Japonsko-český strojový překlad

Dušan Variš, Ondřej Bojar

Univerzita Karlova v Praze, Matematicko-fyzikální fakulta
dvaris@seznam.cz, bojar@ufal.mff.cuni.cz

Abstrakt: Článek popisuje prototyp japonsko-českého strojového překladače založeného na hloubkovém větném rozboru. Tento typ strojového překladu není v současné době ve srovnání s jinými metodami tolik rozšířen, věříme však, že některé jeho aspekty jsou schopny přispět k celkově lepší kvalitě výstupu. Nutnou součástí našeho úkolu je i získání a zpracování potřebných paralelních dat. Jelikož japonsko-česká paralelní data nejsou prakticky vůbec dostupná, snažili jsme se vyzkoušet různé postupy, které by nám pomohly tento nedostatek nahradit. Náš systém je založen na stejném principu jako anglicko-český překladač TectoMT. V naší práci jsme se snažili zachytit alespoň základní jazykové jevy charakteristické pro japonštinu. Náš hloubkový systém též porovnáváme se zavedeným frázovým modelem překladu. Navzdory počátečním očekáváním pracuje frázový překlad lépe i při relativním nedostatku paralelních dat.

1 Úvod

Tato práce se zabývá strojovým překladem (machine translation, MT) z japonštiny do češtiny. Hlavním zaměřením je přitom překlad s využitím hloubkového větného rozboru a jeho porovnání s frázovým překladem. 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, která 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 tzv. tektogramatické rovině, tj. hloubkového větného rozboru [8]. 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 [1].

S rozvojem této metody překladu souvisí i snaha vyzkoušet ji i na dalších jazykových párech, proto jsme se ji rozhodli 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 pro výzkum strojového překladu japonsko-český pár přínosný.

Tento jazykový pár je zajímavý i z pohledu shromažďování vhodných paralelních dat, neboť v současné době neexistují téměř žádné dostatečně velké japonsko-české korpusy ani žádné strojově čitelné slovníky.

1.2 Související práce

Náš systém využívá během překladu stejných principů jako transfer-based systém TectoMT, který pracuje ve třech krocích: nejprve provede analýzu vstupního textu na požadovanou úroveň abstrakce, poté je analyzovaný text převeden na analogickou reprezentaci v cílovém jazyce a nakonec jsou na cílové straně sestaveny přeložené věty. Jako vhodnou úroveň abstrakce jsme přitom po vzoru TectoMT zvolili tektogramatickou rovinu, známou např. z Pražského závislostního korpusu 2.0 [3]. Právě na této úrovni jsou totiž zachyceny hloubkové sémantické vztahy mezi uzly stromu, kterými jsou v tomto případě pouze plnovýznamová slova, což je vhodné pro náš jazykový pár. Stejná úroveň abstrakce nám navíc umožňuje použít během syntézy hotovou kaskádu nástrojů pro vygenerování českých vět.

2 Použité nástroje

Pro naše experimenty používáme systém pro zpracování přirozených jazyků Treex [8],[1] dříve známý pod názvem TectoMT [12]. Jeho modularita nám umožňuje nejen integrovat různorodé externí nástroje pro zpracování přirozených jazyků, ale i kombinovat statistické a pravidlové metody. Scénář našeho japonsko-českého překladu vychází ze vzoru anglicko-českého překladového scénáře používaného v TectoMT a jak již bylo řečeno, syntéza češtiny je identická.

Tokenizaci a značkování slovními druhy (POS tagging) japonských vět provádíme pomocí morfologického analyzéru a taggeru MeCab [7]. MeCab využívá sadu tagů IPADIC, obsahující téměř 70 morfosyntaktických kategorií v hierarchické struktuře (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 navazujícím parserem.

