Problémy recyklování systému automatického překladu*

Ondřej Bojar, Petr Homola, Vladislav Kuboň

Ústav formální a aplikované lingvistiky
ÚFAL MFF UK, Malostranské náměstí 25, Praha 1, CZ-11800
Česká republika
{bojar,homola,vk}@ufal.mff.cuni.cz

Abstract. Tento článek se zabývá problémy, spojenými se snahou využít jednotlivých modulů systému automatického překladu mezi češtinou a ruštinou RUSLAN pro překlad do typově odlišného jazyka – do angličtiny. Tato snaha je motivována hypotézou, že přes veškeré problémy spojené s oživováním původního systému, je použití jeho klíčových modulů jednodušší než vývoj zcela nových modulů. Původní idea projektu počítala s použitím dvou klíčových částí sytému, syntaktické analýzy češtiny a jeho slovníku. V první části článek představuje systém RUSLAN, vyvíjený v letech 1985-90 pro překlad manuálů operačních systémů sálových počítačů z češtiny do ruštiny. V následující kapitole je rozebrána současná situace v oblasti automatického překladu z češtiny do angličtiny a je vysvětlena motivace pro recyklaci částí systému. Klíčovou část článku tvoří popis jednotlivých součástí systému, včetně nových modulů, kterými reagujeme na problémy vzniklé při recyklaci původního systému.

1 Úvod

V minulosti již bylo do nejrůznějších systémů automatického překladu investováno obrovské množství úsilí a prostředků. Výsledky však většinou ani zdaleka nebyly uspokojivé. Samozřejmě, existují výjimky, historie automatického překladu zná mnoho příkladů úspěšných systémů, ať již komerčních nebo experimentálních, ale počet těch systémů, které spolykaly obrovské množství lidské práce a finančních prostředků a které byly potichu opuštěny, když se ukázalo, že jejich výsledky nesplňují očekávání zadavatelů, je mnohem vyšší. Toto tvrzení se vztahuje zvláště na klasické překladové systémy založené na ručně vytvářených pravidlech, které obvykle vyžadují značné investice do vývoje gramatik a slovníků ještě předtím, než začnou produkovat výsledky v přijatelné kvalitě. Není náhodou, že vývoj takových systémů, jako je např. Systran či moskevský Etap, trval několik desetiletí - zkvalitňování takového systému je dlouhý proces, zahrnující zvětšování slovní zásoby, ladění gramatiky a testování celého systému.

* Výzkum popsaný v tomto článku je podporován grantem GAČR 405/03/0914 a zčásti také grantem GAUK 207-10/203330

Poslední desetiletí bylo svědkem několika pokusů o zvýšení kvality systémů automatického překladu prostřednictvím nových metod. Silný důraz na pravděpodobnostní a statistické ve zpracování přirozeného jazyka obecně se projevil i v oblasti automatického překladu. Začaly se objevovat hybridní systémy, překladatelskou veřejností začaly být široce přijímány systémy využívající tzv. překladovou paměť, objevily se úspěšné prototypy systémů na překlad mluvené řeči (i když zatím pouze s velmi omezenou tématickou doménou). Všechny tyto (a také mnoho dalších) trendy v posledních letech dosáhly celé řady povzbuzujících výsledků. Hledání nových metod a jejich využívání v praxi rozhodně posouvá celé odvětví automatického překladu kupředu, neměli bychom ale zapomínat na úsilí, investované do starších systémů. Opětovné použití, recyklování, celých systémů nebo jejich částí by mohlo pomoci snížit náklady na vývoj nových systémů, zvláště pokud se jedná o tzv. malé jazyky, pro které neexistuje takové množství nejrůznějších nástrojů, gramatik a slovníků jako např. pro angličtinu, japonštinu, němčinu nebo španělštinu. V tomto článku bychom rádi popsali jeden takový pokus o recyklování částí existujícího systému pro nový jazykový pár a problémy, na které jsme při práci narazili.

2 RUSLAN

Jeden ze systémů, které byly v tichosti opuštěny na počátku devadesátých let byl systém automatického překladu manuálů k sálovým počítačům, nazývaný RUSLAN [1]. Tento systém byl vyvíjen ve druhé polovině osmdesátých let jako společný projekt mezi MFF UK Praha a Výzkumným ústavem matematických strojů v Praze.

