Jak se Mojžíš s Jozuem učili hindsky

Ondřej Bojar, Pavel Straňák a Dan Zeman
ve spolupráci s Gauravem Jainem, Michalem Hrušeckým
a Michalem Richterem

ÚFAL

23. listopadu 2009

Osnova

Proč právě Hindi?

- Žádný zvláštní důvod jsme neměli
- Na IJCNLP 2008 v Hyderabadu jsme zjistili, že je v Indii zájem o překlad, ale zároveň jej stále dělají pravidlově
- Pokus, jestli opravdu dokážeme díky statistickým metodám překládat do jazyka, o kterém nevíme prakticky nic
- Prakticky jsme se rozhodli, když vyhlásili soutěž v překladu v rámci pravidelného NLP Tools Contest na konferenci ICON (International (really 'Indian') Conference on NLP)
  - čekali jsme, že budeme nejhorší
  - byli jsme spíše mezi lepšími, tak jsme se rozhodli pokračovat

Osnova

Hindská data

- paralelní data
  - paralelní korpusy
    - EILMT (oficiální test data pro ICON 2008 NLP Tools Contest)
    - TIDES (taky z ICONu 2008, ale možno používat i dále)
    - Emille (ELDA)
    - Daniel Pipes (web site)
    - Agro corpus (Mumbai)
  - slovníky
    - polmenované entity z anglické Wikipedie
    - Shabdanjali
- hindská data
  - news korpus z několika hlavních hindských deníků ($>$ 300M slov)

Paralelní korpusy

- EILMT
  - 7k vět, turistika, licence jen na ICON 2008
  - v r. 2008 ofic. testovací data. Nejlepší výsledky byly z trénování jen na EILMT (TIDES škodil)
- TIDES
  - 50k+1k+1k vět, DARPA-TIDES, IIIT Hyderabad
  - cca 1,2M tokenů (anglických trénovacích)
  - automatický převod do dévanágarí, místy neúspěšný
- Emille
  - paralelní část obsahuje 200k anglických slov a překlady do několika indických jazyků
  - data i překlady jsou problematické, nejdou zarovnat
    - 2 pokusy o ruční opravu: Gaurav a Om
    - Omille: Omem vyčištěný Emille, který by již měl být paralelní, ale je taky o dost menší ($< 50\,\%$)
- Daniel Pipes
  - novinářův web, který obsahuje překlady autorových článků v 25 jazycích
  - 322 článků v hindí, 6761 párů vět en-hi

Slovníky a (jen) hindská data

- entity z anglické wikipedie
  - "Ladakh (Tibetan script: ལ་དྭགས་; Wylie: la-dwags, Ladakhi: [*laḍaks*]; Hindi: लद्दाख़, Urdu: لدّاخ, Hindustani pronunciation: [*ləˈdaːx*]; "land of high passes") is a region situated in the disputed state of Jammu and Kashmir which ..."
  - ukládáme dvojice: 1 slovo – text za (Hindi | Devanagari | Marathi | Sanskrit), který je v devanagari
- Shabdanjali
  - anglicko-hindský slovník (licence GPL)
  - také automaticky převedený do dévanágarí
  - cca 26 000 hesel
- hindské deníky
  - ne nezbytně indické: mj. CNN, Deutsche Welle, Dainik Jagran (Yahoo)
  - LM z těchto dat v r. 2008 nepomohl, letos jsme jej nepoužili

Out of Vocabulary

		Tides	Tides+DP	all − Tides
Tokens	Tides-test-en	369	348	2429 (8.940%)
	Tides-test-hi	839	830	3310 (11.584%)
	Tides-dev-en	464	421	1873 (8.330%)
	Tides-dev-hi	619	607	2661 (10.922%)

		Tides	Tides+DP	all − Tides
Types	Tides-test-en	363	343	1901 (32.009%)
	Tides-test-hi	642	633	2465 (41.979%)
	Tides-dev-en	459	418	1608 (28.735%)
	Tides-dev-hi	580	568	2129 (37.735%)

- Ostatní data (bez Tides) pokryjí cca. 90%/60% Tides (tokens/types).
- Tides types a tokens skoro stejné – slova s jedním výskytem.
- Hindi horší – tvarosloví, transkripce, homonyma ...

