

Počítače a přirozený jazyk

Ondřej Bojar
bojar@ufal.mff.cuni.cz

22. únor 2006

Osnova

- Kontakt
- Dohoda o společných nástrojích
- Za co bude zápočet
- Dnešní “výklad” :
 - Trojúhelník strojového překladu
 - Ilustrace: předmět zájmu lingvistů
 - BLEU: standardní metrika kvality překladu
 - Ilustrace předzpracování trénovacích dat
 - Příčiny nízkého skóre BLEU
 - Souhrn experimentů frázového statistického systému pro čj→aj

Kontakt + čím se zabývám

Ondřej Bojar

obo@cuni.cz

<http://www.cuni.cz/~obo>

“Dohoda” o společných nástrojích

- CVS
- make
- latex
- coreutils (grep, sed)
- Perl
- sourceforge?

Za co bude zápočet

Zápočet bude udělen za zápočtovou úlohu a “publikaci”.

Zápočtová úloha je jedno z:

- netriviální program na předem domluvené téma
- netriviální experiment s dostupnými programy (srovnání více programů, více přístupů k nějakému konkrétnímu problému)
- netriviální příprava lingvistických dat (sběr, očištění a strojová anotace ap.)
- netriviální rešerše nad přístupy k nějaké konkrétní otázce

Zápočtové úlohy je nutno ve všech případech předat v plně provozuschopné podobě. Kdokoli musí být schopen program snadno spustit, experiment nechat znovu proběhnout.

Zápočet (II)

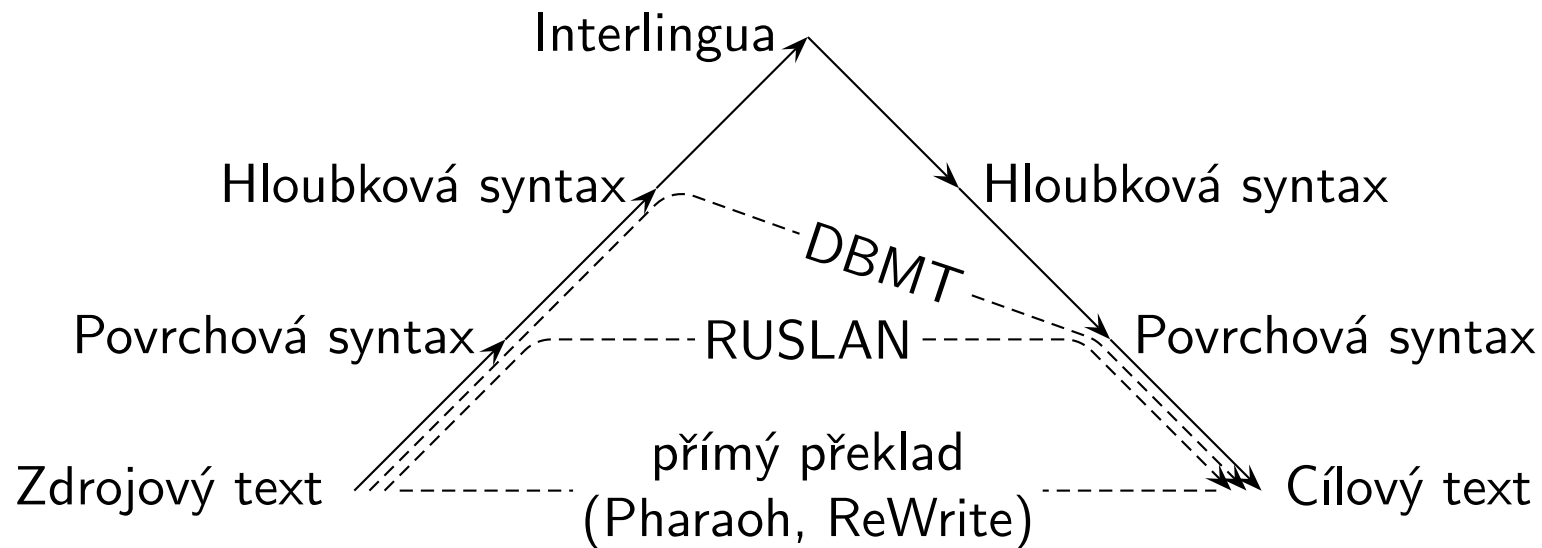
Úlohy je po předchozí dohodě možné a často vhodné řešit ve skupinkách, dvou až čtyřčlenných.

“Publikace” je anglický příspěvek popisující provedený experiment, implementovaný program, sebraná a anotovaná data.

“Publikace” musí splňovat veškeré náležitosti běžného konferenčního příspěvku (struktura, formát, citace, . . .).

