Překlad do hindštiny pro začátečníky

Daniel Zeman
Ondřej Bojar, Pavel Straňák

ÚFAL MFF, Univerzita Karlova, Praha

Granty MSM 0021620838, EU FP6-IST-5-034291-STP

NLP Tools Contest

- Společná úloha: překlad z angličtiny do hindštiny
- Při konferenci ICON, prosinec 2008, Puné, Indie
  - ICON je „International…“, ale hodilo by se „Indian…“
  - Pořádáno s nadšením, ale velmi amatérsky, zejména co do webu
- Nakonec se nám podařilo zaregistrovat i trefit do místa konání ☺
- V létě vyhlášeno, koncem října data, 1. prosince článek (BLUE jsme si počítali sami)

Osoby a obsazení

- Pavel nás do toho namočil a dodával data
- Ondra zajistil výcvik Mojžíše
- Dan předstíral znalost hindštiny

NAMOČITEL

DODAVATEL

PŘEKLADATEL

प्रेद्स्तीरतेल

Osoby a obsazení

- A další, kteří měli smůlu, že
  - Mají zajímavý software
  - Mají indické jméno
  - Potkali Ondru

Dear Bojar

Uméš Mirek श्री गनेश Pradžól

श्रीराम

Ganéš कर्थिक Vašek उमेश Dan

Noah

प्रजोल Pavel Philip Ondra Charlie Šrírám

Šrí Straňák Karthik Zdeněk & Co. శ్రీరాము

संजीव

Zúčastněné týmy

- Univerzita Karlova (Ondra, Dan, Pavel)
- Dublin City University (Andy Way a Indové)
- IIIT Hyderabad (Indové)
- IIT Bombay (Indové)
- IIT Kharagpur (další Indové)

- Jen u nás nikdo hindsky neuměl
  - Statistika se to musela naučit za nás

Něco o hindštině

- Indoevropský jazyk
  - tj. vzdáleně příbuzný češtině (v některých slovech víc než třeba angličtina)
  - Ale spousta slov i z perštiny a arabštiny
- Nápadná podobnost
  - čísla: 1 एक ék 2 दो dó 3 तीन tín 4 चार čár 5 पांच pánč 6 छह čhah 7 सात sát 8 आठ áth 9 नौ nau 10 दस das
  - dveře = दरवाज़ा darvázá
  - piju = पितो pitó
  - padat = पदन padana

Něco o hindštině

- Prý volný slovosled, ale zjevně míň než v češtině
- SOV jazyk:
  - „Ráma Móhana vidí.“
- Sloveso je vždy na konci
  - Často spona / pomocné sloveso být:
    - है (hai) = „je“ … hodně častý konec věty
- Postpozice (záložky) místo předložek

Něco o hindštině

- Tradiční systém pádů *(vibhakti)*
- Skutečné pády 2 (nominativ 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 hai
  - Dillí Indie *genitiv* hlavní-město je

Strategie

- Soustředili jsme se na tři okruhy:
  - Sehnat co nejvíc dat
  - Sehnat morfologii hindštiny (faktorizovaný překlad)
  - Udělat něco se slovosledem angličtiny

- Neplánovaně přibyl ještě čtvrtý:
  - Vhodné nastavení Mosese

Data

- Paralelní (en-hi)
  - TIDES (50k trénovacích vět, 1.2M hi slov)
  - EILMT (7k trénovacích vět, 181k hi slov)
  - EMILLE (12k vět, 200k en slov)
  - Daniel Pipes (322 texts, 15k vět, 300k hi slov)
  - Agriculture (17k en ~ 13k hi slov)
- Jednojazyčná (hi)
  - Hindi news web (18M vět, 309M slov)
  - EMILLE neparalelní (365k odstavců)
  - Hindská strana paralelních dat, viz výše

Testovací data

- Paralelní (en-hi)
  - TIDES
  - EILMT

- EILMT jsou velmi čistá data, specializovaná doména (turistika, hodně pojmenovaných entit)
- TIDES jsou sice několikrát větší, ale se šumem

Vliv přídavných dat

- Přídavná paralelní data pomáhají
  - Testovací data: EILMT
  - Trénovací & dev data:
    - EILMT 18,88 ± 2,05
    - EILMT+TIDES 19,27 ± 2,22
    - EILMT+TIDES+20k web vět 20,07 ± 2,21

Vliv přídavných dat

- Větší hindský jazykový model nepomáhá
  - Testovací data: EILMT
  - Paralelní trénovací data: EILMT + TIDES + 20k web vět
  - Trénovací data pro jazykový model:
    - EILMT + web (>300M slov): 18,82 ± 2,13
    - EILMT (181k slov): 20,07 ± 2,21
  - Mimo doménu
  - Nekompatibilní tokenizace?

