivých fázích překladu. Ve větším detailu je rozebrána fáze analýzy a transferu, neboť bloky používané v těchto částech jsme nově implementovali do rozhraní Treex. Pro úplnost jsou ovšem stručně popsány i kroky syntézy, které jsou stejné jako v anglicko-českém překladu.

V příloze B je pak uveden plný výpis překladového scénáře se všemi bloky, které se během překladu na vstupní text (a jeho vnitřní reprezentace) aplikují.

6.1 Analýza

Vstupní dokument je zpracováván po jednotlivých řádcích. Předpokládáme přitom, že každá věta je na samostatném řádku. Úkolem analýzy je převést vstupní text z povrchové reprezentace na tektogramatickou rovinu, kde je pak prováděn samotný překlad. Převod na tektogramatickou rovinu by bylo obtížné dělat přímo, nejprve je vhodné provést rozbor na analytické rovině.

6.1.1 Z povrchové reprezentace na a-rovinu

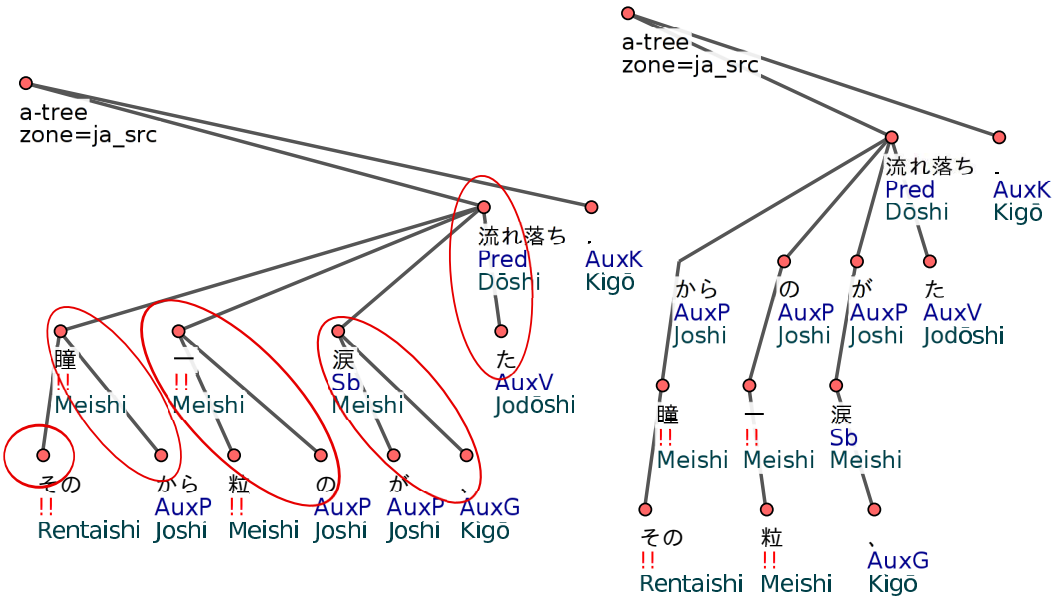
Každá věta je nejprve rozdělena na tokeny, poté je provedeno značkování slovních druhů. Oba kroky má na starost tagger MeCab. Jak už bylo řečeno, používáme sadu tagů IPADIC, která je v oblasti automatického zpracování japonštiny nejrozšířenější. Tagy mají hierarchickou strukturu, obecně se rozlišují ohebné (slovesa, přídavná jména, pomocná slovesa) a neohebné (podstatná jména, příslovce aj.) mluvnické kategorie. Tag se skládá z hlavní kategorie a podle slovního druhu jedné až tří podkategorií, které jej dále specifikují. Během taggingu je provedena i lematizace jednotlivých tokenů. K lematizaci dochází pouze u ohebných slovních druhů, zejména u sloves[1].

Pomocí parseru JDEPP je následně postaven závislostní strom (a-strom). Vzhledem k tomu, že JDEPP pracuje pouze s bunsetsu, jsou zbylé závislosti mezi tokeny dotvořeny následujícím způsobem: na „hlavu“ bunsetsu jsou zavěšeny všechny zbývající tokeny v daném bunsetsu. Za „hlavu“ bunsetsu v tomto případě považujeme plnovýznamové slovo v bunsetsu, které je téměř vždy prvním tokenem zleva (v lineární reprezentaci věty). Další úpravy topologie takto vzniklého stromu jsou podle potřeby provedeny v následujících blocích. Na konci tohoto kroku je provedena romanizace použitých tagů[2].

Aplikací sady heuristik je upravena topologie a-stromu. Vycházíme přitom z konvencí korpusu Verbmobil použitých pro japonský jazyk [8], snažíme se je ovšem aplikovat pro závislostní stromy. Provádíme především přesouvání částic

[1] Je to způsobeno námi zvolenou tokenizací. Kdybychom například použili tokenizaci kde částice nejsou samostatnými tokeny, daly by se za ohebné slovní druhy považovat například i podstatná jména (jejich morfologie by byla dána právě částicemi). Podle IPADIC tagestu jsou částice brány jako samostatné tokeny, které se, dle našeho názoru, svojí funkcí více blíží českým předložkám či spojkám.

[2] Romanizace je prováděna za účelem snadnější práce s tagy v dalších krocích, v budoucnu by ale bylo vhodné zvážit místo romanizace použití vlastních POS značek.



Obrázek 6.1: Porovnání závislostního stromu vygenerovaného JDEPPem (vlevo) a závislostního stromu po všech ostatních úpravách v Treexu (vpravo). Červeně jsou zakroužkovány uzly patřící do stejného bunsetsu.

