Budování česko-anglického slovníku pro strojový překlad*

Ondřej Bojar

Ústav formální a aplikované lingvistiky MFF UK
Malostranské náměstí 25, Praha 1, 118 00, Česká republika
bojar@ufal.mff.cuni.cz

Abstrakt Příspěvek popisuje probíhající práce na česko-anglickém slovníku pro strojový překlad. Slovník vychází z dostupných strojově čitelných slovníků, kombinací ručních a automatických metod jsou doplňovány informace nutné pro nasazení slovníku v automatickém systému. Jedná se zejména o morfologická omezení nutná pro uplatnění hesel, popis syntaktické struktury hesel a překladové ekvivalenty slovesných rámců.

1 Úvod

Strojový překlad je obtížnou úlohou, kterou nelze uspokojivě řešit bez rozsáhlých a kvalitních slovníků, zejména v případě, že chceme dát přednost strukturálnímu přístupu před přístupem čistě statistickým (pro podrobný přehled celé škály metod strojového překladu viz např. [1]). V případě překladu mezi češtinou a angličtinou se strukturálním překladem experimentují zejména [2], kteří překládají z češtiny texty ekonomické povahy. Naše práce je pak spjata s projektem GAČR 405/03/0914, jehož cílem je přizpůsobit překladový systém Ruslan [3] pro překlad do angličtiny místo původní ruštiny.

V současné době není k dispozici žádný vhodný slovník pro strojový překlad. Z toho důvodu např. i [2] používají zjednodušený slovník s překladovými hesly[1] omezenými na jedno slovo na české straně a nejvýše dvouslovné pevné spojení na straně anglické. Je zřejmé, že při takovém omezení systém nemůže správně překládat např. slovesné vazby ap.

Naším cílem je tedy obohatit strukturu slovníku, tak aby vyhovovala potřebám (zejména strukturálního) strojového překladu, a připravit slovník pro překlad z češtiny do angličtiny.

V sekci 2 popisujeme dostupné zdroje dat pro náš úkol a ilustrujeme některé jejich problémy. Sekce 3 se podrobně věnuje našim krokům při čištění slovníků a doplňování informací, které jsou nezbytné pro nasazení v systému automatického překladu. Samostatnou sekci 4 věnujeme našemu novému algoritmu určenému pro automatický sběr překladů slovesných rámců.

* Práce na tomto projektu je podporována granty MŠMT LC536, GAČR 201/05/H014 a GAUK 351/2005.

[1] V celém textu budeme pojmem HESLO označovat překladový pár jednoho českého a jednoho anglického výrazu.

2 Dostupné strojově čitelné slovníky a jejich problémy

Při budování slovníku pro strojový překlad je jistě vhodné v maximální možné míře využít existující elektronicky dostupné zdroje. V našem případě se jedná zejména o slovníky WinGED[2] a [4] a rovněž nelze opomenout EuroWordNet[3].

Dostupné elektronické slovníky s výjimkou EuroWordNetu jsou budovány zásadně pro lidského uživatele a z tohoto důvodu je nelze přímo použít při automatickém zpracování. Problémy zejména působí nekonzistentní anotace informací o slovníkových heslech a chybějící údaje, které jsou nutné pro jednotlivé kroky strojového překladu.

2.1 Nekonzistentní anotace metainformací

Dostupné slovníky mají tvar dvousloupcového textu, kde na levé straně je výraz český a na pravé straně výraz anglický. Samotný text na jednotlivých řádcích však v řadě případů obsahuje nejen čistý základní tvar překládaného hesla, ale též dodatečné informace *o* výrazu, např.:

- stručnou poznámku o stylu či dialektu pro dané konkrétní heslo: *am.*, *mat.*
- oddělovač vyznačující více variant výrazu, z nichž lze při překladu užít vždy jen jednu: *be liable/subordinate to*
- konstrukce označující nepovinnou část výrazu: *(auto)stop*, *(na)líčení tváře*
- poznámky popisující tvaroslovné vlastnosti výrazu: *průdušky pl.* pro plurál, *pyramida f* pro ženský rod
- poznámky charakterizující syntaktické vlastnosti výrazu, zejména syntaktické rámce sloves ap.: *accession to = vstoupení do* nebo *adjudge sb. to be guilty = uznat vinným koho*

Tyto cenné informace bohužel nejsou v heslech vyznačovány žádnou standardizovanou metodou. Pro lidského uživatele není problém takové informaci v heslu porozumět a zejména ji z hesla vyjmout, automatický systém však nemá žádný podklad pro to, aby např. v překládané větě nehledal text *pl.*, když chce použít heslo *průdušky pl.*, nebo aby při generování naopak nepřidal do textu tuto „netextovou“ část hesla.

2.2 Chybějící morfologická informace

Pro úspěšné použití slovníkového hesla v automatickém systému je nutné poskytnout počítači informaci o slovním druhu jednotlivých slov výrazu a případně o dalších povinných morfologických rysech hesel (např. povinný genitiv, povinnou negaci, povinné množné číslo ap.). Až na velmi řídké výjimky však slovníková

[2] `http://www.rewin.cz/`

[3] EuroWordNet ([5]) je možné použít jako překladový slovník, obsahuje však celkem pouze cca 10 000 podstatných jmen a cca 3 000 sloves (bez informace o vazbách, s výjimkou zvratné částice *se,si*), což je o řád méně než ostatní slovníky.

hesla tyto informace neobsahují a s ohledem na vysokou míru tvaroslovné nejednoznačnosti češtiny v řadě případů není morfologická informace pro počítač jasná. Počítač pak nemůže rozhodnout, zda je překladové heslo oprávněn použít, nebo zda musí hledat jinou variantu, protože slovní druhy neodpovídají. Tabulka 1 ilustruje tento problém ukázkou českých výrazů ze slovníkových hesel.

Substantivum a substantivum/adjektivum	**Správná interpretace**
husa divoká	substantivum adjektivum
kniha účetní†	substantivum adjektivum
napětí dovolené†	substantivum adjektivum
chyba měření	substantivum substantivum
plán prací†	substantivum substantivum
rozsah měření	substantivum substantivum
Číslovka/sloveso a substantivum	**Správná interpretace**
tři prdele†	číslovka substantivum
pět švestek	číslovka substantivum
pět chválu	sloveso substantivum

Tabulka 1. Příklady nejednoznačnosti slovních druhů v překladových slovnících. Výrazy označené † umožňují i druhou interpretaci, typicky poněkud směšnou.

2.3 Chybějící syntaktická informace

Tvůrci překladových slovníků typicky neuvádějí žádné informace o syntaktické struktuře překladových hesel a ani nedodržují žádná pevná pravidla, která by umožnila tuto informaci automaticky u hesel doplnit. (Např. řídící slovo výrazu by mohlo být uváděno zásadně na prvním místě, ne vždy se tak však děje.)

Syntaktická informace (vnitřní syntaktická struktura hesla) je však pro strukturální překladový systém velmi podstatná. Systém totiž typicky provede větný rozbor dříve, než vyhledává překladové ekvivalenty. Víceslovný výraz je pak typicky reprezentován více uzly ve stromové struktuře věty a pokud slovníková hesla tuto strukturu nepopisují, je obtížné v této fázi identifikovat, zda se heslo ve větě skutečně vyskytlo.

3 Postup čištění a doplňování potřebných informací

V této sekci stručně popíšeme provedený postup čištění a doplnění potřebných informací do dostupných elektronických slovníků WinGED a GNU-FDL.