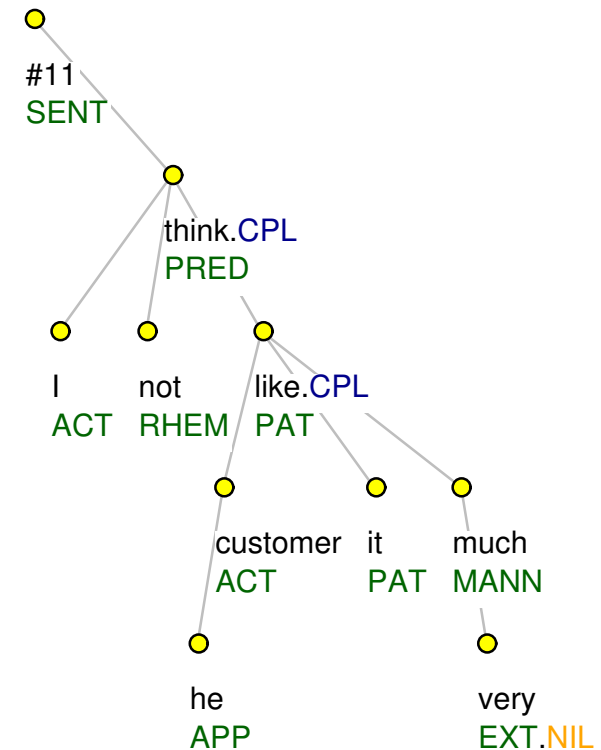
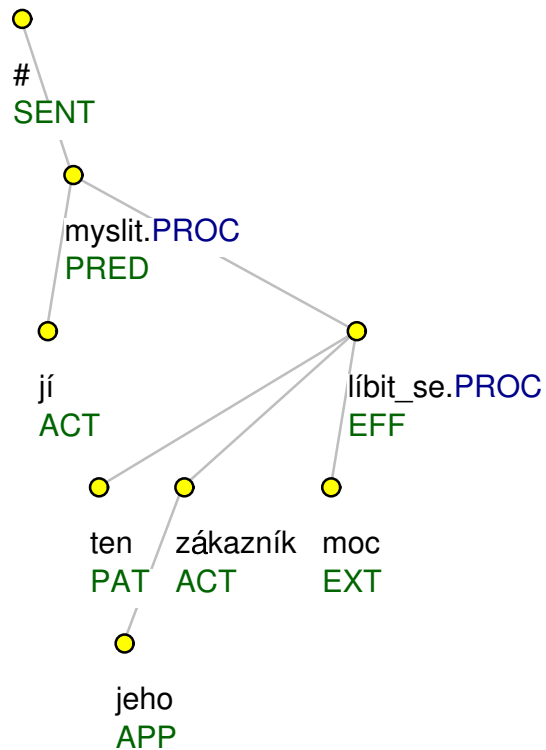
Rozsah upřesníme podle dohodnuté úlohy a kolektivu, obecně bude mezi 4 a 10 stranami dvousloupcového textu.

Trojúhelník strojového překladu (MT)



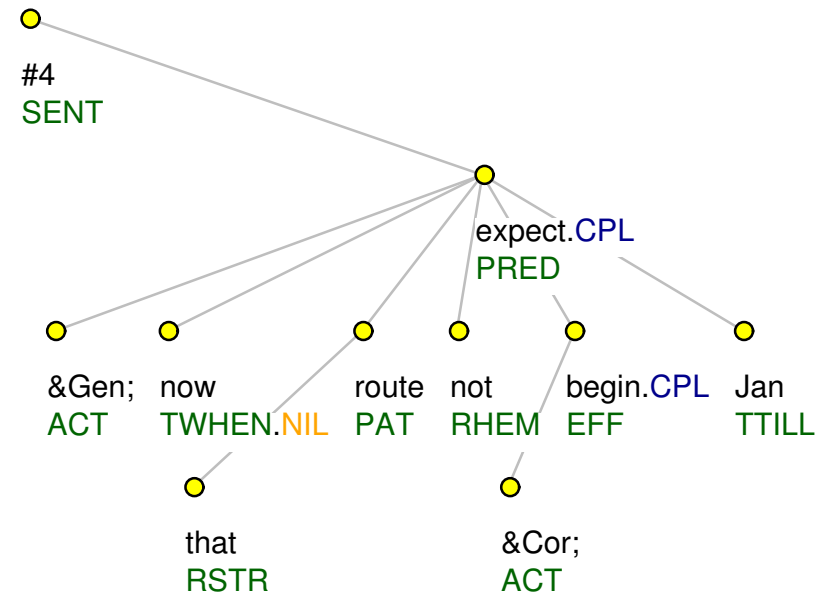
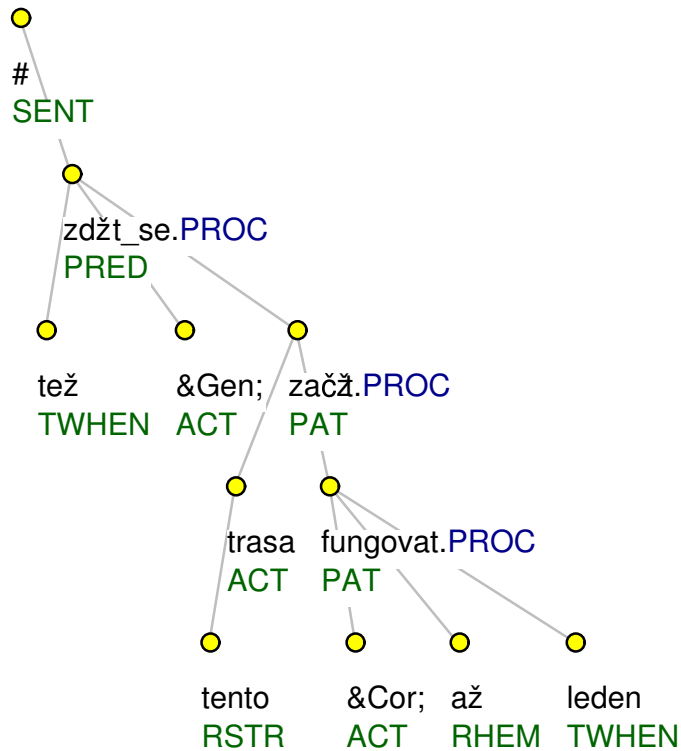
DBMT a ReWrite viz Čmejrek et al. [2003] a další, Pharaoh viz Koehn et al. [2003]

