Využití lingvistických dat ve strojovém překladu

Ondřej Bojar
bojar@ufal.mff.cuni.cz
Ústav formální a aplikované lingvistiky
Matematicko-fyzikální fakulta
Univerzita Karlova v Praze

Obsah prezentace

Širší motivace: vztah lingvistické teorie, dat a aplikací.

- Teorie předpovídá, které rysy jazykových dat jsou podstatné, jak data „komprimovat“ do slovníků ap.
- Podaří se nám tyto rady využít pro lingvistické aplikace?

Omezíme se na:

- jednu teorii: Funkční generativní popis (FGD, Sgall *et al.* (1986)),
- jeden typ dat: valenční slovníky,
- jednu aplikaci: strojový překlad z angličtiny do češtiny.

Struktura disertační práce i této prezentace:

- Extrakce valenčních rámců.
- Překlad se strukturním transferem.
- Frázový překlad o více faktorech.
- Shrnutí přínosu práce.

Část první: Valenční rámce z korpusu

- Valenční rámce slovesa formálně zachycují schopnost vázat další větné členy.
- Valenční slovník považován za významnou komponentu aplikací:
  - Zjednoznačňuje analýzy, pomáhá volit formu doplnění při generování.
- Ručně vytvářený VALLEX 1.5 pokrývá 90 % výskytů sloves, zbývajících 10 % ale představuje téměř 30 tisíc slovesných lemat.
  - V mezidobí byl VALLEX rozšířen, aktuální verze 2.5 je větší, ale má mírně jinou strukturu.

⇒ Lze ušetřit lexikografům práci a navrhovat hesla automaticky dle korpusu?

Tradiční **přesnost** (precision) a **pokrytí** (recall) na úrovni celých rámců příliš hrubé.

Nová metrika: **očekávaná úspora** (expected saving, ES) práce lexikografa:

= úspora v editaci slovníkových hesel díky rámcům navrženým systémem.

Navržené metody extrakce

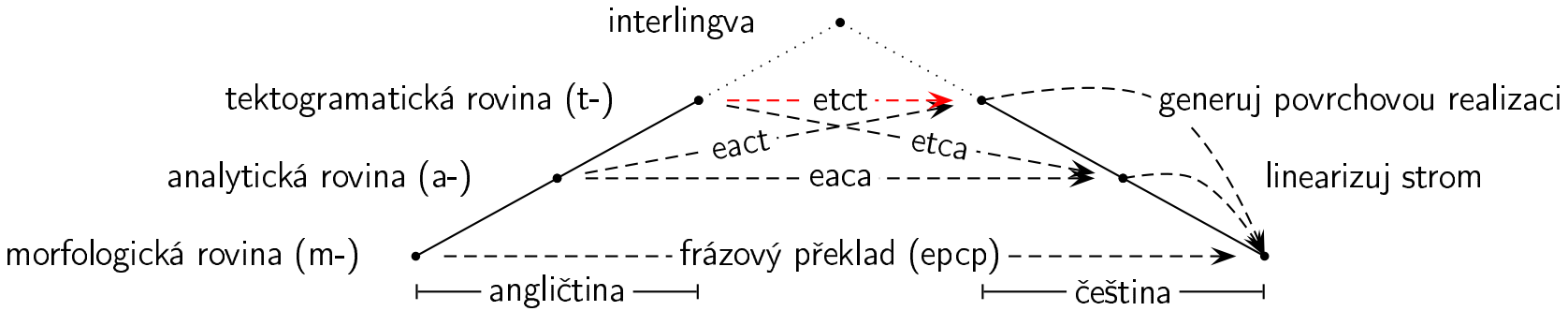
Přímé metody:

- Vstup: korpus s anotovanými slovesnými rámci (např. VALEVAL, PDT).
- Novým slovesům přiřaď rámce na základě podobnosti povrchových konfigurací se známými slovesy.
- Tři varianty podobnosti: WFD, Decomp, DSD.
- Úspora je na úrovni baseline: $\sim 40\,\%$ editačních úprav.