Něco o hindštině

- Indoevropský jazyk
  - Tj. vzdáleně příbuzný češtině (v některých slovech víc než třeba angličtina)
  - Ale spousta slov i z perštiny a arabštiny
- Prý volný slovosled, ale míň než v češtině
- SOV jazyk: „Ráma Móhana vidí.“
- Na konci často spona / pomocné sloveso být:
  - है (hai) = „je“ ... hodně častý konec věty
- Postpozice (záložky) místo předložek

Písmo dévanágarí

- (Polo)slabičné písmo

- क का कि की कु कू कृ के कै को कौ क्
- *ka ká ki kí ku kú kr ké kæ kó kau k*

- दस दिन को चाय पियो
- दरवाज़े के पास अलमारी है
- सर, रेलवे स्टेशन से साइकिल को लेना

Písmo dévanágarí

- Po transliteraci...

- दस दिन को चाय पियो
  das din kó čáj pijó
- दरवाज़े के पास अलमारी है
  darvázé ké pás almárí hæ
- सर, रेलवे स्टेशन से साइकिल को लेना
  sar, rélvé stéšan sé sáikil kó léná

Písmo dévanágarí

- Po transliteraci někdy příjemné překvapení

- **das din** kó **čáj pijó**
  = **deset dní pijte čaj**
- **darvázé** ké pás **almárí** hæ
  = u **dveří** je **skříň**
- **sar, rélvé stéšan** sé **sáikil** kó léná
  = **sir,** take the bi**cycle** from the **railway station**

Hindské pády

- Tradiční systém pádů *vibhakti*
- Skutečné pády jsou 2 (direct a oblique)
- Zbytek tvořen záložkami
  - Záložky dříve přilepené ke slovu, tj. pádové koncovky

- Příklad: genitiv
  - Delhi is the capital of India.
  - दिल्ली भारत का राजधानी है ।
  - dillí bhárat ká rádžadhání hæ.
  - Dillí Indie *genitiv* hlavní-město je.

Normalizace dat

- Různé korpusy prošly různým zpracováním
- Tides:
  - Větu ukončuje tečka (.)
  - Číslice jsou euro-arabské (0123456789)
- Emille:
  - Větu ukončuje danda (।)
  - Číslice jsou z dévanágarí (०१२३४५६७८९)
- Co ještě lze napsat více způsoby:
  - Znaky s *nuktou* (क़ख़ग़ज़ड़ढ़फ़): फ़ vs. फ+़ vs. फ
  - Pořadí kombinované diakritiky: प+ा+ॅ vs. प+ॅ+ा
  - Nahrazení *čandrabindu anusvárem*: पाँच vs. पांच
  - Řídící znaky, zero-width joiners apod.
  - Ne-ASCII interpunkce, např. „—“ vs. „-“
- My se to snažíme v datech sjednotit
- Navíc re-tokenizujeme *(Anglo-American)*

Další hrůzy v datech

- Vsuvka v latince se během konverzí mylně považuje za romanizovaný zápis hindštiny:
  - *Information Commis(s)ioner* => ईन्डोर्मटिओन् छोम्मिसिओनेर् *(īṅṅormaṭion chommisioner)*, skutečná transkripce by byla spíš इन्फ़ोर्मेशन कोमिशनेर *(informešana komišanera)*
- Více než 200 hindských vět v Tides začíná v dévanágarí, pak ale náhle přejdou do nečitelné latinky:
  - प्रादेशिक – जनसंख्या बंगाली बंग्लादेश ह्यपूर्वी बंगालह्ह से आए अधिकांश विस्थापित दक्षिण अंडमान , नेल , हैवलाक , मध्य अंडमान , उ*<arI AMDmaana tqaa ilaiTla AMDmaana maoM basaae gae .*
- Znak *danda* (konec věty) zaměněn za svislítko, to zakódováno jako `|BAR;`, a to nakonec považováno za romanizovanou hindštinu: |भाष्;
- Opakující se záhadná sekvence ऋ–ऊण्श्छष्– (`Q-UNSCR-`; klidně uprostřed hindského slova)