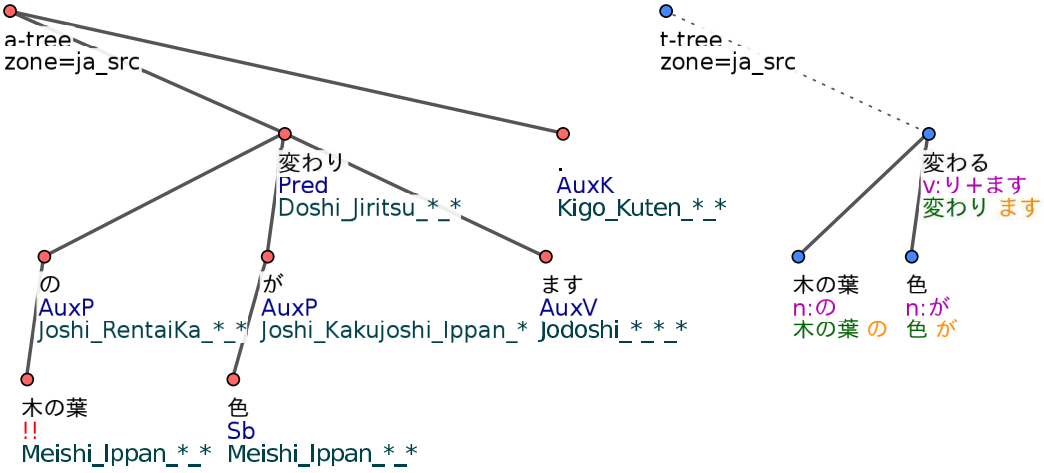
do řídící pozice (slovo, jehož roli ve větě určují, je na nich pak závislé) a stejné přesunutí sponových slov, neboť ty jsou po parsingu závislé na jmenném členu, ale pro překlad potřebujeme reprezentovat opačný vztah. Stejně tak jsou přesunuta nesamostatná slovesa, která bývají po parsingu řídícím členem samostatných (plnovýznamových) sloves. Do budoucna máme v plánu přidat správně přesouvání částic řídících koordinaci a subordinaci ve větě, tyto jevy se ale obecně v závislostních strukturách obtížně reprezentují [17]. Porovnání struktury stromu před a po úpravách topologie je zobrazeno na obrázku 6.1.

Dále jsou nastaveny analytické funkce některých uzlů, nyní pouze za účelem správného převodu na tektogramatickou rovinu. I přesto, že analytické funkce nemají na samotný překlad velký vliv, bylo by vhodné pro úplnost provádět jejich nastavení pro všechny druhy uzlů.

6.1.2 Z a-roviny na t-rovinu

Před samotnou konstrukcí t-stromu jsou označeny uzly pomocných slov, zkráceně pomocné uzly. Jedná se o všechny tokeny, které nereprezentují plnovýznamová slova, tedy částice (vyjma příslovečných částic) a „koncovky“ sloves (ty jsou také segmentovány jako samostatné tokeny a označeny jako pomocná slovesa).

Po těchto úpravách je postaven tektogramatický strom (t-strom). Jeho uzly tvoří pouze plnovýznamová slova. Uzly t-stromu navíc obsahují referenci na svou reprezentaci v rámci a-roviny a některé pomocné uzly označené v předchozím kroku (tj. uzly, které byly při stavbě t-stromu staženy do t-uzlů přes hrany označené blokem `MarkEdgesToCollapse`). Hrany t-stromu jsou odvozeny z hran a-stromu spojujících tyto shluky uzlů. V případě angličtiny nebo češtiny jsou navíc v některých případech upravována t-lemmata, aby lépe zachycovala například frázová slovesa (např. anglické „take_off“). Tento krok ale v případě japonštiny považujeme v tuto chvíli za zbytečný. Příklad reprezentace věty na a- a t- rovině



Obrázek 6.2: Ukázka reprezentace japonské věty na a-rovině a t-rovině. Uzly označené tagem Joshi, Jodoshi a Kigo jsou jakožto pomocné uzly před vytvořením t-stromu označeny k „skrytí“ a na t-rovině nejsou reprezentovány.

lze vidět na obrázku 6.2.

Před samotnou fází transferu jsou ještě všem uzlům t-stromu vyplněny formémy a částečně gramatémy. Funkce a podoba formémů je popsána v kapitole 7. U gramatémů zatím vyplňujeme pouze negaci, ostatní kategorie by ovšem v rámci dalšího vývoje bylo také dobré vyplňovat.

6.2 Transfer

Hlavní úlohou transferové části překladu je tvorba t-stromu cílového jazyka na základě jeho protějšku v jazyce zdrojovém. Topologie zdrojového stromu je zkopírována a následně jsou v cílovém t-stromu vybrány vhodné překlady japonských t-lemmat a formémů.

Výběr je prováděn ve dvou krocích: Nejprve je u každého uzlu vyplněn seznam n nejlepších kandidátů pro překlad. To je provedeno na základě našich statistických překladových modelů. V následujícím kroku jsou pak za pomoci HMTM (Hidden Markov Tree Model) porovnávány jednotlivé kombinace t-lemmat a formémů. U každého uzlu jsou pak vybrány překlady, které byly nejlepší v rámci celé věty (v kombinaci s překlady ostatních uzlů).

Nyní transfer provádíme pouze za pomoci výše zmíněných kroků, ovšem v budoucnu můžeme počítat s přidáním několika pravidlových bloků ošetřujících výjimky či speciální případy. Na mysli máme zejména překlad japonských spon (např. です) na české „být“ (nyní jsou překládány skrze překladový model). Kromě úpravy t-lemmat můžeme uvažovat i modifikaci topologie cílového t-stromu, neboť v některých případech nejsou stromy zdrojového a cílového jazyka zcela izomorfní. V našem případě by se mohlo jednat zejména o generování uzlů, které ve zdrojové větě nejsou vyjádřeny (vyplývají z kontextu). Je ale možné, že tyto úpravy bude potřeba provádět už během analýzy.

6.3 Syntéza

V závěru celého překladu je vygenerována česká věta na základě českého t-stromu vytvořeného během překladu. Je vytvořen a-strom a následně je vyplněna povrchová morfologie (rod, číslo, pád, atd.) s pomocí vyplněných formémů, případně gramatémů. Dále jsou vytvořeny a-uzly odpovídající pomocným slovesům, spojkám, předložkám atd. Kromě jiného dochází k vytvoření výsledných tvarů slov za pomoci generátoru slovních tvarů [4]. Podrobnější popis syntézy českých vět je k dispozici v dokumentaci TectoMT [22].