Přes sémantickou třídu:

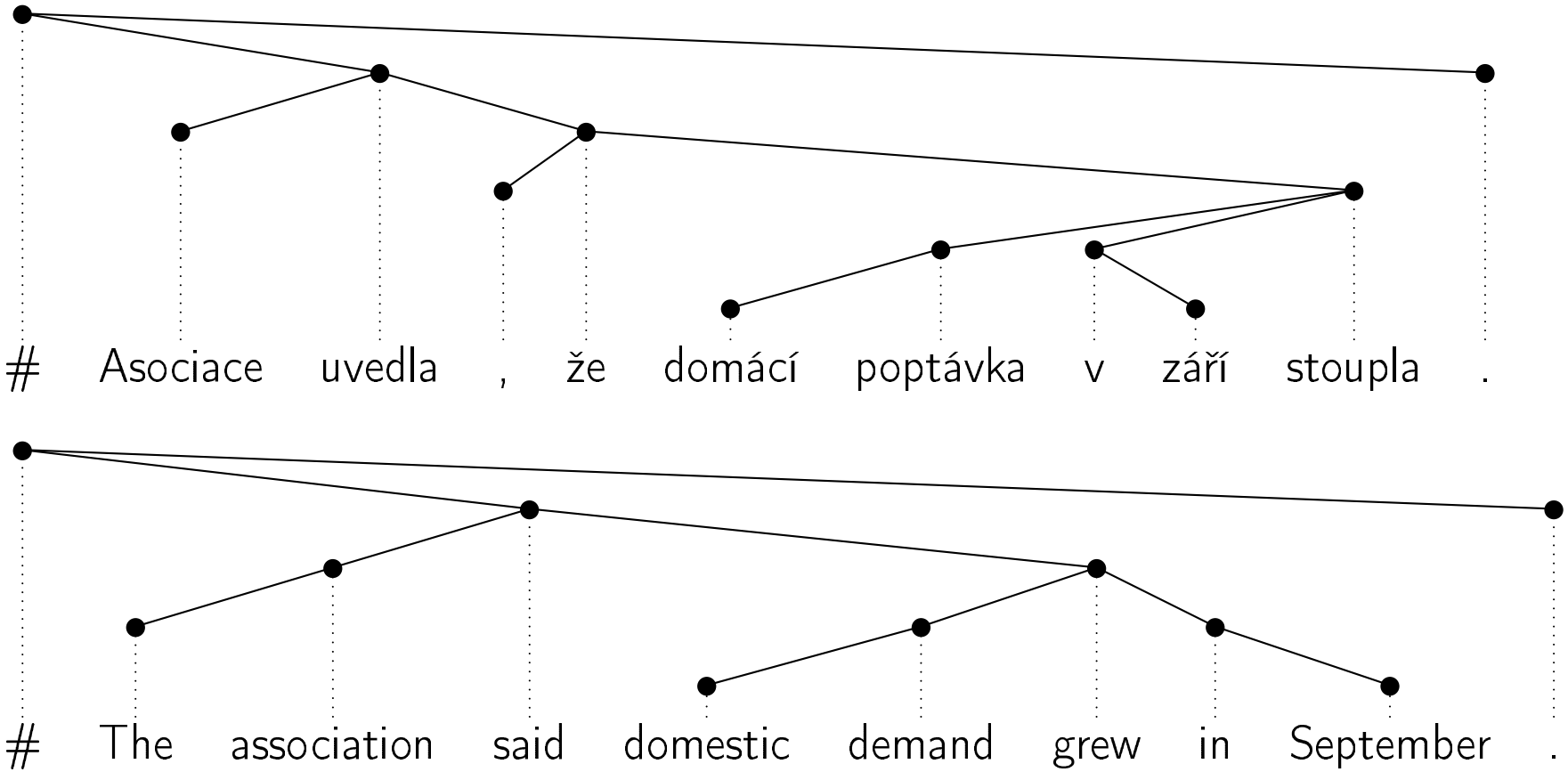
- Slovesa mluvení typicky rozvita: *nom*, *dat/acc*, *subord(že, aby, ...)*
- Nová slovesa s dostatečným počtem výskytů v této konfiguraci prohlásíme za slovesa mluvení, přisoudíme jim rámec ACT-PAT-ADDR.
- Úspora: 38,0±0,2 místo baseline 37,6±0,2 (navržen 1×ACT-PAT).
- Podstatnější je správně poznat počet rámců: úspora 39,1±0,1 při 2×ACT-PAT.