Co už se normalizovat nedá

- Nejednotná transkripce anglických slov do dévanágarí

- स्टैंडर्डज *(sṭaiṁḍarḍaja)*
- स्टैंडर्डस *(sṭaiṁḍarḍasa)*
- स्टैंडर्ड्स *(sṭaiṁḍarḍsa)*

Co už se vůbec normalizovat nedá

- Synonyma podle původu slov

English	Hindi/Persian	Hindi/Sanskrit
language	ज़बान *(zabāna)*	भाषा *(bhāṣā)*
book	किताब *(kitāba)*	पुस्तक *(pustaka)*
newspaper	अख़्बार *(axbāra)*	समाचार–पत्र *(samācāra-patra)*
beautiful	ख़ूब्सूरत *(xūbsūrata)*	सुन्दर *(sundara)*
meat	गोश्त *(gośta)*	माँस *(mām̃sa)*
thank you	शुक्रिया *(śukriyā)*	धन्यवाद *(dhanyavāda)*

Osnova

Přehled pokusů

- Systém Moses (Mojžíš)
  - Faktorizovaný překlad
    - Různé modely morfologie
  - Vícefaktorový jazykový model
  - Různé kombinace dat
- Systém Joshua (Jozue)
  - Hierarchický frázový model
  - Různé kombinace dat

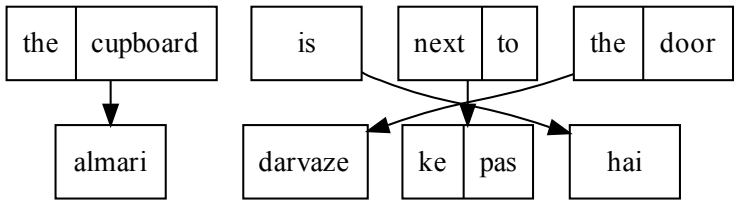
Připomenutí Mojžíšovy roury

1. Paralelní korpus zarovnej po slovech.
2. Extrahuj **fráze** konzistentní se zarovnáním po slovech.
3. Natrénuj hindský jazykový model (LM).
4. Natrénuj hindský reorderovací model.
5. Na vývojových datech vylaď váhy modelů (MERT).

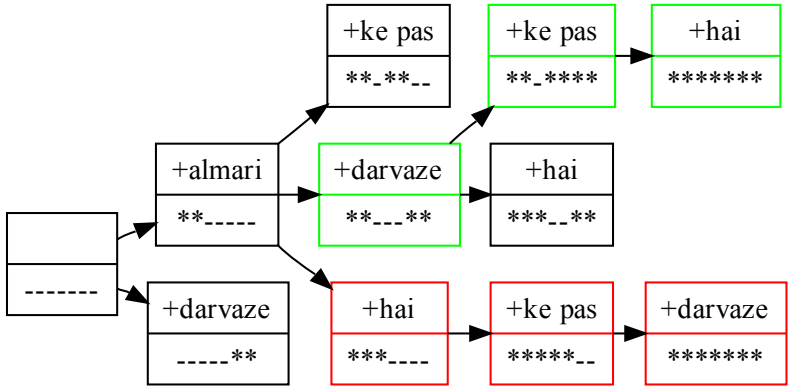
1. Vstupní větu **rozděl na fráze**.
2. Fráze přelož nezávisle.
3. Urči výsledné pořadí frází a spoj je.