7. Formémy

Po vzoru TectoMT používá náš systém formémy, jež byly zavedeny za účelem indikace morfosyntaktických vlastností a vztahů slov reprezentovaných na tektogramatické rovině a přenesení těchto vztahů během překladu. Motivací je stejný cílový jazyk našeho překladače (čeština), pro který se v minulosti zavedení formémů ukázalo z pohledu syntézy jako přínosné. Navíc kromě uspokojivé reprezentace výše zmíněných větných vztahů nám práce s formémy umožňuje velmi snadno s pomocí několika jednoduchých pravidel vytvořit základní překladový systém, schopný v cílovém jazyce vytvářet přinejmenším jednoduchou morfologii překládaných slov.

7.1 Japonské formémy

Vzhledem k tomu, že je množina použitých formému závislá na příslušném jazyce, bylo potřeba sadu japonských formémů vybudovat od základu tak, aby nám požadované morfosyntaktické vlastnosti japonštiny zachytila. Kvůli výrazné odlišnosti japonštiny jsme se některým drobným změnám nevyhnuli. Až na výjimky jsme se ale snažili zachovat následující vlastnosti:

- hodnoty formémů by měly být strojově snadno čitelné,
- měly by také být snadno srozumitelné člověku: v tuto chvíli jsou součástí japonských formémů i japonské znaky, k jejich čtení je tedy potřeba alespoň jejich základní znalost,
- různé množiny formémů jsou použitelné pro t-uzly s různým sémantickým slovním druhem, z hodnoty formému by tedy mělo být přímo čitelné, ke kterému slovnímu druhu patří.

Protože v současné době japonské formémy používáme pouze během analýzy a překladu nebyl kladen velký důraz na zachování vlastností, které by pomohly při syntéze japonských vět.

Přiřazování hodnot formémů je v podstatě určeno POS tagy příslušných plnovýznamových slov a hodnotami k nim náležících pomocných a-uzlů. Způsob přidělování přitom můžeme rozdělit na dvě skupiny podle toho, zdali se jedná o podstatná jména (名詞 - Meishi) a nominální adjektiva (tzv. な-*adjektiva*, neboli 形容動詞 - *Keiyōdōshi*), nebo o slovesa (動詞 - Dōshi) a slovesná adjektiva (tzv. い-*adjektiva*, neboli 形容詞 - *Keiyōshi*).

V tuto chvíli nerozlišujeme podstatná jména od nominálních adjektiv, pro naše potřeby obojí klasifikujeme jako sémantická substantiva. Hodnota formémů podstatných jmen je určena částicemi, které k daným t-uzlům náleží. V případě, že k t-uzlu náleží více částic, jsou uvedeny hodnoty všech. S nominálními adjektivy nakládáme jako s neshodnými přívlastky, hodnota jejich formémů je *n:attr*. Podstatná jména a nominální adjektiva mohou být samozřejmě i součástí sponových sloves, v takovém případě nám ale napomáhá fakt, že sponové slovo です je na t-rovině také reprezentováno. Díky tomu můžeme funkci predikátu nechat

sponě, která je pro účely přidělování formémů považována za sloveso, a jmenné části přiřadíme formém normálním způsobem. Uveďme si příklady některých substantivních formémů:

- n:は — téma (nebo podmět) věty (indikované částicí は - *„wa“*)
- n:の — modifikátor jiného větného členu (vyjádřen částicí の - *„no“*); má podobnou funkci jako český přívlastek
- n:を — předmět (indikovaný částicí を - *„wo“*)

V případě sloves a い-adjektiv přiřazujeme hodnoty formémů jiným způsobem. Jelikož se jedná o slovní druhy s vlastním skloňováním, dochází ke změně tvaru kořenového slova (v případě pravidelných sloves pouze ke změně poslední slabiky) a přidání vhodného suffixu. Jako hodnotu formému tedy bereme podřetězec, ve kterém se slovní forma liší od svého lemmatu. Stačilo by sice značit pouze hodnotu poslední slabiky, chceme ale rovněž pokrýt nepravidelná slovesa くる - *„kuru“* (jít, přicházet) a する - *„suru“* (dělat)[1], kde v některých případech dochází k změně celého tvaru slovesa. Zde je pár příkladů formémů sloves:

- v:り+ます — sloveso v tzv. *zdvořilostní* (ます - *„masu“*) formě
- v:い+て_くださる — sloveso v tzv. て *(„te“)* formě s pomocným slovesem くださる *(„kudasaru“)*, které vyjadřuje formální požadavek
- v:し+た — sloveso v prosté formě v minulém čase (znázorněném koncovkou た - *„ta“*)

Slovesná adjektiva jsou v této skupině zahrnuta proto, že mají stejně jako slovesa vlastní skloňování. To sice není tak bohaté jako v případě sloves, ale pro účely přiřazování formémů s nimi můžeme nakládat podobným způsobem. Příklady formémů slovesných adjektiv:

- adj: — implicitní hodnota formému přiřazovaná *i*-adjektivům
- adj:く+て — *i*-adjektivum v て *(„te“)* formě
- adj:く — *i*-adjektivum v prostém (slovníkovém) tvaru

Formémy přiřazujeme i příslovcím a příslovečným částicím, jež z hlediska sémantických slovních druhů nerozlišujeme. Nyní jim je přiřazována pouze jediná hodnota formému: *adv:*.

[1]Tato slovesa mají v japonštině mnoho dalších významů v závislosti na slovech, která se k nim váží (např. 勉強する - *„studovat“*, 心配する - *„znepokojovat_se)“*.