Část druhá: Strukturní překlad

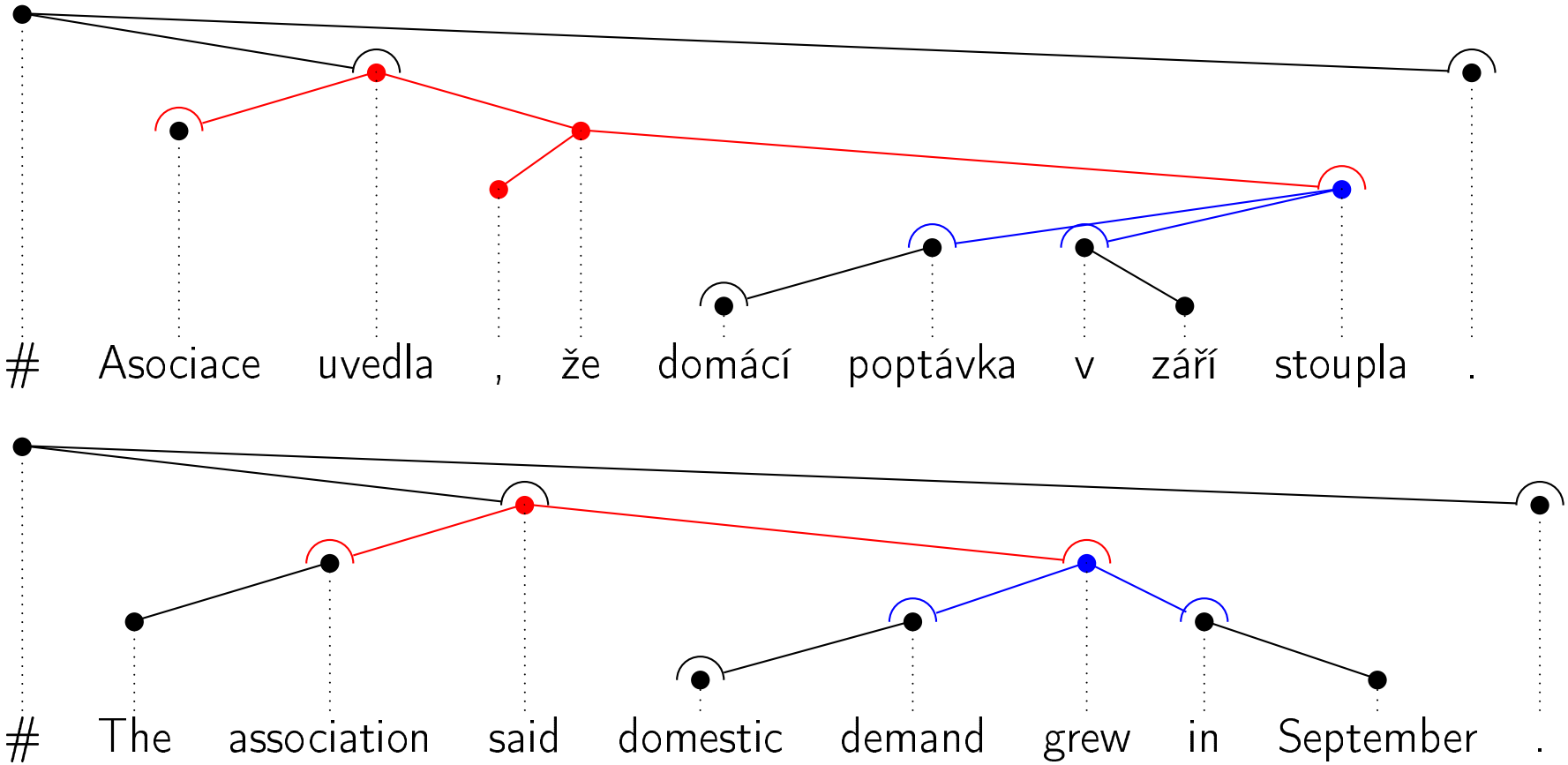


- Vstupní i cílová věta reprezentována jako závislostní strom.
- Hlubší roviny jsou si podobnější.
- Formalismus **Synchronous Tree Substitution Grammar** (Čmejrek, 2006) pro převod stromu na strom.
- Implementovaný dekodér lze užít na kterékoli rovině i napříč rovinami.