Nevýhoda Mojžíše: Slabý frázový model

Potřebujeme, aby nejpravděpodobnější bylo toto:



Mojžíš takto rozvíjí hypotézy:



- Základní model: čím větší délka přesunu, tím dražší ($\Rightarrow$ monotonie).
- Částečně lze kompenzovat lexikalizovaným reorderingem:
  - P(monotone/swap/discontinuous | next to, ke pas)

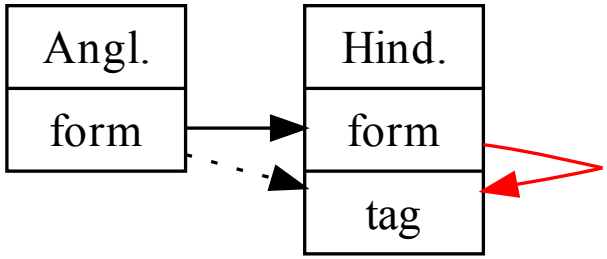
Pokusy z Bojar et al. (2008)	EILMT	TIDES
Baseline Moses, Distance Reordering	18.88±2.05	10.06±0.76
Baseline Moses, Reordering Using en+hi Forms	19.77±2.03	**10.95±0.75**

Výhoda Mojžíše: Zachycení morfologie

- Slova jsou vektory „faktorů“.
- Vybrané modely je tak možno založit na jemnější či hrubší reprezentaci slova.

Hrubší reprezentace na cílové straně:

- Umožňuje zapojit spolehlivější jazykový model (hustší data).



Zachycení tvarosloví

- Morfologie s učitelem (supervised).
  - Hindi POS Tagger (Gupta et al., 2006).
  - Koncovky z učebnice (Snell and Weightman, 2003).
- Morfologie bez učitele (unsupervised).
  - Posledních *n* písmenek slova.
  - Automatické (bigramové) slovní třídy (Brown et al., 1992; Och, 1995).
  - Hindomor (Zeman, 2008).
  - Affisix (Hlaváčová and Hrušecký, 2008).

Učebnicové koncovky

- Primitivní řízený stemming.
- Během 2 hodin jsme proběhli učebnici hindštiny pro samouky a ze všech gramatických tabulek vypsali koncovky skloňování a časování.
- Výsledkem je seznam asi 30 koncovek, včetně duplikátů.
- Pokud byla u slova nalezena známá koncovka, je to jeho „značka“.
- Velmi častá slova ponechána vcelku, jsou sama sobě „značkou“.

Příklady různých morfologií

Ukázková věta: *unheṁ vahāṁ kalakattā śahara dikhāyā gayā .*
Doslova: jim tam Kalkata město ukázáno bylo .
Vstup: They were shown Calcutta City .

Forma	Tag	Učeb.	2 písm.	WC10	hindomor	bbf	bdf	ddf
उन्हें	PRP	उन्हें	◌ें	2	◌ं	—	—	—
वहां	PRP	वहां	◌ां	2	◌ं	—	—	—
कलकत्ता	NNP	आ	ता	3	◌ा	ता	ता	—
शहर	NN	शहर	हर	3	र	र	—	—
दिखाया	VM	आ	या	7	◌ा	◌ा	—	—
गया	VAUX	गया	या	11	◌ा	—	—	—
.	SYM	.	.	6	—	—	—	—

- Tagy velmi chudé pro hidské tvarosloví.
- Automatické třídy v souladu s tagem.
- Různé konfigurace Affisixu (bbf, bdf, ddf) různě jemné.

Výsledky pokusů

Morfologie	BLEU	Morfologie	BLEU
tag	12.03±0.75	hitbsuf	11.58±0.74
wc50	11.97±0.73	hindomor2	11.55±0.74
wc10	11.76±0.74	hindomor1	11.54±0.71
lcsuf3	11.66±0.75	affddf	11.50±0.7
lcsuf1	11.63±0.72	affbdf	11.33±0.72
hindomor3	11.60±0.73	lcsuf2	11.14±0.74

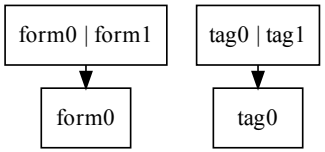
- Baseline bez morfologie: 11.46±0.72.
- Rozdíly mezi všemi konfiguracemi zanedbatelné.