Co jsme s daty nestihli

- Pavlův seznam pojmenovaných entit z Wikipedie
  - Přidali jsme do trénovacích dat, ale nemáme vyhodnoceno
- Hindská Wikipedie
  - Jen jednojazyčná data
  - Dan má staženo, jakž takž vyčištěno
  - Tokenizace zavařila TectoMT: 20k souborů, některé jen 1 věta, některé desetitisíce slov
  - Nutné rozpoznávání jazyků (angličtina, maráthí, nepálí, sanskrtská poezie ☺)

Co jsme s daty nestihli

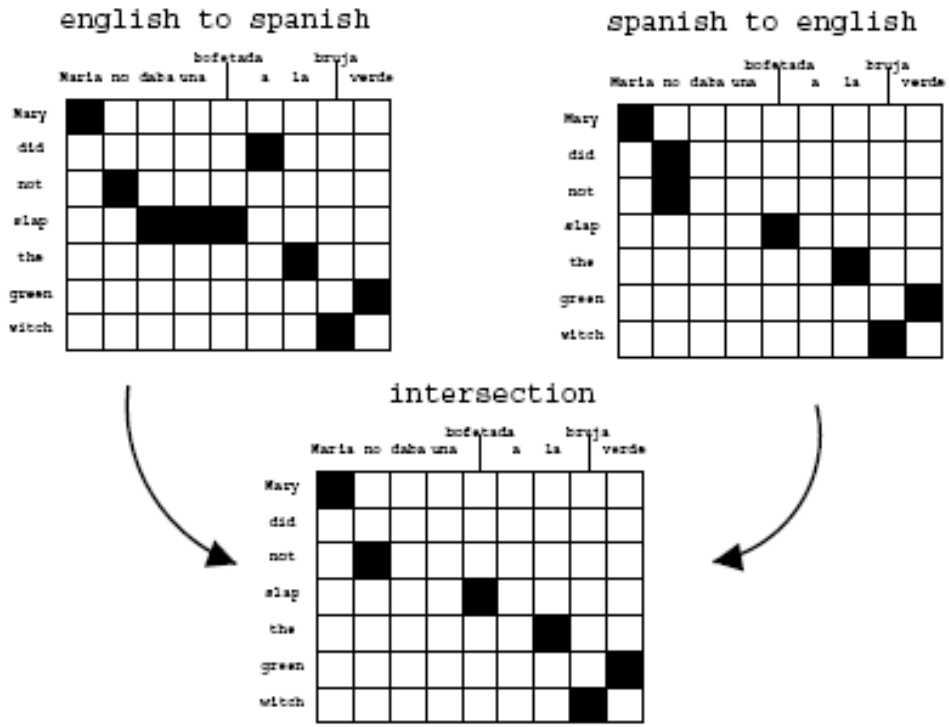
- Hindská Bible, Bhagávadgíta aj.
  - Sice mimo doménu, ale paralelní
    - Navíc velmi dobré párování vět
  - Problém s překódováním z exotických webových kódování („fontů“)
- Paralelní EMILLE
  - Hunalign nebyl schopen spárovat (neodpovídají si záhlaví souborů)
- STRAND (Philip Resnik, Noah Smith)
  - Hledá pomocí webového vyhledávače dvojice paralelních stránek