[1] http://ufal.mff.cuni.cz/treex
[2] http://chasen-legacy.sourceforge.jp/

Vstup	彼は本を読まない人だ							
MeCab	彼	は	本	を	読ま	ない	人	だ
Bunsetsu	彼は		本を		読まない		人だ	
Význam	on		kniha		nečíst		člověk	

Obrázek 1: Příklad tokenizace věty „On je člověk, který nečte knihy“ MeCabem a tokenizace na bunsetsu pro JDEPP.

Závislostní parsing provádí JDEPP [14],[3] jehož přesnost (accuracy) zavěšování jednotlivých uzlů dosahuje zhruba 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 uvnitř jednotlivých bunsetsu dotváříme až v následujících blocích Treexu. Příklad tokenizace na bunsetsu a tokenizace MeCabem je zobrazen na obrázku 1.

3 Použitá data a jejich zpracování

Při tektogramatickém překladu dochází k převodu vybraných atributů mezi uzly zdrojového a cílového tektogramatického stromu (t-stromu), konkrétně tektogramatických lemmat neboli t-lemmat a formémů, viz sekci 5 níže. Volbu vhodných protějšků t-lemmat a formémů v cílovém jazyce zajišťují pravděpodobnostní unigramové překladové modely. K jejich tréninku používáme japonsko-české slovníky obsahující frekvenci výskytu jednotlivých dvojic unigramů t-lemmat a formémů. Tato sekce popisuje extrakci těchto slovníků z dostupných paralelních dat.

V současné době jako zdrojová data používáme paralelní korpusy s větným zarovnáním, viz tabulka 1. Japonsko-anglická data jsou zpracována nezávisle na anglicko-českých datech. V obou případech je prováděna hloubková analýza vstupních vět. U anglicko-českých dat byl tento krok proveden již ve zdrojovém korpusu CzEng a my jen přebíráme hotové anotace. Postup analýzy na t-rovinu je pro jednotlivé jazyky popsán v následující podsekci.

3.1 Lingvistické předzpracování

Analýza anglických a českých vět byla proveda kaskádou nástrojů Treex, stejnou jako používá i překladač TectoMT. Tagging anglických vět provedl tagger Morče [13], u českých vět byl pro tyto účely použit tagger Featurama,[5]. Povrchový parsing pak v obou případech zajistil MST parser [9]. Zbylé kroky zahrnovaly konstrukci t-roviny v závislosti na povrchovém parsingu a konstrukci t-lemmat,

Zdroj	Počet vět	Počet tokenů	
		japonština	angličtina
Kyoto's Wiki articles[6]	500k	11,0M	9,9M
Tanaka Corpus[7]	150k	1,7M	1,1M
Reuters Corpora[8]	56k	1,9M	1,3M
		čeština	angličtina
CzEng 1.0[9]	15 136k	206,4M	232,7M

Tabulka 1: Přehled použitých dat. Počty tokenů byly spočteny na námi tokenizovaných větách.

která byla později použita při slovním zarovnání a samotné stavbě slovníku.

Zpracování japonských vět jsme také prováděli v rámci platformy Treex, stejným způsobem jako v případě analýzy při samotném japonsko-českém překladu. Kroky jsou blíže popsány v sekci 4.1.

3.2 Zarovnání slov

Pro získání slovního zarovnání jsme použili program GIZA++ [10].[10] Spustili jsme jej na linearizované t-stromy, ve kterých každý uzel odpovídá jednomu plnovýznamovému slovu. Tím jsme se snažili vyhnout možnému problému řídkosti dat, který bývá často způsoben bohatou morfologií českého jazyka. Příklad slovního zarovnání je uveden na obrázku 2.

3.3 Stavba slovníku

Pro konstrukci slovníku jsme vyzkoušeli dva různé postupy. V prvním případě jsme vytvořili dílčí slovníky (japonsko-anglický a anglicko-český) z příslušných paralelních dat a ty jsme pak spojili skrze shodující se anglická hesla. Ve druhém případě jsme strojově přeložili anglické věty z japonsko-anglických dat, čímž jsme získali umělá japonsko-česká data. Z těch bylo možné japonsko-český slovník extrahovat přímo. Pro strojový překlad z angličtiny posloužila frázová komponenta soutěžního systému [1].