Ilustrace: hloubkově-syntaktické stromy



" Nemyslím , že by se to jejich zákazníkům moc líbilo . " " I do n't think their customers would like it very much . "

Ilustrace: hloubkově syntaktické stromy



Teď se zdá , že tyto trasy začnou fungovat až v lednu . Now , those routes are n't expected to begin until Jan .

Motto: od začátku pracuj od konce

Chceme spokojeného uživatele. Uživatel bude spokojený, když bude překlad “dobrý”. Co je “dobrý” překlad?

Referenční překlady od profesionála:

- (1) japan remained the biggest trading partner
- (2) japan is still the largest trade partner
- (3) japan still remain the number one trade partner

Návrhy systému:

- (4) japan will continue to be partner big
- (5) japan is still the biggest trading partner
- (6) ukraine won 2:1 against poland

BLEU: standardní metrika kvality překladu

Překlad (hypotéza):

Papineni et al. [2002]

n=1: For example , Fidelity prepares for case market plunge ads several months in advance .

n=2: For example , Fidelity prepares for case market plunge ads several months in advance .

Reference:

Fidelity Investments , for example , created their advertisements several months in advance , just in case the market dropped .

For example , Fidelity prepared advertisements for a potential market slump a few months in advance .

For example , Fidelity prepared ads some months in advance for a case where the market fell .

For instance Fidelity prepared ads for the event of a market plunge several months in advance .

BLEU = podíl 1- až 4-gramů z hypotézy doložených v referenčních překladech

- v rozsahu 0-1, někdy zapisováno jako 0 až 100 %
- lidský překlad proti dalším lidským překladům: cca 60 %
- Google čínština→angličtina: cca 30, arabština→angličtina cca 50.

Existují i další metriky (Word Error Rate, Position-Independent WER, NIST)

Statistický překlad po slovech či frázích

- trénovací soubor **paralelních textů**
- zarovnání po slovech
- extrakce slovníku (překlady slov či frází)
- decoding (překlad) = hledání “nejhladší formulace”
nejhladší \sim 3-gramy v mé hypotéze ať jsou v průměru (součin pstí) co nejběžnější (často spatřeny korpusu cílového jazyka, tzv. **jazykovém modelu**)

	Skóre	Zdrojová fráze	Cílová fráze
funguje	2.30	že bude	it would
reklama	2.79	že bude	he would
zda	3.08	že bude	he will
,	3.08	že bude	it will
Uvidíme	3.48	že bude	it will be
	3.77	že bude	it would be
	4.17	že bude	be
	4.17	že bude	it is
	...		

Ukázka překladu z češtiny do angličtiny

We 'll see whether the campaigns work .

Immediately after Friday 's 190 14-point stock market and a consequent uncertainty excretes several big brokerage firms new ads UNKNOWN_vytrubující usual message : Go on in investing , the market is in order .

Their business is persuade clients from escaping from the market , which individual investors masse fact , after plunging in October .

Uvidíme , zda reklama funguje .

Okamžitě po pátečním 190 bodovém propadu akciového trhu a následné nejistotě vypouští několik velkých brokerských firem nové inzeráty vytrubující obvyklé poselství : Pokračujte v investování , trh je v pořádku .