Nastavení Mosese

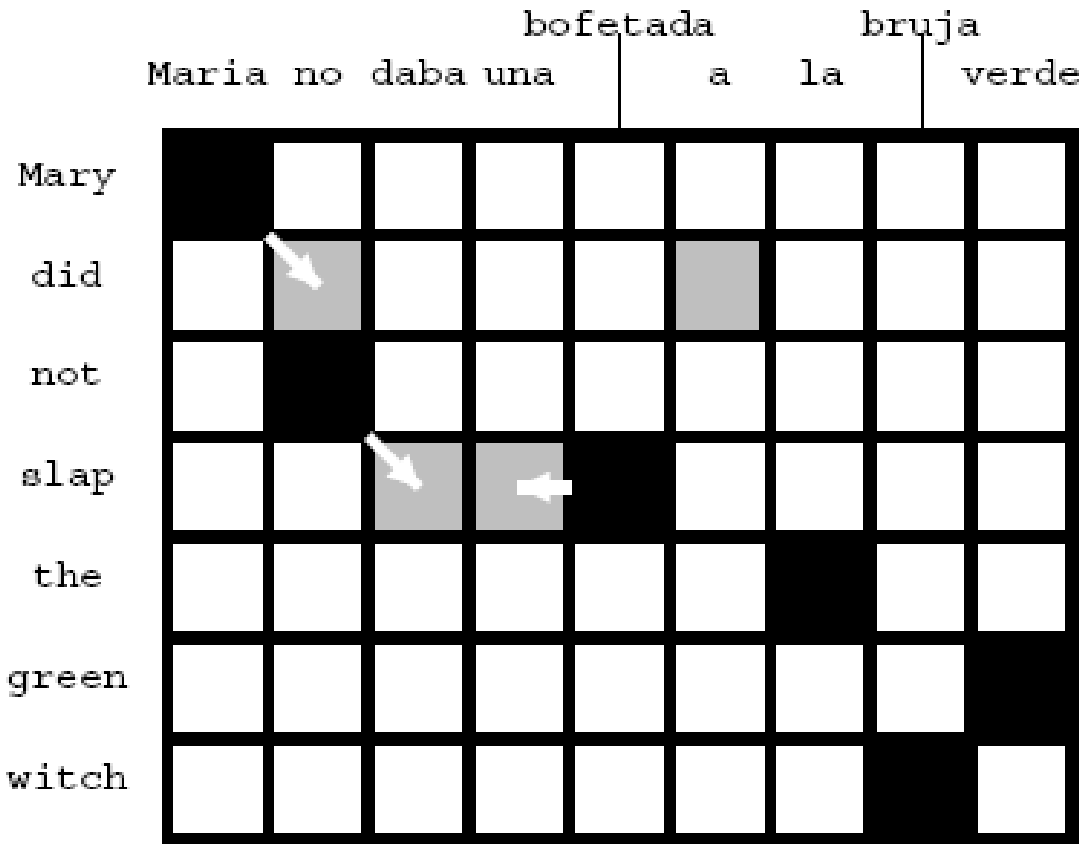
- Heuristika pro symetrizaci párování: grow-diag-final-and (GDFA)
  - 4× víc extrahovaných frází než GDF
  - Rozdíl v BLEU skóre 5 bodů! *(tabulka)*

Symetrizace párování

- Model IBM (Giza++) umí navrhnout párování 1:N
  - Funkce může vrátit stejnou hodnotu pro různé vstupy, ale pro jeden vstup vrátí jen jednu hodnotu
- Potřebujeme párování M:N



Symetrizace: „grow“



Symetrizace párování

```
GROW-DIAG-FINAL-AND(e2f,f2e):
  neighboring = ((-1,0),(0,-1),(1,0),(0,1),(-1,-1),(-1,1),(1,-1),(1,1))
  alignment = intersect(e2f,f2e);
  GROW-DIAG(); FINAL-AND(e2f); FINAL-AND(f2e);

GROW-DIAG():
  iterate until no new points added
    for english word e = 0 ... en
      for foreign word f = 0 ... fn
        if ( e aligned with f )
          for each neighboring point ( e-new, f-new ):
            if ( ( e-new not aligned or f-new not aligned ) and
                 ( e-new, f-new ) in union( e2f, f2e ) )
              add alignment point ( e-new, f-new )
FINAL-AND(a):
  for english word e-new = 0 ... en
    for foreign word f-new = 0 ... fn
      if ( ( e-new not aligned and f-new not aligned ) and
           ( e-new, f-new ) in alignment a )
        add alignment point ( e-new, f-new )
```

Symetrizace párování

- Grow-diag:
  - Nespárované slovo
  - Párovací bod v sousedství existujícího
  - Párovací bod je ve sjednocení obou párování: přidat
- Final-and:
  - Dvojice dosud nespárovaných slov
  - Párovací bod je ve sjednocení: přidat
- Co je final(-non-and)?
  - Možná stačí, aby jen jedno slovo bylo nespárované (anglické nebo hindské)

Heuristika pro párování

	EILMT	all
grow-diag-final	13,82 ± 1,46	14,67 ± 1,46
grow-diag-final-and	18,88 ± 2,05	20,07 ± 2,21

Heuristika pro párování: cs-en

	CS to EN	EN to CS
grow-diag-final	17,37 ± 0,46	14,40 ± 0,88
grow-diag-final-and	17,67 ± 0,44	14,50 ± 0,87

Nastavení Mosese

- Párování pomocí prvních 4 znaků (“light stemming”)
  - pomáhá ve spojení s GDF (nevýznamně)
  - nepomáhá s GDFA (nevýznamně)
- MERT vyvažování překladového modelu, jazykového modelu a dalších rysů
  - (tohle Indové do oficiálního baseline nezahrnuli)

Pravidla pro úpravu slovosledu

- Přesunout určitá slovesa na konec věty (ale ne přes interpunkci, “that”, WH-slova).
- Z předložek udělat záložky

- Tady se uplatnilo TectoMT
  - označkovat angličtinu Morčetem
  - rozebrat McDonaldovým MST parserem
  - na analytickou rovinu pustit přeskládávací blok

Příklad přeskládané věty

Technology is the most obvious part : the telecommunications revolution is far more pervasive and spreading more rapidly than the telegraph or telephone did in their time .