F_{ja}	F_{cs}	$P(F_{cs}\|F_{ja})$
adj:	adj:1	0.1612
adj:	adv	0.1149
n:は	n:1	0.4369
n:は	n:X	0.1815
n:を	n:4	0.2178
n:を	n:1	0.1225
n:を	n:X	0.1392
n:が	n:1	0.3043
n:が	n:X	0.1907
n:が	adj:attr	0.1018
n:が	n:4	0.0857
v:り+なさる	v:inf	0.3148
v:り+なさる	v:fin	0.2778
v:り+なさる	adv	0.2407
n:に_と_の	v:že+fin	0.2608
n:に_と_の	v:fin	0.2173
n:に_と_の	n:s+7	0.1739
v:て_いる_ます	v:fin	0.4754
v:て_いる_ます	adj:1	0.1475
v:て_いる_ます	adv	0.1229

Tabulka 7.1: Ukázka japonsko-českého pravděpodobnostního překladového slovníku formémů. Pro vybrané japonské formémy je zobrazeno několik nejvíce pravděpodobných českých protějšků spolu s podmíněnou pravděpodobností českého formému za předpokladu japonského.

7.2 Překlad formémů

Vzhledem k tomu, že současná sada formémů byla vytvořena intuitivně a s omezenou znalostí japonštiny, nepoužíváme pro jejich překlad žádná ručně psaná pravidla a vycházíme pouze z našich trénovacích dat. Extrakci slovníku formémů přitom provádíme téměř stejným způsobem jako extrakci slovníku t-lemmat. Díky tomu, že formém je stejně jako t-lemma atributem uzlů t-stromů, se postup liší pouze v extrakci jiné hodnoty při linearizaci t-stromů.

V tabulce 7.1 je uveden fragment extrahovaného slovníku. Jde vidět, že překlad formémů podstatných jmen a adjektiv alespoň v některých případech probíhá podle našich představ, v případě sloves jsou výsledky výrazně horší.

7.3 Budoucí práce

Při zkoumání slovníků a překládaných vět jsme se přesvědčili, že je potřeba současnou sadu formémů ještě dále vylepšovat. Ze zkoumaného vzorku dat jsme ochotni tvrdit, že například formémy podstatných jmen (tedy formémy odvozované od částic) jsou v tuto chvíli vyhovující. V případech, kdy překlad substantivních formémů neprobíhal, tak jak bychom to očekávali, lze příčiny neúspěchu

hledat na analytické rovině, neboť kupříkladu stále neošetřujeme částice zajišťující koordinaci ve větě.

Naopak v případě slovesných formémů je potřeba v budoucnu zvolit zcela odlišný přístup. Nejenže jsou součástí slovesných formémů informace, které by měly být ukládány ve zcela odlišných atributech (slovesný čas vyjádřený „koncovkami“ sloves by měl být uložen v gramatémech), ale není ani jisté, zdali například změna kmenového tvaru pomáhá určovat morfosyntaktické vztahy vůči ostatním větným členům. V budoucnu bychom mohli také zkusit na slovesa aplikovat některé formémy používané v angličtině či češtině.

8. Experimenty a měření

V této kapitole se budeme věnovat vyhodnocování kvality našeho překladového systému. V první sekci popíšeme sadu testovacích dat, jež jsme během našeho měření použili, a způsob, jakým byla zkonstruována. Dále popíšeme základní frázový systém, který jsme použili pro srovnání s naším překladačem. V sekci poté jsou prezentovány výsledky našich měření a v závěru této kapitoly provedeme jejich interpretaci.

8.1 Testovací data

Pro účely měření kvality překladu jsme náhodně vybrali 1000 dvojic vět z našich japonsko-anglických paralelních dat, přesněji z korpusu Tanaka a Reuters. Anglické věty jsme strojově přeložili do češtiny (stejným způsobem jako při tvorbě japonsko-českých paralelních dat) a výsledek jsme posléze ještě ručně opravili. Jednalo se zejména o opravu gramatických chyb, které při překladu vznikly, pouze v případě velkých odchylek od japonských protějšků jsme věty celé ručně přepsali. Do testovacích dat jsme nezahrnuli věty z korpusu Kyoto's Wikipedia articles, neboť obsahoval mnoho souvětí se složitou strukturou, důkladná korektura překladu anglických vět by proto byla příliš časově náročná.

Japonské věty byly kvůli frázovému systému tokenizovány MeCabem. Náš překladač pak při samotném překladu tento krok jednoduše přeskočil.

8.2 Frázový překladový systém

Pro porovnání s naším překladovým systémem jsme si vybrali frázový systém Moses [11][1]. Nejenže jakožto zástupce přímého překladu reprezentuje v rámci přístupu ke strojovému překladu zcela odlišné paradigma, konstrukce jednoduchého n-gramového překladače je také velmi snadná.

8.2.1 Použitá data

Vzhledem k tomu, že naše japonsko-anglická a anglicko-česká data mají téměř prázdný průnik přes anglické věty, byla konstrukce trénovacích dat pro frázový překlad spojováním přes prostřední jazyk vyloučena. Místo toho jsme se rozhodli použít náš uměle vytvořený japonsko-český korpus.

Jedná se o stejná data, která jsme použili pro extrakci slovníků našeho hloubkového systému. Z těchto trénovacích dat jsme dále náhodně vyjmuli kolem 2500 větných dvojic, které nám posloužily k vyladění frázového překladového modelu. Tokenizace těchto dat byla provedena stejným způsobem jako u testovací sady vět.

[1] http://www.statmt.org/moses/

8.2.2 Příprava

Nejprve jsme provedli slovní zarovnání na našich umělých japonsko-českých datech. Na rozdíl od extrakce slovníků ale bylo toto zarovnání provedeno pouze na tokenizovaných povrchových reprezentacích vět. Na základě těchto zarovnání jsme vytvořili statistický překladový model. Vedle něj jsme natrénovali i jazykový model cílového jazyka. I když jsme měli k dispozici čistá česká data, zvolili jsme pro trénink jazykového modelu české věty z našich umělých dat. Důvodem byl fakt, že jazykový model vytvořený z čistých českých dat dostal během ladění mnohem menší váhu než jazykový model z umělých dat. Bylo to zřejmě způsobeno charakterem našeho n-gramového překladového modelu a dat určených k ladění (také obsahovala umělé české věty). Kombinací těchto modelů jsme pak získali základní model, který Moses později použil pro překlad testovacích vět. Tento model byl dále s použitím dat určených k ladění upraven pomocí metody MERT. Takto vyladěný model byl pak připraven k testování.