Tento systém využíval architekturu založenou na existenci fáze transferu. Jeho autoři zpočátku tvrdili, že pro tak příbuzné jazyky, jako je čeština a ruština, bude z důvodu syntaktické podobnosti obou jazyků fáze transferu nepatrná. Tento předpoklad se ukázal jako mylný, testování a ladění systému jednoznačně ukázalo, že množství rozdílů mezi oběma jazyky vede k celé řadě speciálních transferových pravidel.

Systém byl prakticky kompletně implementován v Q-systémech, formalismu vytvořeném Alainem Colmerauerem [2]. Q-systémy jsou v podstatě grafovým analyzátorem (chart parserem). Poprvé byly úspěšně použity v systému automatického překladu TAUM-METEO na montrealské univerzitě. Dovolují rozdělit gramatiku do modulů, které na sebe navazují tak, že výstupní graf předchozího modulu slouží jako vstup do modulu následujícího. Každý modul se skládá z množiny pravidel, která v podstatě popisují transformace stromových struktur. Ačkoli byl tento formalismus mnoha odborníky považován za zastaralý, Q-systémy prokázaly, že jsou spolehlivým nástrojem pro prakticky orientované projekty - interpret gramatiky napsané v Q-systémech byl dostatečně rychlý a spolehlivý i pro gramatiky pokrývající velkou část syntaxe obou jazyků.

Kromě gramatiky systém také spoléhal na sadu slovníků, které obsahovaly veškerá data využívaná jednotlivými moduly systému. Každá lexikální jednotka v hlavním (dvojjazyčném) slovníku neobsahovala pouze lexikálně-syntaktická

data (valenční rámce apod.), ale také množinu sémantických rysů, které se primárně používaly k řešení syntaktických víceznačností v průběhu syntaktické analýzy zdrojového jazyka (češtiny). Protože cílem celého systému byl překlad relativně omezené tématické domény, dosahovala velikost slovníku přibližně 8000 položek. Kromě nich však byl systém schopen překládat prostřednictvím speciálního modulu zvaného transdukční slovník dalších 2000 převážně technických a odborných termínů řecko-latinského původu pomocí algoritmu na přímou transkripci českých slovních forem do ruských [3].

Práce na systému RUSLAN byla ukončena v roce 1990 ve fázi konečného ladění a testování. Důvod ukončení prací na systému byl velmi jednoduchý – po roce 1989 nebyla po automatickém překladu z češtiny do ruštiny žádná poptávka.

3 Motivace

Poptávka po překladu z angličtiny nebo do angličtiny v letech po ukončení prací na systému RUSLAN dramaticky vzrostla. Na druhé straně se také podstatně zvětšilo množství metod, nástrojů a jazykových zdrojů pro automatický překlad. Pro češtinu byla vytvořena celá řada korpusů, z nichž nejznámějšími jsou morfologicky anotovaný Český národní korpus a syntakticky označkovaný Pražský závislostní korpus. V roce 2002 jsme zahájili práci na paralelním dvojjazyčném Pražském závislostním česko-anglickém korpusu (PCEDT) [4], který obsahuje asi polovinu textů z amerického korpusu PennTreebank 3, přeloženou do češtiny rodilými mluvčími. Dalším důležitým zdrojem dat, který byl pro češtinu také vyvinut v uplynulém období byl rozsáhlý morfologický slovník češtiny [5], který umožnil vysoce kvalitní morfologickou analýzu češtiny, která byla od té doby ověřena v bezpočtu komerčních aplikací a vědeckých projektů.

Uplynulé desetiletí bylo také svědkem vzniku několika stochastických analyzátorů češtiny, které by mohly být použity pro analýzu zdrojového jazyka v česko-anglických projektech. Ačkoli je kvalita výsledků těchto systémů o mnoho nižší než výsledky zveřejněné pro angličtinu – nejlepší výsledky, publikované Charniakem a Collinsem [6] se pohybují okolo 85% – je možné tyto analyzátory použít pro statistický překlad z češtiny.