Jejich úkolem je odradit klienty od útěku z trhu , což jednotliví investoři hromadně činili po propadu v říjnu .

Ilustrace předzpracování trénovacích dat

		Vocab		Singl/Vocab	
Vstup do automatického hledání zarovnání po slovech		CZ	EN	CZ	EN
Formy	Produkce malých vozů se více než ztrojnásobila .	57k	31k	55.1%	47.6%
Stem4	Prod malý vozů se více než ztro .	17k	14k	36.5%	35.8%
Stem42	Prod/ce malých vozů se více než ztro/la .	52k	28k	51.2%	45.3%
Lem+Sing	produkce malý vůz se hodně než-2 UNK-verb .	15k	13k	0.1%	0.0%
Lemata	produkce malý vůz se hodně než-2 ztrojnásobit .	28k	25k	46.4%	47.5%

	vstup	do překladače	výstup
baseline	na 57,375 dolarech	na 57,375 dolarech	at UNK_57,375 \$
řešení čísel	na 57,375 dolarech	na _NUM dolarech	at \$ 57,375
čísla+začištění	na 57,375 dolarech	na _NUM dolarech	at \$ 57.375

Příčiny nízkého skóre BLEU

Nejvýznamnější chybějící bigramy:		Nejvýznamnější nadbytečné bigramy:					
19	, "	12	" said	26	, ' '	18	' ' .
12	of the	10	Free Europe	14	" said	12	, which
10	Radio Free	7	. "	11	Svobodná Evropa	8	, when
6	L.J. Hooker	6	United States	8	the state	7	, who
6	in the	6	the United	7	J. Hooker	7	L. J.
6	the strike	5	" We	7	company GM	7	firm Hooker
5	, a	5	is a	7	radio Svobodná	7	spokesman for
5	margin calls			7	the company		
4	28 tokens, 7 types			6	18 tokens, 3 types		
3	54 tokens, 18 types			5	35 tokens, 7 types		
2	94 tokens, 47 types			4	40 tokens, 10 types		
1	698 tokens, 698 types			3	117 tokens, 39 types		
				2	342 tokens, 171 types		
				1	3214 tokens, 3214 types		

Chybějící bigram = obsažen ve všech referencích, ale ne hypotéze

Nadbytečný bigram = obsažen v hypotéze, ale v žádné z referencí

Souhrn série experimentů: co zlepšuje BLEU

vhodné zarovnání po slovech	+1.5 až +2.0
morfologické předzpracování (stemming)	+1.0
morfologické předzpracování (plná lemmatizace)	+1.5
přidání nepředzpracovaného slovníku	+0.2
dodatečné paralelní texty, použity i v jazykovém modelu	+0.7 až +1.7
větší jazykový model v doméně	+2.1 až +3.4
ještě větší, ale obecný jazykový model	+4.6
dodatečné paralelní texty, ale jazykový model (větší) v doméně	+5.0 až +6.0
pravidlové zpracování číselných výrazů	+0.5
umělé zvětšování trénovacích dat na základě syntaktické struktury	+0.5
oprava evidentních prohřešků proti referenčním překladům	+1.0 až +1.5
sjednocení tokenizace v hypotéze a referenčních překladech	+10.0

Nadhled nad přístupy k úloze strojového překladu

Modelový lingvista usiluje o popis jazyka, vysvětlení toho, co se děje, když si lidé rozumějí.

Modelový statistik usiluje o řešení dané úlohy s co nejmenší chybou.

- statistik potřebuje úlohu
- statistik potřebuje metriku
- statistik ctí princip Occamovy břitvy
- statistik zohledňuje zákon klesajícího zisku
- statistický systém strojového překladu je snadno portovatelný na jiné jazyky

Domácí úkol

- Společnou silou vytvořit 4 *nezávislé* překlady celkem 515 anglických vět do češtiny.
Pokuste se o *dobrý* překlad, hledejte ve slovnících, hledejte na webu.
- Změřit 4x BLEU jednoho překladu proti třem zbývajícím pomocí “oficiální” implementace BLEU.
skript mt_eval, Google: “NIST machine translation evaluation”

Pro zájemce: najít webové překladače, které překládají do češtiny a vyzkoušet je (ev. změřit).

References

- Martin Čmejrek, Jan Cuřín, and Jiří Havelka. Czech-English Dependency-based Machine Translation. In *EACL 2003 Proceedings of the Conference*, pages 83–90. Association for Computational Linguistics, April 2003. ISBN 1-932432-00-0. MSM113200006, LN00A063.
- Philipp Koehn, Franz Joseph Och, and Daniel Marcu. Statistical phrase based translation. In *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT/NAACL)*, 2003.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In *ACL 2002, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, 2002.