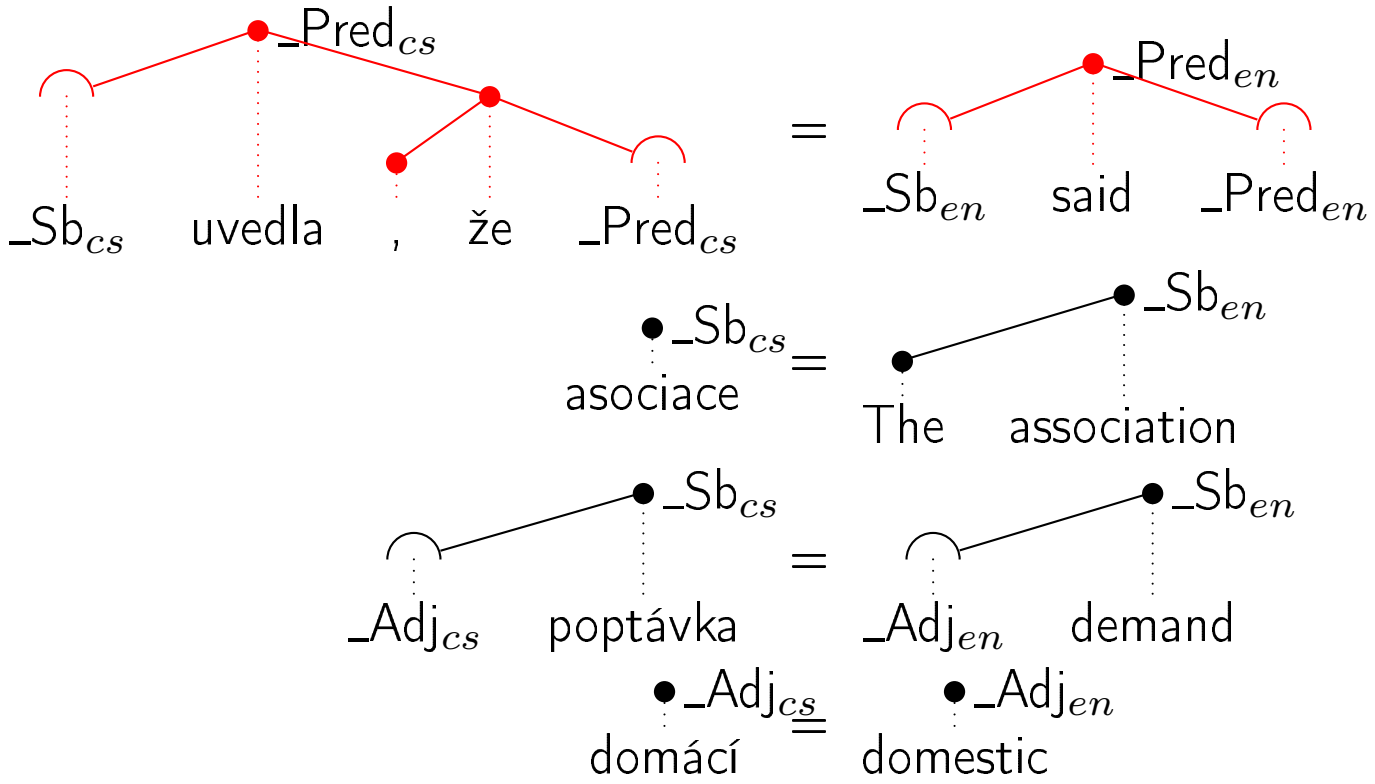
Ilustrace: Dvojici analytických stromů...



...rozložíme na stromečky...

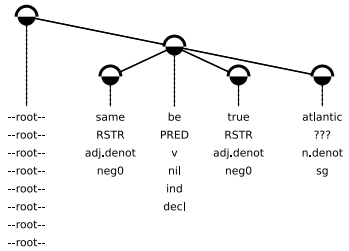


...a sebereme slovník překladů stromečků.



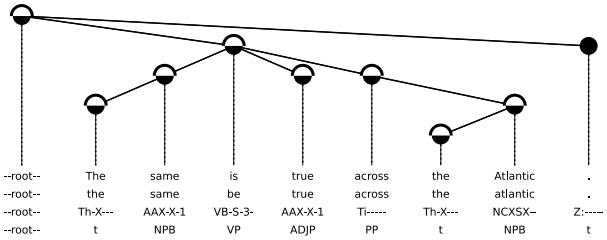
Dosažené skóre (BLEU)