Ve všech případech ručního čištění jsme zásadně postupovali po blocích, metodou velmi podobnou klasickému „rozděl a panuj“, neboť tento postup vede k nejvyšší dosažitelné vnitřní konzistenci a současně je i nejrychlejší. Před samotnou ruční korekcí nejprve shromáždíme k sobě všechna slovníková hesla s daným nejednoznačným nebo nepřesným rysem a pak pracujeme jen na této podmnožině. Můžeme samozřejmě tuto podmnožinu opět rozdělit podle dodatečných kritérií a v řadě případů je možné dopracovat se k podmnožině dostatečně specifické na to, aby bylo možné korekci provést na všech heslech najednou.

3.1 Identifikace metainformací

V prvním kroku jsme identifikovali všechna častá slova (dle výskytu ve slovníku), která typicky kódují určitou informaci o hesle, např. *pl.*, *am.*, *někoho*, *sb.*, *someone* ap. Rovněž jsme prověřili všechna hesla, která končí potenciální předložkou. Ve všech těchto případech bylo na základě překladového ekvivalentu možné rozhodnout, zda dané slovo vyjadřuje morfologickou informaci, informaci o stylu, „slot“ (tj. zástupný symbol pro další rozvití), nebo zda jde o pevnou součást hesla. Např. v hesle *mít o sobě vysoké mínění = think something of oneself* označují slot pouze slova *sobě* a *oneself*, ne však slovo *something*.

V této fázi jsme rovněž roznásobili všechna hesla, která kódovala více překladových variant současně.

3.2 Zjednoznačnění slovních druhů

Ke zjednoznačnění slovních druhů jsme zatím přistoupili pouze na české straně. Slovní druhy na anglické straně bude možné částečně odvodit ze slovního druhu na české straně a ve sporných případech zvolíme analogickou techniku jako pro češtinu.

Slovníková hesla jsme nejprve analyzovali pomocí morfologického analyzátoru [6] a víceznačné případy jsme roztřídili ručně.[4] Postupovali jsme přitom opět v blocích podle konkrétního typu morfologické nejednoznačnosti, jak však ilustruje tabulka 1, není možné úplně se vyhnout čtení slovníkových hesel.

3.3 Doplnění morfologických omezení

Morfologická omezení hesel popisují, které hodnoty morfologických rysů jsou pro která slova v hesle povinná a která volná, případně s dodatečným požadavkem shody daného rysu mezi více slovy hesla. Morfologická omezení pak budou využita při analýze textu ke kontrole, zda výskyt daných slov skutečně vyjadřuje slovníkové heslo, nebo jde o náhodnou kolokaci. S ohledem na náš směr překladu z češtiny do angličtiny jsme se zatím soustředili na identifikaci morfologických omezení pro českou stranu hesel.

Nutná morfologická omezení generujeme automaticky na základě příkladů dostupných v korpusu[5]. Pro každé slovníkové heslo automaticky vyhledáme v korpusu věty, které obsahují všechna lemata (základní tvary slov) daného hesla v blízkém sousedství. Mezi slovy přitom mohou být vložena jiná slova a na pořadí slov nezáleží. Nalezeným výskytům přiřazujeme váhu tak, aby vyšší váhu měly výskyty bez vložených slov a případně se souvislým závislostním grafem.[6] Seznam vážených výskytů je prohledán a je ověřeno, zda dané výskyty splňují

[4] Je zřejmé, že pro tento účel není vhodné použít automatickou desambiguaci (tagger), neboť slovníkovým heslům chybí větný kontext nezbytný pro správný chod taggeru.

[5] Konkrétně používáme Český národní korpus [7].

[6] V případě, že použitý korpus obsahuje již i větnou strukturu, jako např. PDT [8] nebo PCEDT [9], lze ji přímo použít. V případě ostatních korpusů spoléháme na automatickou syntaktickou analýzu, jak ji generuje parser [10] adaptovaný pro češtinu.

některé z předem definovaných typů omezení. Dosud jsou v našem repertoáru předem definovaných omezení omezení unární (např. test typu „pád je akuzativ“ nebo „číslo je jednotné“) a binární („pád u prvního a druhého slova se shodují“ ap.). Pokud alespoň 75 % celkové váhy tvoří výskyty splňující dané omezení, pokládáme omezení za povinně platné pro dané heslo. (Zbylých 25 % přisuzujeme náhodným souvýskytům slov.)