Jedním z posledních stimulů pro naši snahu recyklovat části systému RUSLAN byly experimenty zahájené v roce 2001 M. Čmejrkem a J. Curínem. Jejich projekt (popsaný například v [7] se zaměřil na plně automatický statistický překlad, zahrnující stochastickou analýzu zdrojového textu na tektogramatickou (hloubkově syntaktickou) rovinu. Ačkoli tento projekt původně také měl využívat statistický transfer, novější experimenty jej nahradily modulem založeným z větší části na ručně vytvářených pravidlech. Tento posun od stochastického k ručně vytvořenému transferu otevřel otázku, zda by také nebylo lepší využít ručně psané gramatiky pro syntaktickou analýzu češtinu spíše než statistický analyzátor. Vyšší počet nesprávně přiřazených hran v závislostním stromě nemusí nutně znamenat horší analýzu. Pro systém automatického překladu je možná důležitější schopnost správně zanalyzovat souvislé části vstupních vět a tak umožnit správný překlad jednotlivých klauzí. Z našeho pohledu by systém

automatického překladu mohl sloužit jako jistý testovací nástroj, umožňující porovnání výsledků ručně vytvořených gramatik se statistickými analyzátory. Takové srovnání je obtížné, pokud chceme dostat opravdu objektivní výsledky. Důvodem je to, že standardní míry kvality analýzy jsou příliš spojeny s určitým typem dat a s určitým typem syntaktického značkování, použitého při budování syntaktických korpusů.

Hlavní motivací našeho česko-anglického experimentu bylo otestování několika hypotéz. Nejvýznamnější je hypotéza týkající se úrovně, na které má probíhat transfer. Vzhledem k odlišnostem mezi oběma jazyky není možné provádět transfer bezprostředně po morfologické analýze nebo po fázi mělké (částečné) syntaktické analýzy, jak to provádí systém Česílko, zaměřující se na překlad mezi blízce příbuznými (a podobnými) jazyky [viz [8]]. Na druhé straně je otázkou, zda typologické odlišnosti mezi češtinou a angličtinou opravňují použití transferu na tektogramatické (hloubkově syntaktické) úrovni, jako to bylo realizováno Cuřínem a Čmejrkem v jejich systému. Podle našeho názoru by transfer provedený na úrovni povrchové syntaxe mohl být vhodnější. Hlavním problémem transferu na tektogramatické (hloubkově syntaktické) úrovni je nízká kvalita výsledků analýzy na této úrovni, které jsou mnohem horší než výsledky stochastické analýzy na analytickou (povrchově syntaktickou) úroveň. Všechny výše uvedené výsledky patří analyzátorům zaměřujícím se na analytickou (povrchově syntaktickou) úroveň.

V neposlední řadě bylo naším záměrem vyvinout na pravidlech založený systém s minimálními možnými náklady, ať již metodou recyklace použitelných existujících modulů či snahou o využití (polo)automatických metod tam, kde to jen bude možné, se zaměřením na oblasti, kde by využití lidské práce bylo příliš drahé (například při budování rozsáhlého dvojjazyčného slovníku, viz níže.)

4 Jednotlivé moduly nového systému

Hlavním cílem našeho projektu bylo vyvinout experimentální systém automatického překladu pro překlad textů z korpusu PCEDT do angličtiny. Tento systém zkoumá, zda je možné prostřednictvím znovupoužití existujících zdrojů (gramatiky, slovníku) zkrátit čas potřebný k vývoji systému. Projekt také využívá dvojjazyčný paralelní korpus syntakticky označkovaných textů, i když nikoli jako přímý zdroj trénovacích dat, spíše jako dodatečný zdroj lingvistických dat užitečný zejména pro vývoj slovníku a pro testování celého systému.

4.1 Morfologická analýza

Kromě původního překladového systému RUSLAN a dvojjazyčného korpusu PCEDT můžeme také využít modul morfologické analýzy češtiny [5]. Tento modul pokrývá téměř celou českou slovní zásobu, s několika málo výjimkami (odhaduje se, že tato morfologická analýza je schopná rozeznat 800 000 lemat). Jediným problémem bylo jeho zapojení do systému – původní modul morfologické analýzy v systému RUSLAN byl velmi úzce svázán se slovníkem a s mor-

fologickým modulem. Nová morfologická analýza také používá odlišnou sadu značek než ta původní.