Frázový překladový systém jsme tímto způsobem natrénovali dvakrát, jednou na slovních formách, podruhé na lemmatech (tj. překlad do hrubší podoby češtiny)[2].

8.3 Výsledky měření

Výše uvedené systémy jsme spustili na stejném vzorku testovacích dat. Měření překladu jsme poté provedli jak za pomoci automatických metrik, tak i skrze ruční evaluaci. Oba systémy měly téměř stejnou míru OOV (out-of-vocabulary), kolem 3%. Za nepřeložená slova jsme přitom považovali všechny řetězce ve výstupu obsahující japonské znaky.

8.3.1 Automatická evaluace

Automatickou evaluaci jsme prováděli klasicky pomocí metriky BLEU, dále jsme měřili metriky PER, TER a CDER[3]. Pro účely zobrazení výsledků měření těchto metrik jsme přitom u metrik TER a CDER použili místo míry chybovosti (*error-rate*) míru přesnosti (*accuracy*). Ta se v případě metriky TER dá spočítat následujícím způsobem:

$$TAcc = 1 - TER$$

kde TER značí míru chybovosti TER. Přesnost v případě metriky CDER spočteme analogicky. Z charakteru rovnice vyplývá, že čím vyšší naměříme přesnost (a tím pádem menší chybovost), tím kvalitnější je evaluovaný překladový systém.

U metriky PER jsme přesnost překladu počítali následujícím způsobem:

$$PAcc = (C - max(0, T - R))/R$$

kde C značí počet správně přeložených tokenů, T je délka přeložené věty a R je délka referenční věty. Opět platí, že vyšší PER skóre poukazuje na kvalitnější překlad. Kvalitu překladu jsme měřili na slovních formách a na a-lemmatech.

[2]Lematický výstup je nepoužitelný pro koncového uživatele ale je vhodný pro posouzení, zda překladač zachovává slova bez ohledu na morfologii.

[3]Tyto „metriky“ nesplňují vlastnosti metrik v matematickém smyslu, ale tradičně se jim tak říká.

Použité metriky	Uvádíme jako	Treex	Moses
BLEU	BLEU	0,00±0,00	6,55±0,95
PER	PAcc	23,52±1,32	25,02±2,54
TER	TAcc	7,78±1,04	6,85±2,15
CDER	CDAcc	13,81±0,65	19,48±1,17

Tabulka 8.1: Tabulka výsledků měření jednotlivých automatických metrik našich dvou porovnávaných systémů. Překlad byl proveden na předem tokenizovaných větách. U metrik PER, TER a CDER je místo míry chybovosti (error-rate) uvedena přesnost (accuracy).

Použité metriky	Uvádíme jako	Treex	Moses
BLEU	BLEU	0,00±0,00	15,92±1,45
PER	PAcc	39,11±1,64	49,25±2,23
TER	TAcc	14,78±1,13	29,46±2,04
CDER	CDAcc	21,38±0,71	38,47±1,19

Tabulka 8.2: Tabulka výsledků měření jednotlivých automatických metrik našich dvou porovnávaných systémů. V tomto případě byl překlad proveden mezi a-lemmaty. U metrik PER, TER a CDER je opět uvedena přesnost překladu.

V tabulce 8.1 jsou uvedeny výsledky měření překladu na slovních formách. Bohužel, BLEU skóre našeho překladače bylo nulové. To bylo zřejmě způsobeno tím, že se v přeloženém textu nepodařilo najít ani jeden 4-gram, který by referenční překlad potvrdil. Frázový systém si v tomto ohledu vedl podstatně lépe. Lépe dopadl i v případě metrik PER a CDER, zde byl ovšem rozdíl poměrně malý. Náš systém naopak překvapivě dosáhl lepšího výsledku při měření metrikou TER.

V tabulce 8.2 jsou uvedeny hodnoty, které jsme naměřili při překladu na lemmatech. I když jsme v tomto případě očekávali zlepšení BLEU skóre našeho systému, výsledek byl opět nulový. Ani zde se tedy nepodařilo najít jediný 4-gram potvrzený referencí. Na druhou stranu se zlepšení BLEU skóre potvrdilo u frázového systému. Očekávané lepší výsledky našeho systému při překladu lemmat nám potvrdily teprve metriky PER, TER a CDER. Hloubkový překlad ale nakonec ve srovnání s frázovým systémem prohrál.

Vzhledem k tomu, že výsledky měření BLEU našeho hloubkového systému byly velmi špatné, provedli jsme navíc průzkum přesnosti samostatných n-gramů. Výsledky jsou uvedeny v tabulce 8.3. Můžeme si všimnout, že při překladu na slovních formách se kromě 4-gramů nepodařilo v přeloženém textu najít ani jediný 3-gram, který by referenční překlad potvrdil. V textu, který vznikl překladem a-lemmat pak bylo několik 3-gramů nalezeno, jejich množství je ale zanedbatelné. Je vidět, že překlad se v obou případech dařil alespoň na unigramech. Příčinou byla zřejmě skutečnost, že náš systém nebyl schopen na a-rovině generovat chybějící pomocné uzly.

8.3.2 Ruční evaluace

Ruční ohodnocení jsme prováděli na vzorku 100 vět vybraných z našich testovacích dat. Každý pár byl náhodně zamíchán, aby anotátor nevěděl, která věta byla

Druh překladu	1-gram	2-gram	3-gram	4-gram
Treex (formy)	24,4	0,5	0,0	0,0
Treex (lemmata)	40,5	2,3	0,2	0,0

Tabulka 8.3: Tabulka uvádějící přesnosti jednotlivých n-gramů (tj. jaký podíl ze všech n-gramů v hypotéze byl potvrzen referencí).

	*	**	eq+	eq−
Treex	22	2	10	34
Moses	28	6	10	34

Tabulka 8.4: Tabulka výsledků ručního ohodnocení překladu vybraného vzorku přeložených vět. Je zde uvedeno, kolikrát byl překlad věty jednoho systému lepší než překlad druhého (*), kolikrát byl výrazně lepší (**), kolikrát byly oba zhruba stejně dobré (eq+) a kolikrát byl překlad v obou případech stejně špatný (eq−).

vygenerována kterým systémem. Hodnocení překladu bylo vytvářeno zejména na základě porovnání s naším referenčním překladem, nikoli vstupní věty.