Vícefaktorové jazykové modely (Michal Richter)

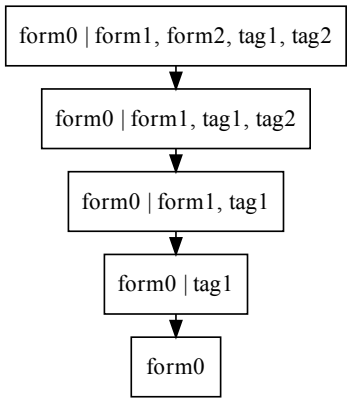
Factored LMs (Bilmes and Kirchhoff, 2003) zobecňují vyhlazování:

- n-gramové LM neviděné n-gramy skórují pomocí $(n-1)$-gramů.
- Faktorové LM pracují s faktorovými slovy $\Rightarrow$ kontext lze omezovat kratší historií i hrubší reprezentací slova (lemma, slovní druh).

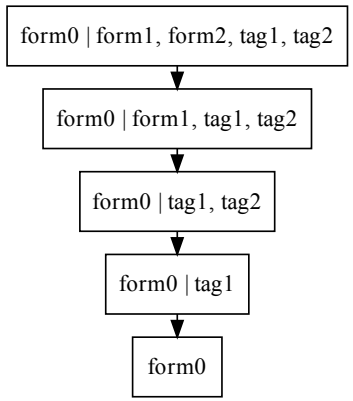
Uživatel definuje graf zapomínání, za běhu je pak pravděpodobnost dána nejpravděpodobnější cestou v grafu.



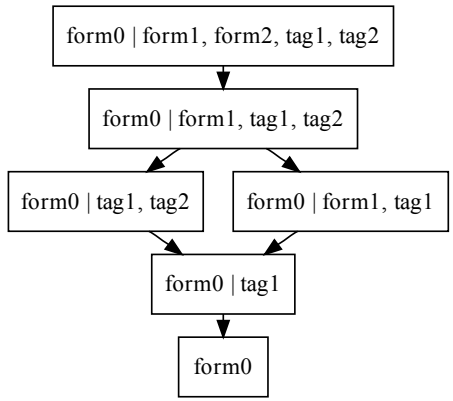
2×bigramový LM



linf



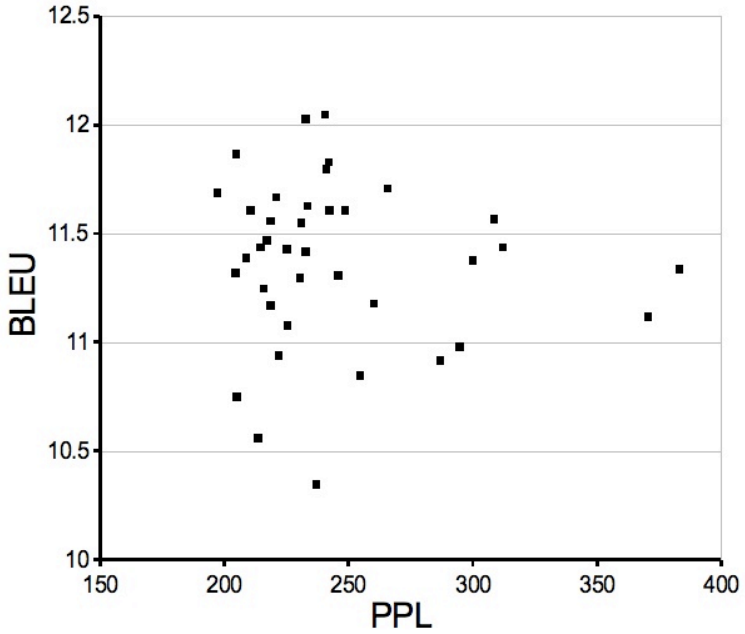
lint



fftt

Výsledky faktorových LM

Tvar + Morfologie	BLEU
fftt + lcsuf3	12,05±0,76
linf + wc10	12,03±0,73
fftt3 + wc50	11,87±0,77
fftt + wc10	11,83±0,73
fftt + hitbsuf	11,80±0,75
fftt3 + lcsuf1	11,67±0,75
2×3gr LM (forma, tag)	12,03±0,73



- Žádná korelace mezi perplexitou FLM a BLEU.
- FLM většinou ublížily BLEU.
- Ani výrazné zlepšení perplexity (197.0 vs. 212.6 baseline).

