Experimenty s frázovým statistickým překladem z češtiny do angličtiny

Ondřej Bojar
obo@cuni.cz

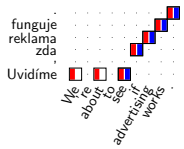
27. únor, 2006

Osnova

- Statistický strojový překlad na 1 slide.
- Dostupná data.
- Zarovnávání po slovech: ruční a série experimentů se strojovým.
- Frázový statistický překlad: metrika, série experimentů.
- Souhrn a varování, výhled.
- Širší zamyšlení a "pracovní návyky".

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 ∼ 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
2.30	že bude	it would
2.79	že bude	he would
3.08	že bude	he will
3.08	že bude	it will
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
...		

Dostupná data pro překlad (říjen 2005)

Korpus a verze	Vět	Tokenů	Slovník	Lemat	Pozn.
Prague Cz-En Dep Tbk (PCEDT 1.0)	22k/49k	0.5M/1.2M	57k/30k	28k/25k	
Reader's Digest (PCEDT 1.0)	44k/44k	658k/755k	84k/36k	?	beletrie
Kačenka	128k/105k	1.5M/1.5M	102k/47k	39k/22k	beletrie
OPUS EU Constitution	11k/10k	127k/164k	?	?	špatná tok.
Kolovratník	107k/107k	1.3M/1.5M	190k/92k	?	netokeniz.

BEAST: kompilace č-a slovníků z webu (400k párů, 235k cs, 225k en hesel; pouze jednoslovná hesla: 138k párů, 58k cs, 53k en)

Ruční zarovnání po slovech

Anotováno: 2x 515 vět z dev. a test. dat PCEDT 1.0.
Typy spojení: žádné/jisté/možné/bez ekvivalentu a žádné/frázové.

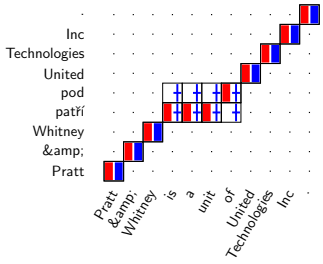
Anotováno celkem 2x16 000=32 000 spojení.
Z toho 18 % (5 800) vytvořil jen jeden anotátor a druhý ne.

Pokud odhlédneme od typu spojení, neshoda klesne na 9 %.

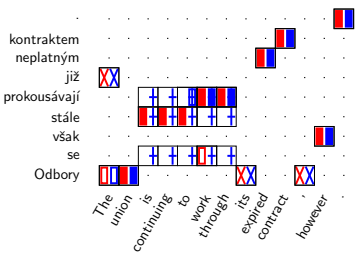
Neshody zejména v:

- anotování součástí analytických predikátů (někdy je konstrukce paralelní, někdy ne)
- anglických členech v případech, kdy na české straně nelze určit řídicí podstatné jméno, např. proto, že při překladu byl změněn slovní druh
- interpunkci (zejména apozice, koordinace ad. v kombinaci)

Příklad téměř ideální



Méně monotónní překlad



Metrika kvality zarovnání

Referenční data standardně neobsahují spojení frázového typu.
Referenční data jsme vytvořili sloučením obou anotací:

- oba anotátoři volí jisté spojení → jisté spojení
- jeden volí jisté a jeden nějaké jiné → jisté spojení
- alespoň jeden volí nějaké spojení → možné spojení

Automat nechť přiřazuje pouze jeden druh spojení (žádné/jisté).

P ... možná spojení, $S \subseteq P$ ... jistá spojení, A ... navrhovaná spojení

precision $= \frac{|A \cap P|}{|A|}$ (chyba uvést falešný, takový, který není ani možný)

recall $= \frac{|A \cap S|}{|S|}$ (chyba zapomenout jistý)

Alignment Error Rate, AER $= 1 - \frac{|A \cap P| + |A \cap S|}{|A| + |S|}$

Výsledky variant předzpracování a spojování

GIZA++ (Och and Ney, 2003) jednomu slovu vždy přiřadí nejvýše jedno odpovídající slovo (alignment je (neprostou) funkcí, 1-n).
Použita ve dvou směrech, konečný alignment lze získat sjednocením či průnikem výsledků z obou směrů.