V obou případech jsme po získání slovního zarovnání provedli extrakci unigramových párů z linearizovaných t-stromů. Takto vzniklé slovníky obsahovaly i počty výskytů jednotlivých překladových dvojic.

Spojení dílčích slovníků bylo prováděno na základě shodných anglických hesel (viz obrázek 3). Poté byly podle vzorce 1 přepočítány „počty výskytů“ nově vzniklých slovních párů.

[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].

[5] http://sourceforge.net/projects/featurama/

[6] http://alaginrc.nict.go.jp/WikiCorpus/

[7] http://www.edrdg.org/wiki/index.php/Tanaka_Corpus

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

[9] http://ufal.mff.cuni.cz/czeng/

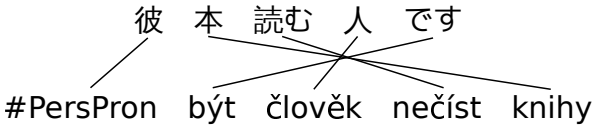
[10] http://code.google.com/p/giza-pp/

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**

Obrázek 3: Příklad japonsko-anglického (tabulka vlevo) a anglicko-českého (uprostřed) dílčího slovníku. Tučně 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 „agentura“ získal kvůli vysoké frekvenci výskytu v en-cs datech vyšší skóre než správný překlad „komunikace“.



Obrázek 2: Příklad slovního zarovnání t-lemmat věty „On je člověk, který nečte knihy“.

$$c(cs|ja) = \sum_{en}(c(en|ja) + w * c(cs|en)) \quad (1)$$

$$P(cs|ja) = \frac{c(cs|ja)}{c(ja)} \quad (2)$$

w udává váhu počtu výskytů dvojic v anglicko-českých datech. Její hodnotu jsme volili dle vlastního odhadu. Vzhledem k tomu, že hodnota $c(cs|ja)$ je vždy nezáporná, můžeme pak pravděpodobnost překladu japonských unigramů počítat klasicky podle vzorce 2.[11]

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 naštěstí t-lemmata zachycují alespoň nejčastěji se vyskytující složeniny. V případě češtiny se jedná zejména o zvratné zájmeno “*se*”, 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.

3.4 Nevýhody prostředního jazyka

Ať už jde o přímou extrakci, nebo spojování dílčích slovníků, v obou případech dochází kvůli prostřednímu 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 částečně omezili 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. Prvotní vhodní kandidáti pro tuto roli jsou bezesporu značky slovních druhů. Za zvážení by stálo i použití vhodných nástrojů pro zjednoznačnění významu (Word-Sense Disambiguation, WSD), kterými by se také daly potřebné příznaky v prostředním jazyce získat.

Dalším problémem je ztráta překladů některých japonských hesel. V japonsko-anglických datech se například mohou vyskytovat překlady pouze pomocí takových anglických hesel, 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.

4 Průběh překladu

V následujících odstavcích jsou popsány kroky aplikované v jednotlivých fázích překladu. Podrobněji 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.

[11] Ve skutečnosti je potřeba hodnotu $c(cs|ja)$ ještě normalizovat, aby byl součet $P(cs|ja)$ přes všechna česká hesla roven jedné.

4.1 Analýza

Každá vstupní věta je nejprve rozdělena na tokeny, a poté je provedeno značkování slovních druhů. Během taggingu je provedena i lematizace jednotlivých tokenů. K lematizaci dochází pouze u ohebných slovních druhů, zejména u sloves.[12]

Následně je postaven závislostní strom (a-strom). Vzhledem k tomu, že použitý parser 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.[13]

Podle podobných zvyklostí a-roviny pro češtinu a angličtinu je upravena topologie a-stromu. Vycházíme přitom též z konvencí korpusu Verbmobil použitých pro japonský jazyk [5]. 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.