Vzhledem k značným nedostatkům obou systémů, jsme byli během hodnocení velmi shovívaví. Byli jsme tolerantní vůči špatnému skloňování, dále jsme tolerovali i nesprávný slovosled. Používali jsme dva stupně hodnocení: pokud byl jeden překlad lepší než druhý, obdržel bod; byl-li jeden z překladů výrazně lepší (až na velmi drobné chyby odpovídal referenci), obdržel dva body. Dále jsme rozlišovali, jestli byly překlady v případě podobné kvality stejně dobré nebo stejně špatně. Výsledky ruční evaluace jsou uvedeny v tabulce 8.4.

Frázový překlad si opět vedl o něco lépe než překlad s hloubkovým rozborem. Rozdíl byl ale tentokrát relativně malý. Dále je vidět, že oba systémy jsou v současné době stále velmi špatné (1/3 překladů byla špatná v obou případech), lze tedy usoudit, že se současnými předními překladači, si náš systém stojí mnohem hůř.

8.4 Shrnutí

Z výše uvedených výsledků našich měření je jednoznačně vidět, že si náš hloubkový překladový systém v případě jazykového páru japonština-čeština vedl hůř než referenční frázový překlad. Přitom je potřeba podotknout, že ani náš frázový překlad zdaleka nedosahoval úrovně současných překladačů. Z ruční evaluace potom vyplývá, že kvalitativní propast mezi našimi dvěma prezentovanými systémy nebyla tak velká, jak ukazovala automatická evaluace.

V následujících sekcích vyjmenujeme nejpodstatnější slabiny obou systémů.

8.4.1 Nedostatky hloubkového překladu

Během ruční kontroly přeložených vět z testovací sady jsme si všimli těchto zásadních nedostatků:

- Náš systém v současné době velmi výrazně selhává během generování slovních forem ve fázi syntézy. To je v první řadě způsobeno nedostatkem vyplněných atributů t-roviny, zejména gramatémů.

- Z předchozího bodu tedy jasně vyplývá, že i když jsou formémy schopny přispět ke kvalitě našeho překladového systému, nejsou sami o sobě dostačující. To je ovšem pochopitelné, protože jejich úkolem je pouze zachycovat morfosyntaktické vztahy ve větě.

- Kromě výše uvedených chyb při generování slovních forem náš systém selhává i při vytváření pomocných uzlů (předložek, spojek atd.) na analytické rovině. Příčina je podobná jako v případě špatné morfologie (nedostatek informací na t-rovině, nevyhovující sada formémů).

- Výrazný přínos zlepšení BLEU skóre by určitě přinesla oprava slovosledu cílových vět. Japonština má totiž například vždy přísudek na konci věty, což ale v případě češtiny už neplatí.

8.4.2 Nedostatky frázového překladu

I když si frázový překlad vedl v pokusu lépe než náš hloubkový překladač, všimli jsme si během ruční kontroly několika slabin, na které by bylo vhodné se v budoucnu zaměřit.

Zdaleka největším problémem našeho frázového překladu byl nedostatek vhodných japonsko-českých paralelních dat. Problém jsme se snažili do jisté míry vyřešit našimi uměle vytvořenými daty, ty ale kvůli způsobu jejich přípravy obsahovaly mnoho podstatných chyb. Dalo by se říci, že s těmito uměle vytvořenými daty dokázal naopak lépe pracovat náš hloubkový systém, protože z nich na rozdíl od frází extrahoval pouze zarovnaná t-lemmata. Tato přednost se ale bohužel během měření nedokázala projevit kvůli výše uvedeným chybám při generování slovních forem a pomocných uzlů na analytické rovině.

9. Závěr

Tato práce popsala naši úvodní verzi japonsko-českého překladače založeného na principu hloubkového překladu. V rámci toho byl tento systém implementován do rozhraní Treex, neboť mnoho postupů přebíral z překladového systému TectoMT, který v minulosti ukázal slibné výsledky.

Naše verze překladače naopak v tuto chvíli při porovnání s frázovým překladem, který je z hlediska strojového překladu nejrozšířenější, neobstála. Jsme si ale vědomi největších nedostatků našeho systému a jeho možného budoucího vylepšení.

Důležitou součástí této práce bylo také získání dostatečného množství japonsko-českých paralelních dat. I přes nedostatek přímých dat jsme byli schopni vytvořit vyhovující překladové modely pro náš hloubkový překlad.

9.1 Budoucí práce

Zřejmě největší slabinou je v současné době nedostatek vyplňovaných atributů na japonské tektogramatické rovině. Dále by bylo potřeba provést revizi japonských formémů; jak totiž bylo uvedeno, v případě sloves je současná sada nevyhovující. Důležitá je i celková revize japonského parsování a přechodu z analytické roviny do tektogramatické. V neposlední řadě by také bylo vhodné (pravděpodobně během transferu) opravovat slovosled cílových vět.

Po dokončení výše uvedených zlepšení by bylo zajisté zajímavé vyzkoušet zkombinovat náš hloubkový překladový systém se systémem frázovým po vzoru překladového systému Chiméra [3].

Literatura

[1] Bojar, O. *Čeština a strojový překlad*. Charles University in Prague, 2012. ISBN 978-80-904571-4-0.

[2] Bojar, O. et al. The Joy of Parallelism with CzEng 1.0. In *Proceedings of the Eighth International Language Resources and Evaluation Conference (LREC'12)*, s. 3921–3928, Istanbul, Turkey, Květen 2012. ELRA, European Language Resources Association. ISBN 978-2-9517408-7-7.

[3] Bojar, O. a Rosa, R. a Tamchyna, A. Chimera – Three Heads for English-to-Czech Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, s. 92–98, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. Dostupné z: <http://www.aclweb.org/anthology/W13-2208>.

[4] Hajič, J. *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Karolinum, Charles University Press, Prague, Czech Republic, 2004.