Přestože je popsaný algoritmus velmi jednoduchý a přestože použité korpusy obsahují často chybnou (automaticky generovanou) morfologickou a závislostní informaci, jsme s výsledky automatické extrakce platných omezení spokojeni. Většina z hesel, u nichž se podařilo v korpusu najít alespoň 10 výskytů, získává správnou sadu omezení. Příklad takto získaných omezení je uveden v tabulce 2. Jako problematická se jeví zejména hesla složená z velmi běžných slov. Běžná slova se totiž poměrně často vyskytnou blízko sebe i v případě, že netvoří dohromady hledané heslo. Značná část pozorovaných výskytů pak neodráží omezení daného hesla a náš práh 75 % tato omezení nepřijme. Např. pro spojení *bohatý člověk* nebylo možné automaticky potvrdit požadavek shody čísla a rodu, shodu pádu se potvrdit podařilo.

Český výraz	**Unární omezení**	**Binární omezení**
za nízkou cenu	RR–4 AAF** NNF*4	pčr:2=3
v jediném dnu	RR–6 AA*S* NNIS*	číslo:2=3
v jiném směru	RR–6 AAI** NNIS6	pád:1=3 číslo:2=3 rod:2=3
v jiném stavu	RR–6 AA*S* NNIS*	
v jistém smyslu	RR–6 AAIS6 NNIS6	pád:1=2 pád:1=3 pčr:2=3
bohatý člověk	AA*** NNM**	číslo:1=2
získané informace	AAFP* NNFP*	pčr:1=2
zkušební provoz	AAIS* NNIS*	pčr:1=2
první světová válka	CrFS* AAFS* NNFS*	pčr:1=2 pčr:1=3 pčr:2=3

Unární omezení vyjadřujeme českou poziční morfologickou značkou (pro reference viz PDT, [8]), kde znak „*“ označuje pozici, jejíž hodnota závisí na kontextu (tj. pozici, na niž se žádné unární omezení nevztahuje).
Binární omezení určují, které hodnoty musí být mezi jednotlivými slovy výrazu sdíleny. Např. „pčr:2=3“ vyjadřuje, že pád, číslo i rod druhého a třetího slova se musí shodovat.

Tabulka 2. Příklady automaticky extrahovaných morfologických omezení.

3.4 Doplnění syntaktických omezení

Syntaktické informace (závislostní vztahy mezi slovy výrazu) jsme zatím doplnili pouze do české části hesel. Ve většině případů bylo možné strukturu stanovit jednotně pro drtivou většinu hesel se shodnou posloupností slovních druhů jednotlivých slov (např. u všech hesel obsahujících adjektivum následované podstatným jménem je řídícím slovem substantivum). Pro zbývající hesla s velmi různorodými posloupnostmi slovních druhů je ruční metoda příliš náročná a dáváme přednost hledání typické struktury korpusu, podobně jako popisujeme v sekci 3.3.

4 Automatické získávání překladových slovesných rámců

Termínem PŘEKLADOVÝ SLOVESNÝ RÁMEC budeme označovat specifický typ překladového hesla.[7] Kromě českého slovesa a jeho anglického protějšku překladový rámec obsahuje též seznam dvojic popisující české a odpovídající anglické formy slovesných doplnění. Příkladem může být heslo: *dělit=divide na+akuzativ=into*.