Hierarchické frázové modely

- Hiero (David Chiang, 2005)
- Joshua (open-source reimplementace z JHU)

- Fráze mohou obsahovat neterminály
  $=>$ synchronní bezkontextové gramatiky
  - Pravidlo má levou stranu a dvě pravé strany, anglickou a hindskou
  - Umožňuje zobecnit nesouvislé fráze, změny slovosledu nebo dokonce rekurzi
  - Typický neterminál: X_i (není to „lingvistická“ gramatika)
- Příklad:

$$X \rightarrow \langle X_1 \cdot of \cdot X_2 \rangle, \langle X_2 \cdot \text{का} \cdot X_1 \rangle$$

Mojžíšova roura

1. Paralelní korpus zarovnej po slovech.
2. Extrahuj **fráze** konzistentní se zarovnáním po slovech.
   - Ke každé dvojici frází známe tři veličiny:
     - Pravděpodobnost překladu zdrojové pravé strany na cílovou.
     - Lexikální pravděpodobnost překladu zdroje cílem po jednotlivých slovech.
     - Lexikální pravděpodobnost překladu cíle zdrojem po jednotlivých slovech.
3. Natrénuj hindský jazykový model (LM).
   - Ke každé hindské větě známe její pravděpodobnost podle LM.
4. Natrénuj hindský reorderovací model.
5. Na vývojových datech vylaď váhy uvedených veličin (MERT)

1. Vstupní větu **rozděl na fráze.**
2. Fráze přelož nezávisle.
3. Urči výsledné pořadí frází a spoj je.

Jozuova roura

1. Paralelní korpus zarovnej po slovech.
2. Extrahuj **pravidla gramatiky** konzistentní se zarovnáním po slovech.
   - Ke každému pravidlu známe tři veličiny:
     - Pravděpodobnost překladu zdrojové pravé strany na cílovou.
     - Lexikální pravděpodobnost překladu zdroje cílem po jednotlivých slovech.
     - Lexikální pravděpodobnost překladu cíle zdrojem po jednotlivých slovech.
3. Natrénuj hindský jazykový model (LM).
   - Ke každé hindské větě známe její pravděpodobnost podle LM.
4. Na vývojových datech vylaď váhy uvedených veličin (MERT).

1. Vstupní větu **rozeber synchronním chart parserem.**
2. Fráze přelož nezávisle.
3. Propoj je podle derivačního stromu.

Mojžíš vs. Jozue

Trénovací data	Joshua	Moses
Tides	12.27±0.83	11.46±0.72
Tides+DP	**12.58±0.77**	11.93±0.75
Tides+DP+Emille	11.32±0.74	10.06±0.72
Tides+DP+Dict	12.43±0.79	11.90±0.78

System	BLEU
Mumbai (Damani et al., 2008)	8.53
Kharagpur (Goswami et al., 2008)	9.76
Prague (Bojar et al., 2008)	10.17
Dublin (Srivastava et al., 2008)	10.49
present Joshua	11.10

Záhadný Emille

- Navzdory faktorům a modelům reorderingu v Mosesovi, Joshua zatím má náskok
- Jak u Joshuy, tak u Mosese se však projevil těžko vysvětlitelný negativní vliv přídavných dat, zejména Emilla
- Rozšíření Tides (50000 vět) o Daniela Pipese (7000 vět) pomohlo
- Naproti tomu přidání vyčištěného Emilla (3500 vět) zřetelně uškodilo
  - Na datech žádný viditelný problém (méně šumu než Tides)
  - Frázové tabulky vypadají OK
  - **Ale:** jasné přetrénování na vývojových datech (po prohození vývojových a testovacích dat problém zmizel)
  - Ukázalo se, že Emille je obsažen v trénovací (2000 vět) a vývojové části (100 vět z 1000) Tidesu!!!

Osnova

Způsob anotace

- Tři nezávislé sondy (100 vět, 100 vět, 43 vět, vždy jiné).
- Zobrazena zdrojová věta a hypotézy.
- Referenční překlad zamíchán mezi hypotézy.