[5] Hajič, J. RUSLAN: an MT system between closely related languages. In *Proceedings of the third conference on European chapter of the Association for Computational Linguistics*, s. 113–117. Association for Computational Linguistics, 1987.

[6] Hajič, J. et al. Prague Czech-English Dependency Treebank 2.0, 2012.

[7] Kawata, Y. *Tagsets for Morphosyntactic Corpus Annotation: The Idea of a 'reference Tagset' for Japanese*. University of Essex, 2005. Dostupné z: <http://books.google.cz/books?id=s_tyHQAACAAJ>.

[8] Kawata, Y. a Bartels, J. Stylebook for the Japanese Treebank in VERBMOBIL. Technical report, 2000.

[9] Kirschner, Z. a Rosen, A. APAC - An experiment in machine translation. *Machine Translation*. 1989, 4, 3, s. 177–193.

[10] Koehn, P. a Och, F. J. a Marcu, D. Statistical phrase based translation. In *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT/NAACL)*, 2003.

[11] Koehn, P. et al. Moses: Open Source Toolkit for Statistical Machine Translation. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, s. 177–180, Prague, Czech Republic, June 2007. Association for Computational Linguistics. Dostupné z: <http://www.aclweb.org/anthology/P/P07/P07-2045>.

[12] Kondo, S. a Duh, K. a Matsumoto, Y. Hidden Markov Tree Model for Word Alignment. In *Proceedings of the Eighth Workshop*

on Statistical Machine Translation, s. 503–511, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. Dostupné z: <http://www.aclweb.org/anthology/W13-2263>.

[13] Kudo, T. MeCab: Yet another part-of-speech and morphological analyzer. http://mecab.sourceforge.net/, 2005.

[14] Mareček, D. a Popel, M. a Žabokrtský, Z. Maximum Entropy Translation Model in Dependency-Based MT Framework. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, s. 207–212, Uppsala, Sweden, July 2010. Association for Computational Linguistics. ISBN 978-1-932432-71-8.

[15] McDonald, R. et al. Non-Projective Dependency Parsing using Spanning Tree Algorithms. In *Proceedings of HLT/EMNLP 2005*, October 2005.

[16] Och, F. J. a Ney, H. A Comparison of Alignment Models for Statistical Machine Translation. In *Proceedings of the 17th conference on Computational linguistics*, s. 1086–1090. Association for Computational Linguistics, 2000. ISBN 1-555-55555-1.

[17] Popel, M. et al. Coordination Structures in Dependency Treebanks. In *ACL (1)*, s. 517–527. The Association for Computer Linguistics, 2013. ISBN 978-1-937284-50-3.

[18] Sgall, P. *Generativní popis jazyka a česká deklinace*. Prague: Academia, 1967.

[19] Sgall, P. a Hajičová, E. a Panevová, J. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Dordrecht: Reidel Publishing Company and Prague: Academia, 1986.

[20] Spoustová, D. et al. The Best of Two Worlds: Cooperation of Statistical and Rule-Based Taggers for Czech. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing, ACL 2007*, s. 67–74, Praha, 2007.

[21] Yoshinaga, N. a Kitsuregawa, M. Kernel slicing: scalable online training with conjunctive features. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, s. 1245–1253, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. Dostupné z: <http://dl.acm.org/citation.cfm?id=1873781.1873921>.

[22] Žabokrtský, Z. *From Treebanking to Machine Translation*. habilitation, Faculty of Mathematics and Physics, Charles University in Prague, Malostranské náměstí 25, Praha 1, 2010.

Seznam tabulek

A. Obsah přiloženého CD

Přiložené CD obsahuje následující položky:

- sources - zdrojový kód námi implementovaných japonských bloků do rozhraní Treex
- data - použitá paralelní data
  1. raw - nezpracované paralelní korpusy
  2. moses - trénovací, testovací a ladící korpusy používané frázovým systémem
  3. treex - jednotlivé slovníky a výsledné překladové modely používané při hloubkovém překladu
- install
  1. README - návod pro ruční checkout Treexu
  2. Makefile - Makefile pro správné umístění odkazů na překladové modely do struktury Treexu
- scenarios - překladový scénář našeho hloubkového překladu
- PDF obsahující tuto práci

B. Scénář japonsko-českého překladu

V této příloze uvádíme překladový scénář používaný naším systémem. Jednotlivé fáze jsou označeny komentáři, bloky pracující s různými rovinami reprezentace jsou vzájemně viditelně odděleny.

```
# read input sentences
Util::SetGlobal language=ja selector=src
Read::Sentences

# analysis
W2A::JA::TagMeCab
W2A::JA::ParseJDEPP

W2A::JA::RomanizeTags
W2A::JA::FixInterpunction
W2A::JA::RehangAuxVerbs
W2A::JA::RehangCopulas
W2A::JA::FixCopulas
W2A::JA::RehangConjunctions
W2A::JA::RehangParticles
W2A::JA::SetAfunParticles
W2A::JA::SetAfun

A2T::MarkEdgesToCollapse
A2T::BuildTtree
A2T::JA::SetFormeme
A2T::JA::SetGrammatemes

# transfer
Util::SetGlobal language=cs selector=tst

T2T::CopyTtree source_language=ja source_selector=src
T2T::JA2CS::TrFAddVariants
T2T::JA2CS::TrLAddVariants
T2T::EN2CS::CutVariants lemma_prob_sum=0.5\
  formeme_prob_sum=0.9 max_lemma\
  variants=7 max_formeme_variants=7
T2T::EN2CS::TrLFTreeViterbi

# syntesis
Util::SetGlobal language=cs selector=tst

T2A::CopyTtree
T2A::CS::DistinguishHomonymousMlemmas
```

```
T2A::CS::ReverseNumberNounDependency
T2A::CS::InitMorphcat
T2A::CS::FixPossessiveAdjs
Util::DefinedAttr tnode=t_lemma,formeme,functor,clause\
  number anode=lemma\
  message="after InitMorphcat and FixPossessiveAdjs"
T2A::CS::MarkSubject
T2A::CS::ImposePronZAgr
T2A::CS::ImposeRelPronAgr
T2A::CS::ImposeSubjpredAgr
T2A::CS::ImposeAttrAgr
T2A::CS::ImposeComplAgr
T2A::CS::DropSubjPersProns
T2A::CS::AddPrepos
T2A::CS::AddSubconjs
T2A::CS::AddReflexParticles
T2A::CS::AddAuxVerbCompoundPassive
T2A::CS::AddAuxVerbModal
T2A::CS::AddAuxVerbCompoundFuture
T2A::CS::AddAuxVerbConditional
T2A::CS::AddAuxVerbCompoundPast
T2A::CS::AddClausalExpletivePronouns
T2A::CS::MoveQuotes
T2A::CS::ResolveVerbs
Util::DefinedAttr anode=clause_number\
  message="after ProjectClauseNumber"
T2A::CS::AddSentFinalPunct
T2A::CS::AddSubordClausePunct
T2A::CS::AddCoordPunct
T2A::CS::AddAppositionPunct
T2A::CS::ChooseMlemmaForPersPron
T2A::CS::GenerateWordforms
T2A::CS::DeleteSuperfluousAuxCP
T2A::CS::MoveCliticsToWackernagel
T2A::CS::DeleteEmptyNouns
T2A::CS::VocalizePrepos
T2A::CS::CapitalizeSentStart
T2A::CS::CapitalizeNamedEntitiesAfterTransfer

A2W::ConcatenateTokens
A2W::CS::ApplySubstitutions
A2W::CS::DetokenizeUsingRules
A2W::CS::RemoveRepeatedTokens
A2W::NormalizePunctuationForWMT

# write translated sentences
Write::Sentences
```