	Průnik (1-1)			Sjednocení (n-n)		
	Prec	Rec	AER	Prec	Rec	AER
Baseline	97,4	57,6	27,4	65,9	86,7	25,5
Lematizace	97,9	75,0	15,0	77,1	89,8	17,2
Lematizace + čísla	97,9	75,2	14,8	77,5	89,9	17,0
Lematizace + singletony	97,4	75,8	14,6	77,8	88,5	17,4

Použitím symetrizace (nejlevnější párování) místo průniku/sjednocení (Matusov, Zens, and Ney, 2004) lze dosáhnout prec 91,4, rec 85,0, AER 11,9 %.

Kde selhává GIZA, měli problémy i lidé

Podíl tokenů, kde se zarovnání shodovalo (OK) nebo neshodovalo (Potíže):

- Lidé proti sobě
- GIZA++ proti referenci vzniklé spojením obou ručních anotací

		Baseline		Lematizace+singletony	
Lidé	GIZA++	en	cs	en	cs
Potíže	Potíže	14,3	15,5	14,3	15,5
Potíže	OK	0,1	0,1	0,2	0,1
OK	Potíže	38,6	35,7	25,2	25,0
OK	OK	46,9	48,7	60,4	59,4

⇒ U pozic, kde GIZA selhala, měli ve 38 % případů potíže i lidé.
⇒ Zlepšení díky lematizaci nepomáhá tam, kde lidé stejně měli potíže.

Hrubá kombinace pravidel a statistiky ubližuje

Členy v češtině nejsou, při jejich ručním zarovnání se postupuje podle předem daných pravidel.

Úvaha: Když členy odstraním, nechám zarovnat ostatní slova a pak členy přivěsím podle pravidel, měl bych dosáhnout lepších výsledků shody.

Zklamání: členy mají "více významů", někdy mají i svůj protiklad v češtině, a pak metoda s jednoduchým pravidlem jen ublíží.

dollar a share = dolar na akcii
the house = tento dům

Pokles o cca 0,5 procentního bodu v prec, rec i AER.

Souhrn zarovnání po slovech

- Úloha zarovnání po slovech by si zasloužila mírně předefinovat, přiřazovat k sobě "tektogramatické uzly", ne jednotlivá slova.
- Při staré definici je kvalita strojového zarovnání po slovech velmi dobrá.
- Vhodným předzpracováním (lematizace+náhrada singletonů slovním druhem) lze chybu snížit na polovinu.
- Nejlepší metodou spojování dvou směrů alignmentu je podle AER symetrizace, z jednoduchých postupů je výrazně lepší průnik než sjednocení.

BLEU: standardní metrika kvality překladu

Překlad (hypotéza):
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)

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 .

Práce s neznámými slovy
Úprava tokenizace referenčních překladů

Neznámá slova	DEV-FIX	TEST-FIX	DEV-ORIG	TEST-ORIG
Přiznat	30.2	25.9	20.8	17.6
Smazat	31	26.5	22.5	19.1
Ponechat nepřeložená	32.4	27.3	21.9	18.4

- ORIG – referenční překlady ponechány v základní podobě
- FIX – referenční překlady automaticky tokenizovány podobně jako trénovací data
  ⇒posun BLEU o ∼10 procentních bodů (1/3 celkového skóre!)

Lematizace > jednoduchý stemming

		DEV-std	TEST-optbleu	TEST-std
	stem42	28.5	26.1	23.5
baseline →	formy	28.6	25.8	23.6
	lemata + singletony	29.3	27.1	24.9
	stem4	29.6	26.7	23.9
	lemata	29.8	27.3	24.6

		Vocab		Singl/Vocab	
Vstup do automatického 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%