etct



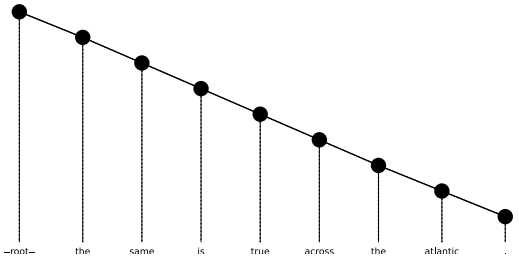
BLEU: 5.61±0.50

eaca



BLEU: 8.75±0.61

epcp



BLEU: 10.90±0.63

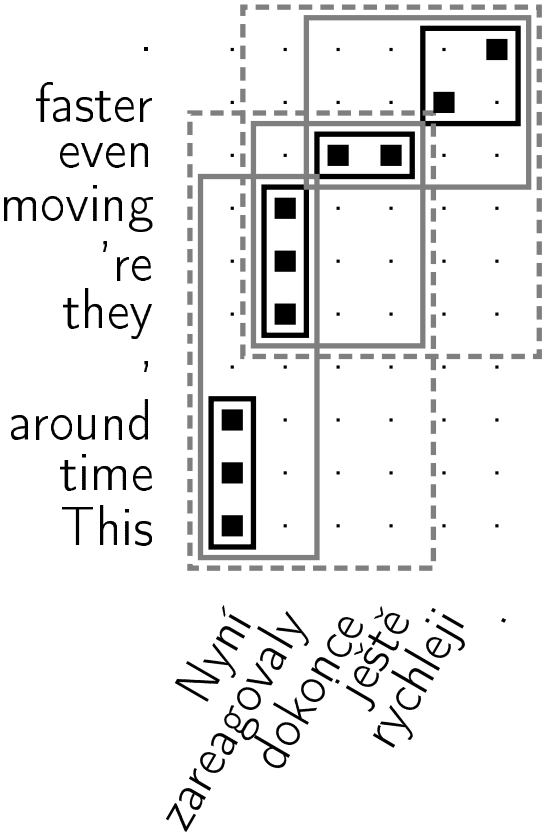
Diskuse: proč mi t-rovina nepomáhá

- **Kumulace chyb** každého kroku analýzy:
  - Např. $93\,\% * 85\,\% * 93\,\% * 92\,\% = 67\,\%$.
- **Výrazná ztráta dat** kvůli neparalelním strukturám:
  - Stačí jedna chyba v českém či anglickém parsingu nebo slovním zarovnání $\Rightarrow$ nelze extrahovat.
  - Volný překlad „včerejší jednání“ místo „meeting yesterday“ $\Rightarrow$ nutno extrahovat po slovech.
- **Kombinatorická exploze** při generování výstupních atributů:
  - Nejprve stavím celé stromečky, včetně všech atributů. Spojuji později.
  - Bez kontextu okolních uzlů těžko rozhodovat detailní atributy, kombinací je moc.

  $\Rightarrow$ lexikální varianty často odsunuty ze zásobníku.

  $\Rightarrow$ n-best list je pestrý v nepodstatných atributech.
- **Deterministické generování**:
  - Připraveno pro ruční stromy, přeložené automatické mají spoustu chyb.
  - Nevyužívá n-gramový jazykový model.

Část třetí: Frázový překlad



This time around	=	Nyní
they 're moving	=	zareagovaly
even	=	dokonce ještě
...	=	...
This time around, they 're moving	=	Nyní zareagovaly
even faster	=	dokonce ještě rychleji
...	=	...

Ve frázovém překladu hledáme:

- takovou segmentaci vstupní věty na úseky („fráze“)
- a takové překlady frází

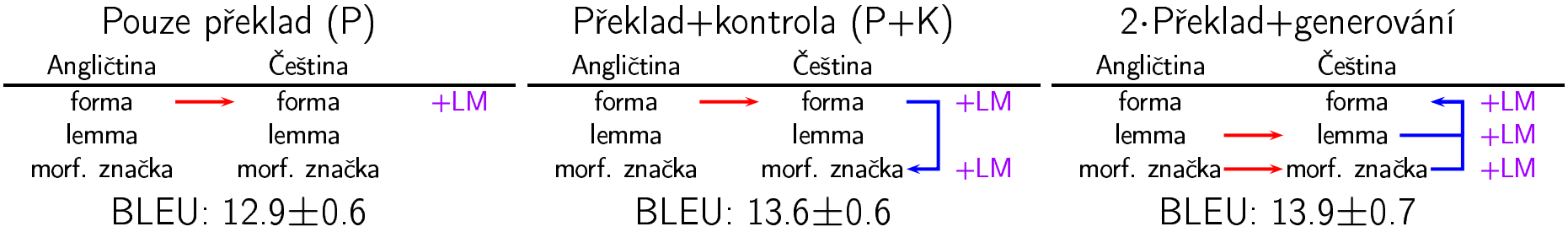
aby byl výstup co nejpravděpodobnější.

Volně šiřitelná implementace: www.statmt.org/moses

Frázový překlad o více faktorech

- Místo prostých slovních forem je každé slovo zapsáno jako n-tice faktorů.
- Možnost zapojit víc jazykových modelů (LM): koherence forem, morf. značek, ...

Provedené experimenty: konfigurace kroků v postupném vyplňování cílových faktorů:



⇒ Explicitní kontrola morfologie pomáhá.