4.2 Dvojjazyčný slovník

Původně jsme předpokládali, že slovník použitý v systému RUSLAN budeme převádět a budeme se snažit využít velkého množství pečlivě zpracovaných lexikálně syntaktických údajů v něm obsažených. Tento předpoklad se ale ukázal jako příliš optimistický. Ačkoli jsou informace z původního slovníku pro modul syntaktické analýzy češtiny nesmírně cenné, rozhodli jsme se tyto informace obětovat. 8000 hesel původního slovníku tvoří příliš malou část slovní zásoby potřebné pro nový systém, než aby stálo zato kvůli těmto informacím obětovat konzistenci údajů v novém slovníku. Proto jsme se rozhodli tento slovník vytvořit znovu.

Budování česko-anglických překladových slovníků stálo již mnoho úsilí a je jistě vhodné raději znovu použít dostupné slovníky než budovat slovníky zcela nové. Pro češtinu a angličtinu jsou však dostupné pouze slovníky určené pro lidské uživatele (např. WinGED[1] nebo [9]), které trpí mnohými problémy:

- jedno překladové heslo[2] v některých případech kóduje více překladových variant současně
- cenné morfologické nebo i syntaktické informace jsou ve slovníku dostupné jen zřídka a navíc způsobem, který není nijak standardizován (Některá slovesná hesla např. obsahují nejen sloveso, ale též formy doplnění slovesa. Bohužel se způsob zápisu formy liší heslo od hesla, někde je uvedena jen předložka, někde předložka a klíčové slovo *(ně)koho* ap. Klíčová slova se navíc nijak formálně neliší od běžných slov, která jsou povinnou součástí hesla.)
- většina hesel neobsahuje vůbec morfologickou a syntaktickou informaci

Pro použití slovníku v systému strojového překladu je třeba tyto informace doplnit, jinak systém např. nebude vědět, zda pro překlad podstatného jména *stát* ve vstupní větě smí použít heslo *stát = stand.* Podobně důležité je, aby víceslovná hesla obsahovala popis syntaktické struktury zdrojového výrazu, protože bez ní je velmi obtížné ve fázi transferu ve větě, která již byla syntakticky analyzována, spolehlivě poznat, kde je které překladové heslo.

Čištění a doplňování anotace do slovníku probíhá v několika oddělených fázích, částečně manuálně a částečně automaticky za použití dostupných paralelních nebo jednojazyčných korpusů:

Identifikace metainformací Všechna slovníková hesla, která obsahovala slova typická pro označení nějaké speciální informace (např. *někoho*, *sg.* ap.), jsme

[1] `http://www.rewin.cz/`
[2] dvojice český výraz–anglický výraz

ručně prošli po blocích podle nalezeného podezřelého slova. Tento způsob zpracování nám umožnil soustředit se vždy na jeden konkrétní jev a zpracovat velké množství hesel současně.

Během této fáze jsme tedy formalizovali všechny metainformace a rovněž jsme roznásobili jednotlivé varianty hesel, které byly dosud sloučeny do jednoho hesla.

Identifikace slovních druhů Po automatické morfologické analýze hesel pomocí analyzátoru [5] bylo nutné provést zjednoznačnění slovního druhu u sporných slov. Toto zjednoznačnění bylo nutné provést opět ručně, po blocích se stejným typem nejednoznačnosti, neboť řada spojení umožňuje více různých interpretací, z nichž však jen jedna odpovídá anglickému překladu. (Např. spojení *kniha účetní* je buď *account book*, nebo prostě kniha, která patří nějaké účetní.)

Přidání morfologických omezení Zejména pro víceslovná hesla je nutné doplnit, které morfologické rysy jsou ve výrazu povinné a které je možné libovolně měnit. Např. hesla typu adjektivum a substantivum typicky vyžadují shodu pádu, čísla a rodu. U bohatších konstrukcí může být seznam omezení složitější.

Morfologická omezení doplňujeme automaticky. Předem jsme definovali repertoár typických omezení (např. pevně daný pád jména nebo zmiňovaná shoda v některé morfologické kategorii) a poté využili Českého národního korpusu (ČNK[3]) k tomu, abychom pro každé heslo na základě pozorovaných (pravděpodobných) výskytů identifikovali, jaká omezení heslo vždy splňuje. Tento postup naráží na problémy jednak u velmi řídkých hesel (kde ČNK neposkytuje dostatek výskytů na to, aby bylo možné spolehlivě rozhodnout, zda dané omezení platí) a jednak u hesel složených z velmi četných slov. V takovém případě totiž najdeme příliš velké množství náhodných kolokací všech slov hesla, nicméně tato slova navzájem nesplňují morfologická omezení (protože koneckonců nejde o výskyt daného hesla), takže algoritmus morfologické omezení neodhalí.