Více Dat? LM>paralelní korpus>slovník

		DEV-std	TEST-optbleu	TEST-std
	pcedt5k ali:lemata	22.7	21.5	19.1
	pcedt5k lmpcedt ali:lemata	25.6	24	21.2
	pcedt10k ali:lemata	26.6	23.7	21.2
baseline →	pcedt20k ali:lemata	29.8	27.3	24.6
ník je horší →	pcedt20k+dict ali:stem4	29.8	27.5	24.6
par. korp. →	pcedt20k+stories ali:stem4	31.6	28	25.9
	pcedt20k+dict lmpcedt ali:stem4	32.7	29.6	26.9
a než LM →	pcedt20k lmpcedt ali:lemata	33.2	29.4	26.4
	pcedt20k lm600M4grKN ali:lemata	33.4	31.9	27.3
	pcedt20k+stories lmpcedt ali:stem4	35.9	32.3	29.7

pcedt 5k 10k 20k	základní paralelní korpus, různé množství trénovacích vět
dict	nerozgenerovaný č-a slovník z webu, 116k hesel, 198/202k tokenů, 20k/30k vocab.
stories	dodatečné paralelní texty, 85k vět, 1.5/1.7M tokenů, 118/44k vocab.
lmpcedt	LM v dané doméně, (Čmejrek, Cuřín, and Havelka, 2003), n-gram vocab. 0.4:5:7M
lm600M4grKN	"obecný" jazykový model, 600M tokenů, n-gram vocab. 1.7:26:38:63M

Pravidlové řešení vlastních jmen a čísel

Ponechat vlastní jména v originále ubližuje (skloňování a tokenizace).
Pravidlové ošetření čísel mírně pomáhá.

	DEV-std	TEST-optbleu	TEST-std
jména+čísla	25.1	23.4	21.3
jména+čísla+začištění čísel	25.5	24.9	22.9
jména	25.8	-	21.4
čísla	29.2	27.1	24.2
čísla+začištění čísel	29.7	28.6	25.8
baseline	29.8	27.3	24.6

	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**

Umělé rozšiřování trénovacích dat podle závislostí

Nápad vytvořit nové trénovací věty (věty s novými ngramy) promazáním listů v závislostních stromech ("redukce" vět).

- Off-line: vypiš všechny možné věty, které lze získat postupnými redukcemi trénovacích vět.
  ⇒nepoužitelné, vede k explozi dat
- On-line: pro dané testovací zdrojové věty (tj. množinu "potřebných" ngramů)
  - Prohledej trénovací korpus s cílem najít *nesouvislé* ukázkové výskyty potřebných ngramů.
  - Označ nalezené uzly, alignované uzly v cílovém jazyce a též všechny sousedy v závislostních stromech tak, aby bylo dosaženo určité úrovně gramatičnosti.
  - Vypiš označené uzly (pokud nebyla nakonec označena celá věta).

Detail o rozšiřování trénovacích dat

263 testovacích vět obsahuje 5146 bigramů.

- 60 % má v trénovacích datech alespoň jeden nesouvislý výskyt
- 33 % nemá žádný výskyt
- 7 % má jen souvislé výskyty.

Z celkem 440 tisíc dohledaných příkladů je:

- 20 % ignorováno (jsou souvislé)
- 60 % spíše náhodné souvýskyty (příliš vzdálené v závislostním stromě)
- Zbývajících 20 % (93 tisíc) se zdá být k užitku.

Po dodání uzlů nutných pro zlepšení gramatičnosti ovšem 92 % z 93 tisíc příkladů svou užitečnost ztrácí, protože se stanou opět nesouvislými. Nakonec je tedy použito 7800 částí vět (jen 2000 unikátních) jako dodatečná trénovací data.

Celkový přínos metody je zanedbatelný

	pcedt 20k	pcedt 10k	pcedt 5k
Baseline	27.3	23.7	21.5
Rozšířená trénovací data	27.4	23.4	21.2
Rozšířená po odfiltrování "L.J. Hooker"	27.8	-	-

Zarovnání bylo vytvořeno pomocí sjednocení a lematizovaných vět. Výsledky jsou uvedeny na testovacích datech při optimalizaci na BLEU.

Souhrnný dojem: rozšiřování korpusu podle závislostí mírně pomáhá, pokud

- zajistíme gramatičnost dogenerovaných vět (pravidla závislá na jazyce
- získané věty ještě pečlivě profiltrujeme od podezřelých vzorků

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

Oprava evidentních prohřešků proti referencím

	DEV-std	TEST-optbleu	TEST-std
pcedt5k	22.7	21.5	19.1
pcedt5k s opravou	24.5	22.2	20
pcedt20k	29.8	27.3	24.6
pcedt20k s opravou	31.6	28.2	25.6
pcedt20k lm600M4grKN	33.4	31.9	27.3
pcedt20k lm600M4grKN s opravou	35.1	32.9	28.4