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 a koordinačních částic) a „koncovky“ sloves (ty jsou také reprezentovány jako samostatné tokeny a označeny jako pomocná slovesa).

Po těchto úpravách je postaven tektogramatický strom (t-strom). Jeho uzly tvoří pouze plnovýznamová slova. Jak je zvykem, ponecháváme u t-uzlů reference na všechny a-uzly, které daný t-uzel reprezentuje, vztah mezi povrchovou a hloubkovou realizací je tedy možné i dodatečně studovat. 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 lze vidět na obrázku 4.

Před samotnou fází transferu jsou ještě všem uzlům t-stromu vyplněny formémy a částečně gramatémy. Funkce a podoba formémů je popsána v sekci 5. U gramatémů zatím vyplňujeme pouze negaci, ostatní kategorie by ovšem v rámci dalšího vývoje bylo také dobré vyplňovat.

4.2 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, [16]) 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).

V současné verzi 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.

4.3 Syntéza

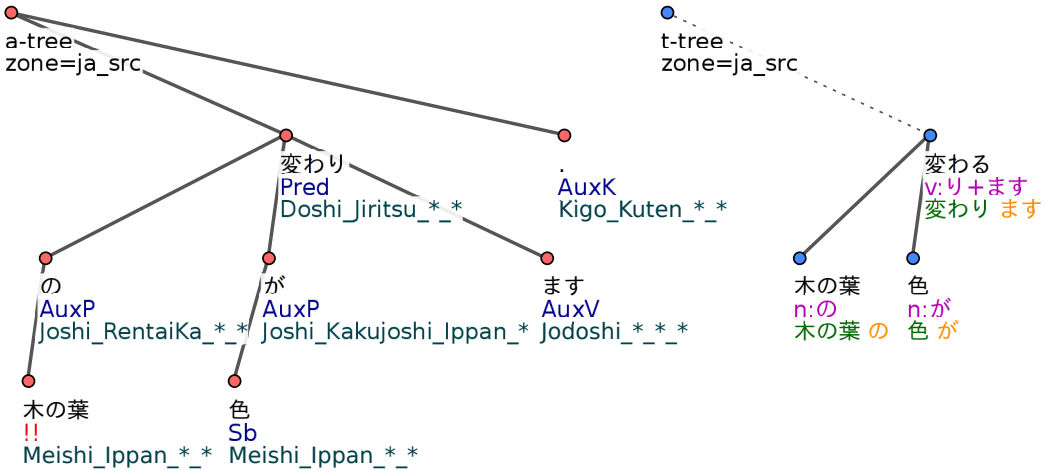
V závěru celého překladu je vygenerována česká věta na základě českého t-stromu. Je vytvořen a-strom a následně je vyplněna povrchová morfologie (rod, číslo, pád, atd.) s pomocí vyplněných formémů a gramatémů. Dále jsou vytvořeny a-uzly odpovídající pomocným slovesům, spojkám, předložkám atd. Kromě jiného dochází k vytvoření výsledných tvarů slov za pomoci generátoru slovních tvarů [2]. Syntézu českých vět podrobněji popisuje Žabokrtský [15].

5 Formémy

Po vzoru TectoMT používá náš systém formémy. Formémy popisují morfosyntaktické vlastnosti slov, tj. nesou např. informaci o tom, v jakém pádě bylo dané podstatné jméno vyjádřeno. Tektogramatická rovina sama o sobě záměrně od těchto vlastností abstrahuje (a je tak vhodná např. pro generování větných parafrází), pro věrný překlad je však vhodné původní formu výrazu ve vstupní větě zohlednit. Na české straně používáme zavedenou sadu formémů [17], japonské formémy v pracovní verzi navrhujeme sami. Protože v současné době japonské formémy používáme pouze během analýzy a překladu, nebyl kladen

[12] 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 tagsetu IPADIC jsou částice brány jako samostatné tokeny, které se, dle našeho názoru, svojí funkcí více blíží českým předložkám či spojkám.