Doplnění syntaktické informace Vzájemné závislostní vztahy mezi slovy hesla přidáváme opět ručně, po blocích se všech hesel se stejnými slovními druhy jednotlivých slov. Bylo by rovněž možné postupovat podobně jako při přidávání morfologických omezení, bylo by však nutné použít syntakticky analyzovaný korpus nebo korpus automaticky syntakticky analyzovat.

4.3 Modul rozpoznávající pojmenované entity

Tento relativně nezávislý modul se stará o zpracování idiomů, terminologických jednotek a ostatních typů pojmenovaných entit.

[3] `http://ucnk.ff.cuni.cz/`

5 Pojmenované entity

Pojmenované entity (*named entities*, NE) jsou atomické jednotky, např. vlastní jména, časové výrazy či množství. Vyskytují se často v různých typech textů a nesou důležitou informaci. Proto má odpovídající analýza pojmenovaných entit a jejich správný překlad podstatný dopad na kvalitu strojového překladu [10].

Překlad NE zahrnuje jak sémantický překlad, tak fonetickou transliteraci. Různé typy NE se překládají různě. Například osobní jména se nepřekládají sémanticky (nutná je v některých případech pouze transliterace), zatímco některé tituly či označení profese se překládají (např. *první dáma Laura Bushová → first lady Laura Bush*). V případě organizací obvykle stačí regulární pravidla pro analýzu jmenných frází (NP) (např. *Ústav formální a aplikované lingvistiky → Institute of formal and applied linguistics*), ačkoliv někdy je vhodnější idiomatický překlad. Co se týče geografických názvů, používáme zvláštní glosář.

Pro rozpoznávání NE jsme vyvinuli gramatiku založenou na regulárních výrazech, jež pracuje s typovanými strukturami rysů. Interpretr této gramatiky, podobný formálně poněkud slabší platformě SProUT [11], používá řetězcové grafy a unifikaci, gramatika tedy sestává z pravidel, kde levá strana je regulární výraz nad typovanými strukturami rysů s proměnnými, jenž reprezentuje rozpoznaný podřetězec vstupu, a pravá strana definuje výstupní strukturu rysů po aplikaci pravidla. Hierarchie typů je definována globálně, jednoduchý příklad je na obr. 1.

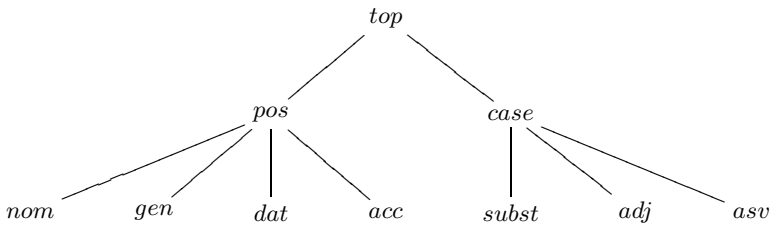

Fig. 1. Příklad jednoduché hierarchie typů

Gramatika rozpoznávající NE je inspirována pravidly popsanými v [12]. Příkladem jednoduchého pravidla je:

```
#subst[LEMMA: ministerstvo]$s1
+ #top[CASE: gen, PHRASE: $phr]$s2
== $s1#ministry[ATTR: $s2,
PHRASE: &('ministerstvo ' + phr)]
```
(1)

První struktura rysů reprezentuje libovolnou morfologickou variantu slova *ministerstvo* následovanou jmennou frází v genitivu. Proměnné $s1, $s2 a $phr zajišťují dynamické přiřazení hodnot a umožňují přenést tyto hodnoty na

příslušné místo ve výstupní struktuře rysů typu *ministry*. Tato struktura obsahuje nový atribut PHRASE s lematizovanou hodnotou celé fráze.