SRC	the private sector units are thirty to forty years old .
	तीस हैं निजी आधार पर क्षेत्र यूनिट्स को को 40 वर्ष की आयु से ऊपर हैं ।
*	निजी क्षेत्र के स्थान पर हैं , 30 से 40 वर्ष से अधिक आयु के थे .
**	निजी क्षेत्र की इकाइयां 30 से 40 वर्ष तक पुरानी हैं .
**	निजी क्षेत्र की इकाइयों में 30 से 40 साल पुरानी हैं .

- Značky: *nic* pro nesrozumitelné, „*“ pro náznaky překladu, „**“ pro akceptovatelné a zachovávající většinu významu, byť s chybami.
- Kontrast s Ramanathan et al. (2009), kde zlepšili v průměru
  - z „*little meaning conveyed, disfluent Hindi, most phrases correct, ungrammatical overall*“
  - na „*much of meaning conveyed, non-native Hindi, few minor grammatical errors*“

Sonda 1: Mojžíš mimo doménu, s morfologií či více daty?

OOD mimo doménu: trénováno na všem mimo Tides

TIDP Tides + Daniel Pipes, bez morfologie

WC10 Tides + trigramový LM na automatických slovních třídách (10 tříd)

Systém	0	*	**	BLEU
REF	6	11	83	—
OOD	80	17	3	1.85±0.24
TIDP	26	44	30	11.93±0.75
WC10	38	46	16	11.76±0.74

- Šest (procent) referenčních překladů nepřijatelných!
- Doména velmi podstatná, OOD propadlo nejen v BLEU, ale i ručně.
- Lepší data navíc než automatická morfologie (TIDP>WC10).
- BLEU ovšem TIDP vs. WC10 neodliší.

Sonda 2: Mojžíš vs. Jozue

Systém	0	*	**	BLEU
REF	6	10	84	—
Joshua	32	37	31	12.58±0.77
Moses	35	35	30	11.93±0.75
Moses-DPipes+POStags	32	42	26	12.03±0.75

- Identická trénovací data (Tides + Daniel Pipes, bez morfologie).
- Jozue (nesignif.) lepší podle BLEU i lidského hodnocení.
- I druhý test Mosese ukazuje, že víc dat spíše lepší než morfologie.
  - Tentokrát užit POS tagger, nikoli automatické slovní třídy.
  - Ne zcela jednoznačný výsledek: klesne počet „**“, ale i „0“.
    Záleží na cílové aplikaci: přesnost vs. pokrytí.

Sonda 3: Jak je to s Emillem a Mojžíšem?

Systém	**0**	*	**	**BLEU**
REF	0	8	45	—
TI DP	20	14	19	11.89±0.76
TI DP EM	22	19	12	9.61±0.75
TI DP EM oth	17	25	11	10.97±0.79
TI DP EM oth DICTFilt	23	17	13	10.96±0.75
TI DP EM oth DICTFull	22	16	15	10.89±0.69

- BLEU tentokrát téměř souhlasí s lidmi.
- Přidání Emilla citelně sníží kvalitu.
- Další data tu ztrátu postupně kompenzují.

Osnova

Shrnutí

- Dosáhli jsme nejlepšího publikovaného BLEU skóre na testovacích datech TIDES
  - Srovnej ICON 2008 NLP Tools Contest
  - Obecně je srovnání en-hi překladu problematické, každý testuje na jiných datech
- Hierarchické modely dávají lepší BLEU než Mojžíšovy faktory a reordering
  - Při ručním vyhodnocení je ale jejich náskok méně přesvědčivý
- Poučení o datech
  - Získat data může být snadnější než je vyčistit
  - Dva různé korpusy z různých zdrojů nemusí být nutně různé!