[13] 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 4: 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.

velký důraz na zachování vlastností, které by pomohly při syntéze japonských vět.

Přiřazování hodnot japonských 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.

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),[14] kde v některých případech dochází k změně celého tvaru slovesa.

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.

Formémy přiřazujeme i příslovcím a příslovečným částicím, jež z hlediska sémantických slovních druhů nerozlišujeme.

V tabulce 2 je uveden fragment extrahovaného slovníku formémů. Jde vidět, že překlad formémů podstatných jmen a adjektiv alespoň v některých případech probíhá podle našich představ (viz české ekvivalenty formémů pro podmět a předmět, kde podle očekávání jako první možnost vychází `n:1`, tj. podstatné jméno v nominativu, resp. `n:4`, tj. akuzativu), v případě sloves jsou výsledky výrazně horší.

6 Experimenty a měření

V této sekci empiricky vyhodnocujeme kvalitu výstupu našeho překladového systému. Nejprve popíšeme sadu testovacích vět, jež jsme během našeho měření použili, a způsob, jakým byla zkonstruována. Dále představíme základní frázový systém, který jsme použili pro srovnání s naším překladačem. Následují výsledky našich měření a jejich interpretace v závěrečné diskusi.

6.1 Testovací data

Pro účely měření kvality překladu jsme náhodně vybrali 1000 dvojic vět, které se nepřekrývaly s našimi trénovacími daty, z našich japonsko-anglických paralelních dat, přesněji z korpusů 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

[14] 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. - základní hodnota	adj:1 adv	0.1612 0.1149
n:は subst. - téma nebo podmět	n:1 n:X	0.4369 0.1815
n:を subst. - předmět	n:4 n:1 n:X	0.2178 0.1225 0.1392
n:が subst. - podmět	n:1 n:X adj:attr n:4	0.3043 0.1907 0.1018 0.0857
v:り+なさる sloveso - zdvořilostní forma (stupeň „sonkeigo“)	v:inf v:fin adv	0.3148 0.2778 0.2407
n:に_と_の subst. se třemi částicemi に, と a の	v:že+fin v:fin n:s+7	0.2608 0.2173 0.1739
v:て_いる_ます sloveso - průběhový čas s pomocným slovesem v tzv. ます-tvaru	v:fin adj:1 adv	0.4754 0.1475 0.1229

Tabulka 2: 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.

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 hloubkový překladač pak při samotném překladu tento krok jednoduše přeskočil.

6.2 Frázový překladový systém

Pro porovnání s naším překladovým systémem jsme si vybrali frázový systém Moses [6].[15] Nejenže jakožto zástupce přímého překladu reprezentuje zcela odlišné paradigma přístupu k MT, konstrukce jednoduchého n-gramového překladače je také velmi snadná.

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, viz sekce 3.3.

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.

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.

Pro přípravu jazykového modelu jsme použili cílovou stranu našeho paralelního korpusu, tj. syntetickou češtinu. Očekáváme, že lepších výsledků by bylo možné dosáhnout při použití čistých českých dat. V prvním takovém experimentu však jazykový model založený na opravdické češtině dostal v automatickém ladění velmi nízkou váhu, a proto jsme jej nakonec nepoužili. Důvodem je pravděpodobně to, že i korpus pro ladění (2500 vět, viz výše) má cílovou stranu syntetickou, bez ruční korektury. Jakmile bude k dispozici více kvalitních japonsko-českých dat, pokus zopakujeme.