Je-li vstupem například

informace ministerstva zahraničí
o cestování do ohrožených oblastí (2)

potom fráze "ministerstva zahraničí" bude rozpoznána jako NE a v dalším procesu překladu s ní bude zacházeno jako s atomickou (nedělitelnou) jednotkou:

$$\begin{bmatrix} \textit{ministry} & \\ \text{LEMMA} & \text{ministerstvo} \\ \text{FORM} & \text{ministerstva} \\ \text{PHRASE} & \text{ministerstvo zahraničí} \\ \text{ATTR} & \begin{bmatrix} \textit{subst} & \\ \text{LEMMA} & \text{zahraničí} \\ \text{PHRASE} & \text{zahraničí} \\ \text{FORM} & \text{zahraničí} \\ \text{CASE} & \text{gen} \\ \text{NUMBER} & \text{sg} \\ \text{GENDER} & \text{n} \end{bmatrix} \\ \text{CASE} & \text{gen} \\ \text{NUMBER} & \text{sg} \\ \text{GENDER} & \text{n} \end{bmatrix} \tag{3}$$

Lematizace NE je klíčovým předpokladem pro strojový překlad. Zejména u jazyků s bohatou flexí může být při analýze problémem strukturní víceznačnost, např. vnitřní uzávorkování komplexních jmenných frází. Jádro gramatiky použité v tomto projektu vychází z gramatiky systému strojového překladu Česílko [8].

5.1 Syntaktická analýza češtiny

Ačkoli jsme původně předpokládali, že modul syntaktické analýzy češtiny bude vyžadovat pouze velmi malé úpravy, ukázalo se, že bude nutné jeho gramatiku doplnit zejména o taková pravidla, která se uplatní na novém vstupním textu. Jedná se například o pravidla zpracovávající obrovské množství číselných výrazů, kterými jsou texty z Wall Street Journalu doslova protkány. Přes tyto potíže však modul syntaktické analýzy zůstal jako jediný, který bylo možno plně uplatnit v novém systému.

5.2 Transfer

Hlavním úkolem tohoto modulu je převod syntaktických struktur (závislostních stromů) popisujících české věty ze vstupního textu do syntaktických struktur (stromů) odpovídajících anglických vět. Transfer se nezabývá překladem jednoduchých lexikálních jednotek (jednotlivých slov), tento překlad obstarává hlavní dvojjazyčný slovník systému v dřívějších fázích zpracování. Transfer se soustředí na tři hlavní úkoly:

– Převod českých závislostních stromů na anglické, zejména s ohledem na slovosledné změny, vyžadované cílovým jazykem.

– Odhalování a překlad těch českých konstrukcí, které vyžadují speciální překlad do angličtiny.
– Vkládání určitých a neurčitých členů do vět cílového jazyka.

Vývoj tohoto modulu stále pokračuje, počáteční testy ukázaly, že se v něm skrývají velké rezervy, které v budoucnu mohou pomoci zlepšit kvalitu výsledného překladu.

5.3 Syntaktická syntéza angličtiny

Syntaktická syntéza ruštiny je v systému RUSLAN velmi úzce svázána s transferem, proto se snažíme využít co největší část gramatiky, kdekoli je to možné, a doplnit nová pravidla podobně, jako je tomu u modulu syntaktické analýzy. Práce na tomto modulu stále pokračují.

5.4 Morfologická syntéza angličtiny

Vzhledem k jednoduchosti anglické morfologie má tento modul v našem systému pouze velmi omezenou roli. V podstatě se jeho úloha v této fázi projektu redukuje pouze na generování správných tvarů plurálu, třetí osoby a nepravidelných slovesných tvarů.

6 Výsledky překladu

Pro vyhodnocení výsledků jsme se rozhodli alespoň zpočátku používat metodu, popsanou v [8]. Tato metoda měří podobnost původního a přeloženého textu (původním textem zde označujeme opravdu původní text z Wall Street Journalu, který máme k dispozici v korpusu PCEDT). Tato podobnost se měří pomocí aplikace TRADOS Translation Workbench. Systém TRADOS dokáže totiž vyhodnotit procentuální shodu mezi dvěma větami, v našem případě tedy mezi původní větou z Wall Street Journalu a automatickým překladem jejího českého překladu zpět do angličtiny. Výhodou této metody je její snadná použitelnost pro monitorování pokroku při vývoji systému, v budoucnu ale plánujeme její nahrazení standardními a přesnějšími metodami, např. použitím tzv. BLEU skóre.