⇒ Složitější scénáře již nejsou výrazně výhodnější.

Srovnání strukturního a frázového překladu

Metoda	Jazykový model	BLEU
Moses P+K, CzEng	4-gramy slov + 7-gramy značek, SYN2006	15.3±0.9
Moses P+K, CzEng	3-gramy slov + 7-gramy značek	14.2±0.7
Moses P+K	3-gramy slov + 7-gramy značek	13.9±0.9
Moses P	3-gramový	12.9±0.6
epcp bez faktorů	3-gramový	10.9±0.6
epcp bez faktorů	žádný	8.7±0.6
nejlepší etct	binodový	5.6±0.5

- Moses lepší než „epcp“:
  - „epcp“ neumožňuje měnit pořadí frází.
  - Moses má řádně implementován MERT (optimalizace vah modelu na BLEU).
- n-gramový LM očividně pomáhá, ale i „epcp“ bez LM > „etct“.

ACL WMT08: Lidská hodnocení překladu

Procento vět, kdy byl daný systém hodnocen stejně nebo lépe než ostatní systémy:

Systém	Commentary (v doméně)	News (mimo doménu)
Můj P+K (cu-bojar)	**71.4 %**	63.4 %
PC Translator	66.3 %	**71.5 %**
TectoMT (cu-tectomt)	48.8 %	49.4 %
Moses, bez faktorů (uedin)	48.6 %	50.2 %

- TectoMT (Žabokrtský, 2008) srovnatelné s prostým frázovým překladem.
  - Transfer na t-rovině jako posloupnost ručně vyladěných kroků, nikoli uniformní prohledávání.

  $\Rightarrow$ Tektogramatická rovina stále dává dobrou naději.
- Frázový překlad o více faktorech (a s více daty) podstatně lepší.
  - Ve známé doméně zvítězil nad komerčním systémem, v obecném překladu horší.

Shrnutí přínosu práce

V disertační práci jsem:

- Navrhl automatickou metriku pro odhad úspory práce lexikografa.
- Navrhl a vyhodnotil několik metod získávání valenčních rámců z korpusu.
- Implementoval systém překladu se strukturním transferem, zapojil jej do kompletní kaskády nástrojů a vyhodnotil jej v několika konfiguracích.
- Zlepšil slovní zarovnání vhodným předzpracováním (Bojar *et al.*, 2006).
- Zlepšil tvaroslovnou soudržnost frázového překladu pomocí více faktorů.

Ve spolupráci s řadou kolegů jsem též připravil a zpřístupnil data:

- Paralelní korpus CzEng (Bojar and Žabokrtský, 2006; Bojar *et al.*, 2008).
- Soubor vět ručně zarovnaný po slovech (Bojar and Prokopová, 2006).
- Zlatý VALEVAL, anotace slovesnými rámci (Bojar *et al.*, 2005).
- Mírně pročištěné česko-anglické překladové slovníky. (Bojar and Prokopová, 2007).

Proč t-rovina není interlingva

...the t-layer includes nodes for entities that were not explicitly expressed in the sentence but the language syntax and lexicon indicate their presence in the described situation. This is one of several reasons that make the t-layer language dependent and not an Interlingua. (str. 19)

Není mi jasné, proč by právě doplnění nevyjádřených větných členů znemožňovalo považovat tektogramatickou rovinu za interlingvu. Prosím o vyjasnění.

- Otázka vůbec je, jak charakterizujeme interlingvu. (Umožňuje inference? Nebo je to jen nejmenší zjemnění významu vyžadované množinou předem daných jazyků?) Obecně o tom, proč t-rovina nestačí pro inference viz obhajoba Václava Nováka.
- Už i samotné lexikální hodnoty (t-lemata) t-uzlů jsou jazykově závislé, viz PDT a PCEDT.
- To, které tektogramatické uzly přidáme pro nevyjádřené větné členy, je dáno valenčním slovníkem: musíme přidat všechny obligatorní.
- Může existovat sloveso+rámec, které nemá ve druhém jazyce ekvivalent se stejnou množinou obligatorních doplnění.

Problémy shody ve frázovém překladu

Obr. 7.2 naznačuje, že výsledek není zaručen, i když se podmět i přísudek vejdou do okénka: firmy vyběhl. A co když větné členy, u kterých se shoda/valence morfologicky projevuje, stojí dál od sebe? Je to o hodně horší?