Slovníková hesla po ručním očištění (viz sekce 3.1) sice v řadě případů obsahují u slovesných hesel informaci o vazbách a jejich překladech, specifické zaměření ekonomických textů však způsobuje, že řada užívaných vazeb ve slovníku vůbec není zahrnuta. Z tohoto důvodu se věnujeme i vývoji automatické metody, která z paralelního korpusu překladová slovesná hesla získá.

Extrakcí (povrchových) slovesných rámců z různých typů korpusů se zabývá též [12] a další práce, na něž odkazuje. Náš cíl je však odlišný. Zatímco citovaní autoři usilují o zodpovězení otázky, která z pozorovaných slovesných doplnění jsou nedílnou součástí slovesného rámce (complements), a která jsou na slovese nezávislá (adjuncts), my potřebujeme vědět: „Jakou vazbu je třeba užít na anglické straně, jestliže na české straně bylo doplnění daného druhu“.

4.1 Metoda získávání překladových rámců

Pro získávání překladových rámců užíváme paralelní závislostní korpus [9], který jsme navíc obohatili automatickou metodou (GIZA++, [13]) o párování jednotlivých slov k sobě. Obecně je možné naši metodu nasadit na jakýkoli paralelní text po doplnění (automatické) závislostní analýzy v obou jazycích a spárování textů po slovech.

V prvním kroku získáváme POZOROVANÉ PŘEKLADOVÉ RÁMCE takto: zaměříme se na všechny výskyty českých sloves, k nimž je párováním jednoznačně připojeno sloveso anglické. Pro všechna doplnění českého slovesa pak zjistíme odpovídající doplnění anglického slovesa: sledujeme, ke kterým uzlům v anglické větě vedou párovací hrany z celého podstromu studovaného českého doplnění. Jako odpovídající anglické doplnění zvolíme to doplnění, do jehož podstromu vede nejvíce sledovaných párovacích hran. V sekci 4.3 popisujeme problémy, na něž tento jednoduchý přístup naráží.

Ve druhém kroku získané překladové rámce automaticky očišťujeme, a to několika různými způsoby:

- Bez očištění (označení `raw`): použijeme přímo pozorované překladové rámce.
- Odstranění málo četných typů doplnění (`freq`): ze všech rámců odstraníme všechny překladové páry forem doplnění, které nebyly pozorovány (napříč všemi slovesy) dostatečně často.

[7] Je třeba upozornit na odlišnost naší definice překladového rámce od valenčních rámců zachycovaných ve valenčním slovníku češtiny VALLEX (nejnověji viz. [11]) nebo PDT-VALLEX (odkaz viz VALLEX), které se soustředí na zachycení hloubkových syntaktických rysů.

– Odstranění nespolehlivých párování (`giza`): na základě dodatečné informace ze systému GIZA++ odstraníme všechny rámce pozorované ve větách, kde si GIZA++ nebyla výsledným párováním slov dostatečně jista.
– Pouze „velmi jednoduché“ české věty (`vjv`): Použitím algoritmu popsaném v práci [14] odstraníme všechny rámce pozorované ve složitých českých větách. Složité české věty představují pro automatickou syntaktickou analýzu větší problém a [14] ukazuje, jak je možné kvalitu pozorovaných českých rámců zlepšit, pokud se omezíme na jednodušší větné konstrukce.

Třetí volitelný krok navazuje na očištěné rámce a používá statistické metody nebo metody strojového učení k tomu, aby dále zjednodušil množinu očištěných rámců a dospěl k finální verzi překladových rámců určené pro nasazení v překladovém systému. Dosud jsme vyzkoušeli jednu z možných metod, algoritmus Apriori ([15])[8], jinou možností je využít některou z metod navrhovaných v práci [12] pro identifikaci těch doplnění, která jsou pro dané sloveso typická. V překladových rámcích pro dané sloveso bychom pak uváděli jen překlady typických doplnění a překlady ostatních doplnění bychom překládali pro všechna slovesa společně.