- Co dál?
  - Opravdu nemůže morfologie pomoct víc?
  - Přeskládání slovosledu angličtiny
  - Z vybraných značek (např. *subject*) udělat tokeny

Zkušenosti se zahraničním studentem

- student magisterského studia na IIT v Bombaji
- ⊕ pilný, poslušný
- ⊖ neiniciativní, nevzal (žádný) úkol za svůj
  naše chyba: nečekali jsme to a příliš dlouho dávali komplexní (ne nutně těžké) úkoly
- ⊖ nedokončil ruční hodnocení ani dodatečně
  naše chyba: data dostal krátce před odjezdem, stihl jen část
  ovšem ani doma v klidu dlouho po návratu hodnocení nedokončil
- ⊖ skrývání problémů nebo spíše *nečekaně jiná rozlišovací úroveň*
  např. jsme se ptali na konkrétní větu, je-li přeložena dobře. Napřed byla, ale když jsme pojali podezření a zeptali se na konkrétní jevy (koncovka, slovosled), „přiznal“ chyby

Pro příště

- mít připravený seznam přesných malých úkolů
- laťku pro samostatnost v práci postupně zvyšovat, volné řízení se neosvědčilo

Děkujeme za podporu z těchto grantů:

- MSM0021620838 (Výzkumný záměr informační sekce MFF UK 2005–2010),
- FP7-ICT-2007-3-231720 (EuroMatrix Plus)

Literatura I

Jeff A. Bilmes and Katrin Kirchhoff. 2003. Factored language models and generalized parallel backoff. In *NAACL '03: Proc. of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 4–6, Morristown, NJ, USA. Association for Computational Linguistics.

Ondřej Bojar, Pavel Straňák, and Daniel Zeman. 2008. English-Hindi Translation in 21 Days. In *Proc. of ICON-2008 NLP Tools Contest*.

Ondřej Bojar, Pavel Straňák, Daniel Zeman, Gaurav Jain, Michal Hrušecký, Michal Richter, and Jan Hajič. 2009. English-Hindi Translation—Obtaining Mediocre Results with Bad Data and Fancy Models. In *Proceedings of the 7th International Conference On Natural Language Processing (ICON-2009)*, Hyderabad, India, December. NLP Association of India.

Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jennifer C. Lai, and Robert L. Mercer. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479.

Om P. Damani, Vasudevan N., and Amit Sangodkar. 2008. Statistical machine translation with rule based re-ordering of source sentences. In *Proc. of ICON-2008 NLP Tools Contest*.

Sumit Goswami, Nirav Shah, Devshri Roy, and Sudeshna Sarkar. 2008. NLP Tools Contest: Statistical Machine Translation (English to Hindi). In *Proc. of ICON-2008 NLP Tools Contest*.

Kuhoo Gupta, Manish Shrivastava, Smriti Singh, and Pushpak Bhattacharyya. 2006. Morphological richness offsets resource poverty- an experience in building a pos tagger for hindi. In *Proc. of COLING/ACL-2006*.

Literatura II

Jaroslava Hlaváčová and Michal Hrušecký. 2008. Affisix: Tool for Prefix Recognition. In *Proc. of Text, Speech and Dialogue, LNAI 5246*, pages 85–92. Springer.

Franz Josef Och. 1995. Maximum-Likelihood-Schiätzung von Wortkategorien mit Verfahren der kombinatorischen Optimierung. Studienarbeit, Universität Erlangen-Nürnberg, Germany.

Ananthakrishnan Ramanathan, Hansraj Choudhary, Avishek Ghosh, and Pushpak Bhattacharyya. 2009. Case markers and morphology: Addressing the crux of the fluency problem in english-hindi smt. In *Proc. of ACL/IJCNLP*.

Rupert Snell and Simon Weightman. 2003. *Teach Yourself Hindi*. Hodder Education, London, UK.

Ankit Kumar Srivastava, Rejwanul Haque, Sudip Kumar Naskar, and Andy Way. 2008. MaTrEx: The DCU Machine Translation System for ICON 2008. In *Proc. of ICON-2008 NLP Tools Contest*.

Daniel Zeman. 2008. Unsupervised acquiring of morphological paradigms from tokenized text. In *Advances in Multilingual and Multimodal Information Retrieval, 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007. LNCS 5152*, pages 892–899. Springer.