Frázový překladový systém jsme tímto způsobem natrénovali dvakrát, jednou na slovních formách, podruhé na lemmatech (tj. překlad do hrubší podoby češtiny).[16]

6.3 Automatické vyhodnocení

Výše uvedené systémy jsme spustili na stejném vzorku testovacích dat. Oba systémy měly téměř stejnou míru *OOV* (out-of-vocabulary, tj. podíl nepřeložených slov), kolem 3%. Za nepřeložená slova jsme přitom považovali všechny řetězce ve výstupu obsahující japonské znaky.

Automatické vyhodnocení jsme prováděli klasicky pomocí metriky BLEU [11]. V tabulce 3 uvádíme nejen celé BLEU, ale i přesnosti jednolivých *n*-gramů (kolik *n*-gramů z výstupu systému bylo nalezeno i v referenční větě). BLEU skóre hloubkového překladu vyšlo bohužel nulové. To je 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.

Všimněme si, že pouze v případě unigramů si hloubkový překlad vedl relativně dobře, stále ale hůře než frázový překladač. Jednou z příčin je nedostatek informací v japonské t-rovině, což po překladu ve fázi syntézy způsobuje, že nedochází k vygenerování všech potřebných pomocných slov. Vyšší *n*-gramy pak trpí tím, že v současné době neupravujeme slovosled, japonsko-český jazykový pár se ovšem slovosledem výrazně liší.

Co se týče kvality připravených slovníků, lepších výsledků jsme dosáhli se slovníky vytvořenými z našich umělých japonsko-českých dat. Metoda spojování dílčích slovníků dopadla výrazně hůř.

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

[16] Lematický výstup je nepoužitelný pro koncového uživatele ale je vhodný pro posouzení, zda překladač zachovává slova bez ohledu na morfologii, tj. lépe odráží přenos základního významu vět.

Druh překladu	1-gram	2-gram	3-gram	4-gram	BLEU
Slovní formy					
Treex (ja-en-cs)	13,2	0,0	0,0	0,0	0,00
Treex (ja-cs)	24,4	0,5	0,0	0,0	0,00
Moses	31,0	9,3	3,7	1,7	6,57
Lemmata					
Treex (ja-en-cs)	17,7	0,0	0,0	0,0	0,00
Treex (ja-cs)	40,5	2,3	0,2	0,0	0,00
Moses	53,2	21,5	10,6	5,3	15,95

Tabulka 3: Přesnosti jednotlivých *n*-gramů a celkové BLEU. Porovnáváme hloubkový překlad se spojovanými slovníky (ja-en-cs), s přímými slovníky (ja-cs) a frázový překlad (Moses).

	Lepší	Stejně dobré	Stejně špatné
Treex	24	10	34
Moses	32		

Tabulka 4: Ruční vyhodnocení na vzorku 100 vět. Tabulka uvádí, kolikrát byl překlad dané věty od jednoho systému lepší než od druhého, kolikrát byly oba překlady zhruba stejně dobré a kolikrát zhruba stejně špatné.

6.4 Ruční vyhodnocení

Ruční vyhodnocení se opírá o vzorek 100 vět z našich testovacích dat. Hodnotili jsme, který systém přeložil větu lépe, v případě podobné kvality jsme rozlišovali, zdali byly oba překlady stejně dobré nebo stejně špatně. Anotátor přitom nevěděl, která věta byla vygenerována kterým systémem. Hodnocení překladu vycházelo zejména z porovnání s naším referenčním překladem, nikoli vstupní větou.

Vzhledem ke značným nedostatkům obou systémů jsme byli během hodnocení velmi shovívaví a pomíjeli např. špatné skloňování nebo slovosled. Výsledky ruční evaluace jsou uvedeny v tabulce 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).

6.5 Diskuse

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. Uveďme několik příkladů kratších vět a zkusme na nich ilustrovat slabiny našeho systému.

(1a) SRC すぐに戻ります .
(1b) REF Brzy se vrátím .
(1c) Treex Dříve vrátí se .