- Dosud nejúspěšnější konfigurace stále neumí vybrat *libovolnou* platnou formu zvoleného slova (tj. např. „vyběhly“), je omezena na formy spatřené na cílové straně paralelního korpusu. Špatně tedy mohou dopadnout i slova v rámci povoleného n-gramového okénka.
- V brzké době plánuji povolit i formy spatřené ve (výrazně větším) jednojazyčném korpusu.
- Lze uvažovat i o generování formy z lematu a značky pomocí morfologického slovníku místo korpusu. De Gispert *et al.* (2005) to dělá pro španělštinu.
- Otázka shody na delší vzdálenost je stále palčivá. Morfologické jazykové modely zaměřené právě na shodu podmětu s přísudkem a s delším dosahem (byť stále pevně omezeným, např. 11-gram) zatím nepřinesly signifikantní zlepšení.

Připomínky k metrice podobnosti rámců

Nebylo by tedy vzhledem k principu posouvání (shifting), který je zmíněn v kapitole 2.2.2 na straně 19, vhodné upravit zavedenou metriku tak, aby přidání uvedených aktantů (aktor a patient) mělo nulovou váhu? Tedy tato dvě vnitřní doplnění v každém novém rámci předvyplňovat?

- Jsou slovesa, která mají méně než dvě doplnění. Technika předvyplňování by tedy (byť v poměrně málo případech) praktickému nasazení ublížila.
- Ze stejného důvodu by metrika i nadále měla penalizovat nadbytečného či chybějícího aktora či patient.
- Výtka je ale velmi platná: jednotlivé metody by bylo velmi vhodné vylepšit a např. v dodatečné opravné fázi provést shifting navržených funktorů.

Lze spekulovat o tom, zda ušetření editační práce ulehčí i intelektuální námaze lexikografa při verifikaci automaticky navrženého rámce.

- Naprosto souhlasím. Navíc to může podstatně zhoršit kvalitu hesel (člověk odkýve).

Extrakce stromečků dle alignmentu

Nemůže při trénování, t.j. párování treeletů ve dvou paralelních stromových strukturách nastat situace, že treelet v jednom stromě nelze namapovat na žádný treelet ve stromě paralelním? Jak dopadne párování treeletů se slovy, která jsou podle GIZA++ spárována s tzv. nulovým slovem?

Při trénování (viz kap. 3.5) extrahujeme všechny dvojice stromečků, které:

- splňují omezení na maximální velikost,
- splňují podmínku STSG, a
- jsou (v konfigurovatelném smyslu) konzistentní se zarovnáním.

Dosud nejvhodnější konfigurace: požaduj interní uzly jednoho stromečku zarovnané na interní uzly druhého stromečku. Přitom dvojice stromečků musí obsahovat alespoň jednu hranu slovního zarovnání.

⇒ Neparalelní části stromů (GIZA přiřadila NULL) nepřispějí do slovníku samostatně, ale jedině jako součást větších (alignovaných) stromečků.

Počty extrahovaných stromečků

Pro úplnost by bylo zajímavé uvést statistiku velikostí treeletů v trénovacím korpusu (co do počtu vnitřních a hraničních uzlů).

	epcp	eaca	etct
Trénovacích vět (tisíce)	84,1	84,1	84,1
Celkem extrahováno dvojic stromečků (tisíce)	2879,2	1148,7	1125,1

Detail podle počtu interních (i) a hraničních (h) uzlů:

		1i 0h	1i 1h	2i 0h	2i 1h	3i 1h	3i 0h	3i 2h	4i 1h	1i 2h	2i 2h	Ostatní
Čj	epcp	80,9	**882,3**	46,3	677,9	529,5	43,3	-	369,1	-	-	249,9
	eaca	**456,4**	137,3	84,0	81,3	49,4	44,9	31,1	31,0	29,9	29,1	174,3
	etct	**303,9**	94,9	104,3	68,1	54,4	59,6	38,2	46,1	28,1	33,1	294,4
Aj	epcp	81,1	**803,4**	44,5	615,5	506,6	37,4	-	417,1	-	-	373,5
	eaca	**446,8**	126,4	79,3	69,1	51,9	45,7	27,6	39,8	25,7	26,4	209,8
	etct	**362,9**	108,1	83,4	72,8	51,5	44,5	38,4	40,0	33,5	35,6	254,2

- Tučně uvedena velikost „základního slovníku“ (tj. počet slov překládaných jedna k jedné).
- epcp i eaca na slovních formách, etct atomicky všechny atributy $\Rightarrow$ srovnatelné počty „forem“.

Lematizované BLEU

Bylo by zajímavé pro jednotlivé ukázky dopočítat i lematizované BLEU, to by v určitém smyslu kategorizovalo jednotlivé systémy podle úspěšnosti v morfologii a v samotném překladu.