4.2 Vyhodnocení

Abychom vyhodnotili, který z navržených postupů získávání a čištění překladových rámců poslouží v překladovém systému pravděpodobně nejlépe, připravili jsme ruční korekcí sadu 140 testovacích vět. Testovací věty obsahují celkem 400 výskytů 200 různých sloves a tato slovesa mají dohromady celkem 1005 doplnění.

Vzhledem k tomu, že celý systém překladu ještě není zprovozněn, implementovali jsme tři jednoduché algoritmy, které mají za úkol pro dané sloveso a jeho pozorovaná česká doplnění odpovědět, jaké formy budou mít odpovídající doplnění anglická.[9] Všechny tři se opírají o očištěný seznam rámců (pro stručnost říkejme prostě „slovník“), liší se však v tom, jakým způsobem konstruují pro pozorovaný český rámec anglickou odpověď:

– Algoritmus A – překlad slot po slotu bez ohledu na sloveso. Tento hladový algoritmus napřed předzpracuje slovník a pro každý český slot si poznamená,

[8] Algoritmus Apriori a další podobné algoritmy byly navrženy pro podporu prodeje: rozborem množiny uskutečněných obchodních transakcí, kde každá transakce představuje množinu zakoupených artiklů, algoritmus identifikuje časté vztahy mezi vybranými artikly. (Kdo si koupil chléb, koupil si často i máslo.) Alternativně může být výstup tohoto algoritmu vyjádřen jako typické podmnožiny uskutečněných transakcí. Analogie s naší situací je zřejmá: jedné transakci odpovídá jeden výskyt slovesa v paralelním korpusu a každá překladová dvojice český slot–anglický slot pak představuje jeden artikl. Algoritmus identifikuje, které podmnožiny překladových dvojic jsou pro dané sloveso typické.

[9] Je zřejmé, že české věty je možné přeložit do angličtiny více různými způsoby, volit různá slovesa s rozdílnou sadou forem doplnění ap. Pro jednoduchost však tento problém necháváme zatím stranou a opíráme se o jediný referenční překlad.

jakou formu má nejčastěji příslušný anglický slot. Při samotném běhu pracuje slot po slotu a zásadně volí tento vítězný ekvivalent.

– Algoritmus B – překlad slot po slotu s ohledem na sloveso. Podobně jako A předzpracuje slovník a překlad volí slot po slotu, vítězný překlad však volí jen ze slotů pozorovaných u daného českého slovesa. V případě, že ve slovníku u daného slovesa nějaká česká forma zcela chybí, můžeme algoritmus B kombinovat s A a zkusit navrhnout překlad slotu bez ohledu na slovesa. U jiného slovesa totiž daná česká forma viděna být mohla a možná i její překlad dobře poslouží zde. Tento kombinovaný algoritmus označme „BA“.
– Algoritmus C – překlad podle vítězného rámce. Pro vstupní pozorovaný rámec algoritmus nejprve najde mezi slovníkovými rámci daného slovesa takový rámec, který je pozorovanému rámci nejpodobnější (ev. více rámců se stejnou mírou podobnosti). Pak překládá slot po slotu, ale překladové ekvivalenty volí jen na základě vybraných rámců. V případě, kdy se nepodaří nějaký slot přeložit, protože vítězný rámec překlad daného slotu vůbec neobsahoval, je opět možné překlad navrhnout některou z jednodušších metod. Konkrétně používáme kombinace CBA a CB.

Pročištění	**Apriori**	**Algoritmus**	**F-score**	**Přesnost**	**Pokrytí**
`giza`	Apriori	A	68.4	52.9	96.7 ← nejlepší F-score
`giza`	Apriori	CBA	66.4	50.5	96.7
`giza`	Apriori	BA	66.1	50.2	96.7
`giza`	Bez Apriori	BA	66.1	49.8	98.0
`raw`	Bez Apriori	A	58.2	41.3	98.9 ← baseline
`freq`	Apriori	BA	56.5	52.9	60.6 ← nejlepší přesnost
`vjv`	Apriori	BA	55.9	41.4	86.3 ← nejlepší z `vjv`
`vjv`	Apriori	C	30.0	33.5	27.2 ← nejhorší výsledek

Tabulka 3. Přesnost a pokrytí algoritmů pro překlad forem slovesných doplnění.