Při porovnání s referenčním překladem by se mohlo zdát, že náš systém v případě této věty úplně selhal při generování slovních tvarů. Je ale potřeba podotknout, že ve zdrojové větě není explicitně uvedena osoba u slovesa „vrátit se“. Pomocí bloků s ručními pravidly by se dalo v těchto případech přiřadovat implicitně první osobu čísla jednotného, která se při nedostatku vhodného kontextu při překladu používá. Až na slovosled a drobnou chybu při překladu výrazu „すぐに“ (*sugu ni* - „brzy“), překlad dopadl obstojně.

(2a) SRC 夕方の五時です .
(2b) REF Je pět hodin večer .
(2c) Treex Večer páté době je .

Ve větě 2 došlo k nejvýraznější chybě při překladu slova „時“ (*toki* - „čas, doba“), které ovšem ve spojení se slovem „五“ (*go* - „pět“) nabývá významu jednotky času (五時 - „pět hodin“). Chybu tedy hledejme v našem překladovém modelu, dále pak do určité míry v HMTM, který měl v závislosti na kontextu („pět hodin“) nalézt vhodnou alternativu z kandidátů na překlad. Mimo jiné byl opět zachován slovosled zdrojové věty.

(3a) SRC 由美は , 私の友達のひとりです .
(3b) REF Yumi je jednou z mých přátel .
(3c) Treex Jumi má přátel sám je .
(3d) Moses Jumi je jeden z mých přátel je

Příklad 3 ukazuje, že alespoň v některých případech byl náš systém schopný konkurovat frázovému překladu (Moses). U frázového překladu došlo v tomto případě k vygenerování většího množství tokenů než bylo potřeba. Hloubkový systém v překladu japonského výrazu „私の“ (*watashi no* - „můj“)[17], zvolil naprosto špatné cílové t-lemma („mít“). Tato chyba je zřejmě důsledkem filtrování našich překladových slovníků, neboť předpokládaný správný překlad na obecné zájmenné t-lemma („#PersPron“) byl ze slovníku odstraněn. Je tedy potřeba v budoucnu zvážit, zdali jsou automatické filtrace spojených slovníků žádoucí. Překlad slova „ひとり“ (*hitori* - „jeden“) byl také v daném kontextu špatný („sám“).

(4a) SRC 良い言葉は教育の結果である .
(4b) REF Dobrá řeč je výsledkem vzdělávání .
(4c) Treex Dobré slovo vzdělávání výsledky je .
(4d) Moses Dobrá slova , a výsledek je , že je vzdělání .

Jako poslední příklad 4 uvádíme mírně lepší výsledek našeho překladu. V tomto případě hloubkový překlad dokonce předčil naši verzi frázového překladu. Tak jako ve všech ostatních případech má po hloubkovém překladu výsledná věta špatný slovosled, který v tomto případě citelně zhoršuje srozumitelnost.

V případě složitějších vět a souvětí dopadl překlad vždy výrazně hůř. U hloubkového překladu se totiž se zvyšující komplexitou analyzovaných závislostních struktur zvyšovala i šance na vnesení nových chyb.

[17] Přesněji se jedná o zájmeno „私“ (*watashi* - „já“) uvedené částicí „の“ (*no*) do pozice atributu.

7 Budoucí práce

Z výsledků vyhodnocení kvality našeho překladu usuzujeme, že by v současné době největší zlepšení přineslo především pečlivé automatické vyplňování všech potřebných atributů t-roviny během fáze analýzy. Také je nutné do budoucna provést důkladnější revizi japonské sady formémů, které jsou nyní například u sémanických sloves nevyhovující. K lepší čitelnosti a srozumitelnosti cílových vět by určitě přispěla i úprava jejich slovosledu.