"Oprava" je přitom jen čtveřice pevných náhrad:

' ' .	→	. "
' '	→	"
L. J. Hooker	→	L.J. Hooker
the U.S.	→	the United States

Souhrn série experimentů: co zlepšuje BLEU

zarovnání jiné než průnikové	+1.5 až +2.0
morfologické předzpracování (stemming)	+1.0
morfologické předzpracování (plná lematizace)	+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

Shrnutí a varování

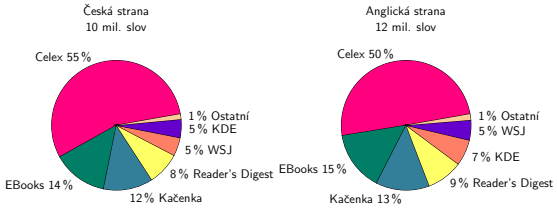
- Od začátku pracuj od konce.
  Jinak se plýtvá časem na minoritní problémy.
- Dílčí metrika podúlohy nemusí korelovat s celkovým hodnocením.
  AER doporučuje průnik alignmentů, BLEU říká, že průnik překladu škodí.
- BLEU je příliš citlivé na detaily.
  ⇒ pomáhá "normalizace" dat (Leusch et al., 2005).
- PCEDT není realistický zdroj dat pro překlad z češtiny do angličtiny.
  Čeština je příliš anglická, překlad do angličtiny nespravedlivě snadný.
- Komunikujte! Komunikujte! Komunikujte!
  Ruční zarovnání po slovech na stejných datech nezávisle a současně dělala Ivana Kruijff-Korbayová a Klára Chvátalová, aniž bychom o sobě věděli.

Výhled / přání

- Referenční překlady *do* češtiny. (Např. PCEDT.)
  Pokouším se vytvořit společně se studenty na FJFI, ale kvalita bude nevalná.
- Hledá se lepší metrika.
  Hodnotit chyby v závislostech, specificky hodnotit chyby ve slovním tvaru. Odstranit přílišnou citlivost na detaily (určitého typu). Kontrolovat konzistenci věty jako celku.
- Hledají se data pro vyhodnocení kvality metriky.
  Je potřeba soubor řady lidských hodnocení nad množinou referenčních překladů. Dobrá metrika je taková, která kandidátské/referenční překlady uspořádá podobně jako lidé.

CzEng (pre-release)

Paralelní korpus, který jsme shromáždili se Zdeňkem Žabokrtským.



Širší zamyšlení

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
- povaha práce na SMT je velmi jiná, řeší se zejména inženýrské problémy, jak rychle zpracovat velké množství dat ⇒ více informatiky než lingvistiky.

Pracovní návyky (jak se dělá špičkový ústav)

- Odborně vysoce fundovaný ředitel, mírně psí režim.
- "Žádný krok mimo".
- Lidé maximálně využívající strojové síly. (Makra na každém kroku.)
- Práce nad společným softwarovým dílem, všichni přispívají.
- Komplexní nástroj téměř zcela vlastní provenience (i vlastní FSA).
  ⇒ lze velmi rychle adaptovat a testovat nové věci.
- Kvalitní implementace (rychlá a úsporná):
  ⇒ umožňuje mnoho vývojových cyklů za jednotku času
- Vysoce kvalitní infrastruktura.
  Paralelní výpočty s minimální režií: rychlý síťový souborový systém, uživatel nerozhoduje, na kterém počítači se úloha spustí.

Jednoduché je krásné. Kratší je lepší.

Literatura

Čmejrek, Martin, Jan Cuřín, and Jiří Havelka. 2003. Czech-English Dependency-based Machine Translation. In *EACL 2003 Proceedings of the Conference*, pages 83–90. Association for Computational Linguistics, April. MSM113200006, LN00A063.

Leusch, Gregor, Nicola Ueffing, David Vilar, and Hermann Ney. 2005. Preprocessing and Normalization for Automatic Evaluation of Machine Translation. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 17–24, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Matusov, E., R. Zens, and H. Ney. 2004. Symmetric Word Alignments for Statistical Machine Translation. In *Proceedings of COLING 2004*, pages 219–225, Geneva, Switzerland, August 23–27.

Och, Franz Josef and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29(1):19–51.