Článek [8] uvádí, že komerční systém automatického překladu PC Translator testovaný na 256 větách z korpusu PCEDT dosáhl váženého průměru shody ve výši 30%. Naše testy na odlišné (a menší) sadě vět ze stejného zdroje naznačují nepatrně lepší výsledky (okolo 35%). Toto měření však bylo pouze orientační a nelze z něj zatím odvozovat žádné závěry, systém stále prochází dalším vývojem a testováním.

7 Závěr

Ačkoli první testy ukázaly povzbuzující výsledky, stále existuje celá řada možností, jak systém dále vylepšit. Kromě práce na současných modulech

se nabízejí nejméně tři další směry pro další výzkum. Jedním z velkých problémů je morfologická víceznačnost jednotlivých českých slovních forem. Cesta, kterou jsme se rozhodli v budoucnu vyzkoušet, nevede přes přes stochastické značkování, ale přes částečnou (ale bezchybnou) desambiguaci výsledků morfologické analýzy češtiny. Ta by měla odstranit značné množství nesprávných značek a tím velmi ulehčit roli modulu syntaktické analýzy češtiny.

Dalším způsobem, jak bude možné snížit víceznačnost slov ze vstupního textu, je zapojení speciálního modulu, který bude schopen vyřešit lexikální víceznačnost v těch případech, kdy slovník nabídne více lexikálních ekvivalentů ke vstupním slovům. Tento stochastický model se bude rozhodovat na základě kontextu a pokusí se navrhnout nejlepší překlad.

Třetím směrem možného vývoje prací na systému je opuštění dědictví systému RUSLAN a nahrazení jeho ručně vytvořené syntaktické analýzy některým z dostupných stochastických analyzátorů češtiny. Takový experiment by mohl dát odpověď na klíčovou otázku celého projektu – opravdu se vyplatí recyklovat staré systémy nebo ne?

References

1. Oliva, K.: A Parser for Czech Implemented in Systems Q. Explizite Beschreibung der Sprache und automatische Textbearbeitung (1989)
2. Colmerauer, A.: Les Systemes Q ou un formalisme pour analyser et synthetiser des phrases sur ordinateur. (1969)
3. Bémová, A., Kuboň, V.: Czech-to-Russian Transducing Dictionary. In: Proceedings of the 13th International Conference COLING, Vol. 3. (1990) 314–316
4. Cuřín, J., Čmejrek, M., Havelka, J., Kuboň, V.: Building a Parallel Bilingual Syntactically Annotated Corpus. In: Proceedings of the 1st International Joint Conference on NLP. (2004)
5. Hajič, J.: Disambiguation of Rich Inflection - Computational Morphology of Czech. Volume I. Prague Karolinum, Charles University Press (2001) 334 pp.
6. Collins, M., Hajič, J., Brill, E., Ramshaw, L., Tillmann, C.: A Statistical Parser of Czech. In: Proceedings of 37th ACL Conference, University of Maryland, College Park, USA (1999) 505–512
7. Čmejrek, M., Cuřín, J., Havelka, J.: Czech-English Dependency-based Machine Translation. In: Proceedings of the 10th Conference of The European Chapter of the Association for Computational Linguistics, Budapest, Hungary (2003) 83–90
8. Hajič, J., Homola, P., Kuboň, V.: A simple multilingual machine translation system. In: In: Proceedings of the MT Summit IX, New Orleans (2003)
9. Svoboda, M.: GNU/FDL English-Czech Dictionary (2001) `http://slovnik.zcu.cz/`.
10. Babych, B., Hartley, A.: Selecting translation strategies in MT using automatic named entity recognition. In: Proceedings of the Ninth EAMT Workshop, Valetta, Malta. (2004)
11. Bering, C., Drożdżyński, W., Erbach, G., Guasch, C., Homola, P., Lehmann, S., Li, H., Krieger, H.U., Piskorski, J., Schaefer, U., Shimada, A., Siegel, M., Xu, F., Ziegler-Eisele, D.: Corpora and evaluation tools for multilingual named entity grammar development. (2003)

12. Piskorski, J., Homola, P., Marciniak, M., Mykowiecka, A., Przepiórkowski, A., Woliński, M.: Information extraction for Polish using the SProUT platform. In: Proceedings of the International IIS:IIP WM'04 Conference, Zakopane, Poland. (2004)