Tabulka 3 shrnuje základní ukazatele o kvalitě predikce pro vybrané kombinace předzpracování a použité metody. Přesnost je počet správně zvolených překladů z celkového počtu odevzdaných překladů, pokrytí je počet slotů, k nimž algoritmus navrhl nějakou odpověď, z celkového počtu slotů. Výsledky jsou uspořádány podle tzv. F-score, které představuje geometrický průměr přesnosti a pokrytí.

Nejlepšího výsledku jsme dosáhli při použití nejprostší metody A opírající se o data po odstranění vět s nespolehlivým párováním a po použití algoritmu Apriori pro zjednodušení pozorovaných překladových rámců. S mírným odstupem následují kombinace CBA a BA na stejně předzpracovaných rámcích. Již samotné odstranění nespolehlivě spárovaných vět (bez nasazení metody Apriori) stačí algoritmu BA k dosažení velmi podobných výsledků. Tyto nejlepší výsledky představují zlepšení o 10 procentních bodů v přesnosti za cenu nepatrného po-

klesu v pokrytí proti metodě, kterou lze považovat za zcela základní (baseline): metoda A na nijak nefiltrované pozorované rámce.

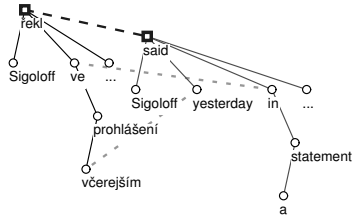
Jak je z výsledků zřejmé, největší vliv na kvalitu predikce překladové formy má kvalitní párování slov. Čištění pomocí Apriori má rovněž svůj význam, z implementovaných metod A, B a C však nejlépe funguje nejjednodušší metoda A. Kvalitě predikce naopak nejvíce ublížilo, když jsme překladové rámce získávali pouze z vět, kde česká věta je „velmi jednoduchá“, `vjv`. Vysvětlení pro tento neúspěch spočívá pravděpodobně v tom, že popsaný algoritmus kombinující dva závislostní stromy pomocí párovacích hran nenajde v (typicky krátkých) velmi jednoduchých větách dostatek podkladů pro rozumné přiřazení slotů k sobě.

4.3 Pomíjené problémy překladu syntaktického okolí sloves

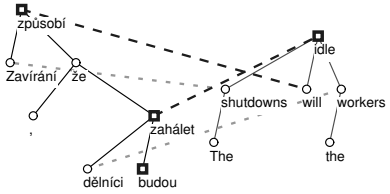
Při našich prvních experimentech pro jednoduchost přehlížíme významné problémy způsobené syntaktickou odlišností češtiny angličtiny, jak ji pozorujeme v korpusu ekonomických textů (PCEDT). V dalším výzkumu se však na tyto problémy zaměříme podrobněji a zohledníme je v našem algoritmu automatizované extrakce.

Posun doplnění. Jak ukazuje obrázek 1, v některých případech mírně odlišný překlad způsobí nemožnost přesného párování jednotlivých doplnění na sebe. Vhodným řešením by bylo takové případy identifikovat a do slovníku zařadit jen ty sloty, které si odpovídají.

Záměna řídících uzlů (Head Switching). U některých konstrukcí dochází při přechodu od českého závislostního stromu k anglickému k záměně řídících uzlů. (Např. ve větě *Nákupní firmy se* mohou *stát ziskovými. = Malls can* become *profitable.* je jako kořen v češtině použito způsobové sloveso, v angličtině se však užívá typicky sloveso plnovýznamové.) Je třeba pečlivě odlišovat situace, kdy je záměna způsobena pouze jinou dohodou o anotaci a kdy jde o závažnější syntaktickou odchylku (viz obrázek 2). Dobrým rámcem pro toto odlišení by mohla být závislostní redukční analýza, [16].