Z hlediska využití pivotního jazyka kvůli nedostatku přímých dat stojí za úvahu překlad přes anglickou t-rovinu. Systém by provedl analýzu japonské věty, transfer na anglický t-strom a místo generování rovnou další transfer na český t-strom. Teprve zde by následovalo standardní generování výstupní věty. Tímto způsobem bychom se vyhnuli zejména problémům, které souvisí se spojováním dílčích slovníků či extrakcí slovníků z umělých japonsko-českých dat.

8 Závěr

Tato práce popsala naši prvotní verzi japonsko-českého překladače založeného na principu hloubkového překladu. Překladač byl implementován do prostředí Treex. V porovnání s frázovým překladem náš systém bohužel stále zaostává, jsme si ale vědomi jeho nedostatků a možných budoucích vylepšení.

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

Poděkování

Práce na tomto projektu byla podpořena grantem FP7-ICT-2011-7-288487 (MosesCore) Evropské unie.

Reference

[1] Ondřej Bojar, Rudolf Rosa, and Aleš Tamchyna. Chimera – Three Heads for English-to-Czech Translation. In *Proc. of the WMT*, pages 92–98, Sofia, Bulgaria, 2013. ACL.

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

[3] Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. Prague Czech-English Dependency Treebank 2.0, 2012. `http://hdl.handle.net/11858/00-097C-0000-0015-8DAF-4`.

[4] Yasuhiro Kawata. *Tagsets for Morphosyntactic Corpus Annotation: The Idea of a 'reference Tagset' for Japanese*. University of Essex, 2005.

[5] Yasuhiro Kawata and Julia Bartels. Stylebook for the Japanese Treebank in VERBMOBIL. Technical report, 2000.

[6] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open Source Toolkit for Statistical Machine Translation. In *ACL Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June 2007. ACL.

[7] Taku Kudo. Mecab: Yet another part-of-speech and morphological analyzer. `http://mecab.sourceforge.net/`, 2005.

[8] David Mareček, Martin Popel, and Zdeněk Žabokrtský. Maximum Entropy Translation Model in Dependency-Based MT Framework. In *Proc. of WMT and MetricsMATR*, pages 207–212, Uppsala, Sweden, 2010. ACL.

[9] Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. Non-Projective Dependency Parsing using Spanning Tree Algorithms. In *Proc. of HLT/EMNLP*, 2005.

[10] Franz Josef Och and Hermann Ney. A Comparison of Alignment Models for Statistical Machine Translation. In *Proc. of COLING*, pages 1086–1090. ACL, 2000.

[11] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proc. of ACL*, pages 311–318, Philadelphia, Pennsylvania, 2002.

[12] Martin Popel and Zdeněk Žabokrtský. Tectomt: Modular nlp framework. In *Proceedings of the 7th International Conference on Advances in Natural Language Processing*, IceTAL'10, pages 293–304, Berlin, Heidelberg, 2010. Springer-Verlag.

[13] Drahomíra Spoustová, Jan Hajič, Jan Votrubec, Pavel Krbec, and Pavel Květoň. The best of two worlds: Cooperation of statistical and rule-based taggers for czech. In *Proc. of the Workshop on Balto-Slavonic Natural Language Processing, ACL 2007*, pages 67–74, Praha, 2007.

[14] Naoki Yoshinaga and Masaru Kitsuregawa. Kernel slicing: scalable online training with conjunctive features. In *Proc. of COLING*, pages 1245–1253, Beijing, China, 2010. ACL.

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

[16] Zdeněk Žabokrtský and Martin Popel. Hidden Markov Tree Model in Dependency-based Machine Translation. In *Proc. of the ACL-IJCNLP Short Papers*, pages 145–148, Suntec, Singapore, 2009. ACL.

[17] Zdeněk Žabokrtský, Jan Ptáček, and Petr Pajas. TectoMT: Highly Modular Hybrid MT System with Tectogrammatics Used as Transfer Layer. In *Proc. of WMT*, pages 167–170, Columbus, Ohio, USA, 2008.