C. Shrnutí vybraných knihoven

V této sekci jsou stručně popsány funkce bloků z překladového scénáře, které jsme implementovali v rámci této práce. Bloky úzce souvisejí se zpracováním japonského textu a s fází transferu t-lemmat a formémů. Bloky používané pro generování českých vět jsou vynechány, jejich popis lze najít v dokumentaci TectoMT.

`W2A::JA::TagMeCab`, `Tool::Tagger::MeCab`

- Provádí tokenizaci, značkování slovních druhů a výběr lemmat.
- Vytváří stromovou strukturu a-roviny bez vyplněných závislostí mezi uzly.
- Uzlům a-stromu jsou nastaveny hodnoty atributů `a_lemma` a `tag`.

`W2A::JA::ParseJDEPP`, `Tool::Parser::JDEPP`

- Bloky mají za úkol na základě vyplněných hodnot `a_lemma` a `tag` provést "hrubý" závislostní parsing nikoliv přes samotné tokeny, ale přes bunsetsu (viz výše).
- Poté, co jsou určeny větné závislosti mezi jednotlivými bunsetsu, jsou dodělány závislosti mezi samotnými a-uzly.
- Mimo jiné také převádí číslování vrcholů používané externím parserem na číslování kompatibilní s platformou Treex.

`W2A::JA::RomanizeTags`

- Pomocí pevně daných substitučních pravidel provádí romanizaci (tj. převod japonských znaků do latinky) používaných tagů.

`W2A::JA::FixInterpunction`

- Blok sloužící k substituci UTF-8 reprezentace japonské tečky (znak 。), otazníku (znak ？) a vykřičníku (znak ！) na konci věty za korespondující ASCII znaky.
- Uzly intrepunkce jsou navíc převěšeny na kořen a-stromu.

`W2A::JA::RehangAuxVerbs`

- Provádí prohození závislostí mezi samostatnými slovesy (動詞_自立 - *Dōshi_Jiritsu*) a pomocnými slovesy (動詞_非自立 - *Dōshi_HiJiritsu*).
- Pomocné sloveso by mělo být závislé na samostatném plnovýznamovém slovesu a ne naopak.

`W2A::JA::RehangCopulas`

- Mění zavěšení japonských sponových sloves (př. です).

`W2A::JA::FixCopulas`

- Upravuje lemmata neformálních tvarů sponových sloves (např. だ).

`W2A::JA::RehangConjunctions`

- Blok starající se o změnu topologie koordinačních a subordinačních částic.
- V tuto chvíli provádí převěšování stejně jako u ostatních částic, v budoucnu ale máme v plánu provést potřebné úpravy, aby převěšování bylo specifičtější.

`W2A::JA::RehangParticles`

- Převěšuje zbývající částice (助詞 - *Joshi*). Blok by měl být volán až po aplikaci všech specifických bloků manipulujících s částicemi v a-stromě.
- Vzhledem k tomu, že japonské částice zastávají podobnou funkci jako předložky, chceme, aby měly ve stromové struktuře stejné umístění.

`W2A::JA::SetAfun`

- Nastavuje hodnotu `afun` pro většinu částic. Defaultní hodnota je AuxP (s částicí je nakládáno jako s předložkou).

`W2A::JA::SetAfunParticles`

- Nastavuje `afun` pro zbylé uzly. Blok by měl být volán stejně jako `W2A::JA::RehangParticles` až ve chvili, kdy již byly zavolány všechny specifičtější bloky.
- Modifikuje zejména `afun` dále používané v bloku `A2T::MarkEdgesToCollapse`.

`A2T::JA::SetFormeme`

- Vyplňuje hodnotu `formeme` uzlů t-stromu podle pravidel popsaných výše.

`A2T::JA::SetGrammatemes`

- Vyplňuje hodnoty gramatémů uzlů t-stromu podle pravidel specifických pravidel.
- V současné době nastavuje pouze hodnotu `gram/sempos` a `gram/negation` u sloves.

`T2T::JA2CS::TrLAddVariants`

- Blok, který provádí překlad japonských t-lemmat do češtiny s použitím statistického překladového modelu.
- Uzlům českého t-stromu nastavuje hodnotu atributů `t_lemma`, `t_lemma_origin` a `t_lemma_variants`.

`T2T::JA2CS::TrFAddVariants`

- Blok provádějící pravděpodobnostní překlad japonských formémů na české podobným způsobem jako u t-lemmat.
- Uzlům českého stromu jsou nastaveny hodnoty atributů `formeme`, `formeme_origin` a `formeme_variants`.