	Commentary (v doméně)			News (mimo doménu)		
	BLEU	lemBLEU	nárůst [%]	BLEU	lemBLEU	nárůst [%]
Google	21.14	26.54	25.54	12.82	18.07	40.95
Moses T+C SYN2006	15.91	23.51	47.77	11.93	16.93	41.91
Moses T+C	14.64	22.50	53.69	9.75	14.60	49.74
PC Translator	8.48	14.13	66.63	8.41	12.88	53.15
TectoMT	9.28	14.02	51.08	6.94	10.99	58.36
etct	4.98	9.59	92.57	3.36	5.96	77.38

- Podle BLEU lematický i běžný výstup kvalitou korelují $\Rightarrow$ nemáme novou kategorizaci.
- Větší jazykový model (SYN2006 i Google) vede k lepším volbám slovních tvarů
  (Nižší možný nárůst, tj. skutečně jsme ubrali kousek prostoru pro zlepšení.)
- Google je podezřele dobře vyladěn na slovní tvary domény Commentary, PC Translator naopak celkem špatně.

Nepodporované korespondence uzlů

...autor v základním modelu neuvažuje několik možných korespondencí uzlů, které byly zpracovávány v původní implementaci Eisnerově (např. 0:1), resp. v práci není dostatečně zdůvodněno, proč tyto korespondence autor vynechal.

- Aktuální implementace (zásobníkový beam search) požaduje invariant, aby po expanzi hypotéza postoupila do nějakého dalšího zásobníku. Zásobníky jsou uspořádány podle počtu pokrytých slov. Dvojice stromečků 0:1 by hypotézu neposunula do dalšího zásobníku. (Dvojice 1:0 povolit lze, zahození části vstupu cyklem nehrozí.)
- Podpora korespondence 0:1 potenciálně vede k zacyklení: na základě ničeho se přidává do výstupu uzel. Jediným mechanizmem, který zabraňuje nekonečnému bobtnání výstupu, je zhoršení skóre takové delší hypotézy. (S ohledem na skóre tedy lze i případy 0:1 povolit, při nízké penalizaci ale přesto hrozí výrazné zpomalení.)
- Aktuální implementace problematické situace „obejde“ užitím větších pravidel: např. 1:2 místo 1:1+0:1.

Literatura

Ondřej Bojar and Magdalena Prokopová. Czech-English Word Alignment. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1236–1239. ELRA, May 2006.

Ondřej Bojar and Magdalena Prokopová. Czech-English Machine Translation Dictionary. Technical report, ÚFAL MFF UK, Prague, Czech Republic, April 2007.

Ondřej Bojar and Zdeněk Žabokrtský. CzEng: Czech-English Parallel Corpus, Release version 0.5. *Prague Bulletin of Mathematical Linguistics*, 86:59–62, 2006.

Ondřej Bojar, Jiří Semecký, and Václava Benešová. VALEVAL: Testing VALLEX Consistency and Experimenting with Word-Frame Disambiguation. *Prague Bulletin of Mathematical Linguistics*, 83:5–17, 2005.

Ondřej Bojar, Evgeny Matusov, and Hermann Ney. Czech-English Phrase-Based Machine Translation. In *FinTAL 2006*, volume LNAI 4139, pages 214–224, Turku, Finland, August 2006. Springer.

Ondřej Bojar, Miroslav Janíček, Zdeněk Žabokrtský, Pavel Češka, and Peter Beňa. CzEng 0.7: Parallel Corpus with Community-Supplied Translations. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008. ELRA.

Martin Čmejrek. *Using Dependency Tree Structure for Czech-English Machine Translation*. PhD thesis, ÚFAL, MFF UK, Prague, Czech Republic, 2006.

Adrià de Gispert, José B. Mariño, and Josep M. Crego. Improving statistical machine translation by classifying and generalizing inflected verb forms. In *Eurospeech 2005*, pages 3185–3188, Lisbon, Portugal, September 2005.

Petr Sgall, Eva Hajičová, and Jarmila Panevová. *The Meaning of the Sentence and Its Semantic and Pragmatic Aspects*. Academia/Reidel Publishing Company, Prague, Czech Republic/Dordrecht, Netherlands, 1986.

Zdeněk Žabokrtský. TectoMT: Highly Modular Hybrid MT System

Literatura

with Tectogrammatics Used as Transfer Layer. In *Proc. of the ACL Workshop on Statistical Machine Translation*, page In print, Columbus, Ohio, USA, 2008.