Obrázek 1. Sigoloff řekl ve včerejším prohlášení ... Sigoloff said yesterday in a statement ...



Obrázek 2. Zavírání způsobí , že dělníci budou zahálet ... The shutdowns will idle the workers ...

5 Shrnutí a další plány

Popsali jsme postup prací, které probíhají na přípravě česko-anglického slovníku pro strojový překlad. Dostupné strojové slovníky kombinací ručních a automatických metod obohacujeme o potřebné morfologické a syntaktické údaje.

V dalším výzkumu se kromě dalšího zlepšování metod automatické identifikace slovesných rámců zaměříme též na možnosti formálního zachycení pozorovaných syntaktických odlišností ve slovníku.

Reference

1. Dorr, B.J., Jordan, P.W., Benoit, J.W.: A survey of current paradigms in machine translation. Technical Report CS-TR-3961 (1998)
2. Čmejrek, M., Cuřín, J., Havelka, J.: Czech-English Dependency-based Machine Translation. In: EACL 2003 Proceedings of the Conference, Association for Computational Linguistics (2003) 83–90 MSM113200006, LN00A063.
3. 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, Association for Computational Linguistics (1987) 113–117
4. Svoboda, M.: GNU/FDL English-Czech Dictionary (2001) `http://slovnik.zcu.cz/`.
5. Pala, K., Smrž, P.: Building Czech Wordnet. ROMANIAN JOURNAL OF INFORMATION, SCIENCE AND TECHNOLOGY **7** (2004) 79–88
6. Hajič, J.: Disambiguation of Rich Inflection - Computational Morphology of Czech. Volume I. Prague Karolinum, Charles University Press (2001) 334 pp.
7. Kocek, J., Kopřivová, M., Kučera, K., eds.: Český národní korpus - úvod a příručka uživatele. FF UK - ÚČNK, Praha (2000)
8. Böhmová, A., Hajič, J., Hajičová, E., Hladká, B.: The Prague Dependency Treebank: Three-Level Annotation Scenario. In Abeillé, A., ed.: Treebanks: Building and Using Syntactically Annotated Corpora. Kluwer Academic Publishers (2001)
9. Čmejrek, M., Cuřín, J., Havelka, J., Hajič, J., Kuboň, V.: Prague Czech-English Dependecy Treebank: Syntactically Annotated Resources for Machine Translation. In: Proceedings of LREC 2004, Lisbon (2004)
10. Charniak, E.: A Maximum-Entropy-Inspired Parser. In: Proceedings of NAACL-2000, Seattle, Washington, USA (2000) 132–139
11. Lopatková, M., Bojar, O., Semecký, J., Benešová, V., Žabokrtský, Z.: Valency Lexicon of Czech Verbs VALLEX: Recent Experiments with Frame Disambiguation. In: Proceedings of TSD 2005. (2005) (in press).
12. Zeman, D., Sarkar, A.: Learning Verb Subcategorization from Corpora: Counting Frame Subsets. In: Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000), Athens, Greece, ELRA (2000)
13. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Comput. Linguist. **29** (2003) 19–51
14. Bojar, O.: Towards Automatic Extraction of Verb Frames. Prague Bulletin of Mathematical Linguistics (2003) 101–120
15. Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large databases. In: SIGMOD '93: Proceedings of the 1993 ACM SIGMOD international conference on Management of data, New York, NY, USA, ACM Press (1993) 207–216
16. Lopatková, M., Plátek, M., Kuboň, V.: Závislostní redukční analýza přirozených jazyků. In: Proceedings of ITAT 2004, Košice, Slovakia, University of P. J. Šafařík (2005) 165–174