Technology the most obvious part is : the telecommunications revolution far more pervasive is and spreading more rapidly than the telegraph or telephone their time in did .

Neřízené sekání kmene a koncovky

- Faktory v Mosesovi
  - Lemma + tag: ale nemáme hindský tagger
  - Kmen + koncovka: neřízené učení morfémů

  - Danův nástroj z Morpho Challenge 2007, 2008

Základní myšlenka

- Předpoklad: jen 2 morfémy: kmen+koncovka
  - Koncovka může být prázdná
- Všechna možná dělení všech slov
  - (na kmen a koncovku)
- Vzor = soubor koncovek viděných se stejným kmenem
  - V širším smyslu vzor = sada koncovek + sada kmenů, které k nim patří

Filtrování vzorů

- Odstranit vzor, jestliže:
  - Má víc koncovek než kmenů
  - Všechny koncovky začínají stejným písmenem
  - Obsahuje pouze jednu koncovku
- Sloučit vzory A a B, jestliže:
  - B je podmnožinou A
  - A je jedinou nadmnožinou B

Příklady vzorů (en)

- Koncovky: e, ed, es, ing, ion, ions, or
- Kmeny: calibrat, decimat, equivocat, …

- Koncovky: e, ed, es, ing, ion, or, ors
- Kmeny: aerat, authenticat, disseminat, …

- Koncovky: 0, d, r, r’s, rs, s
- Kmeny: analyze, chain-smoke, collide, …

Příklady vzorů (hi)

- Koncovky: 0, ◌ा, ◌े, ◌ों
- Kmeny: अहात, खांच, घुटन, चढाव, …

- Koncovky: 0, ◌ं, ◌ंगे, गा
- Kmeny: कराए, दर्शाए, फेंके, बदले, …

- Koncovky: 0, ◌ि, ◌ियां, ◌ियों
- Kmeny: अनुभूत, अभिव्यक्त, …

Příklady vzorů (hi)

- Koncovky: 0, á, é, ón
- Kmeny: ahát, kháńč, ghutan, čadáv, …

- Koncovky: 0, n, ngé, gá
- Kmeny: karáé, daršáé, fénké, badalé, …

- Koncovky: 0, i, ián, ión
- Kmeny: anubhút, abhivjakt, …

Výstup trénování

- Seznam vzorů
- Seznam známých kmenů
- Seznam známých koncovek
- Seznam dvojic kmen-koncovka viděných spolu

- Jak tohle využít k segmentaci slova?

Morfematická segmentace

- Zkusit všechna možná dělení slova
  1. Kmen i koncovka jsou známé a dovolené dohromady
  2. Kmen i koncovka jsou známé, ale ne pohromadě
  3. Kmen je známý
  4. Koncovka je známá
  5. Obojí je neznámé

- Používáme 4 (nejdelší známá koncovka)

Vyhodnocení

- Zatím pouze BLEU score, počítal si ho každý tým sám

- Údajně se pracuje i na ručním vyhodnocení, výsledky zatím neznámé

Vliv našeho předzpracování

	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**
Suffix LM+Reord	20.09±2.18	10.18±0.74
Rule-based Reordering + Suffix LM+Reord	**21.01±2.18**	10.29±0.69

Ostatní dle shrnutí (EILMT)

Tým	Baseline	Nejlepší BLUE
Hajdarábád	0,2018	0,2260
Praha	0,1888	0,2101
Kharagpur	0,1649	0,1873
Mumbaí	0,1450	0,1751
Dublin	0,1635	0,1741

Ostatní dle jejich článků

- Hajdarábád
  - Má ve svém článku maximální BLUE 0,1998 (?!), v prezentaci už mají 0,2260
  - Odstraňování nepohodlných frází. (Možná i automatické – tečky za větou.)
  - Změna slovosledu angličtiny (Libinův dep parser)
  - Zkoušeli faktorizovaný překlad s POS taggerem
  - Anglicko-hindský slovník na neznámá slova

Ostatní dle jejich článků

- Kharagpur
  - Soustředili se na slovník (uvádějí i BLEU-1)
    - Přidali slovník do frázové tabulky
  - Faktorizovaný překlad
  - Uvádějí vyhodnocení pouze na TIDES (ta horší data)
- Mumbaí
  - Taky uvádějí jiné číslo (0,1601)
  - Změna slovosledu pomocí Stanfordského parseru

Ostatní dle jejich článků

- Dublin
  - Použili svůj systém Matrex, což je nějaká nadstavba Mosese
  - Statistická transliterace neznámých slov

Shrnutí

- Relativně slibné výsledky
  - S naprostým minimem znalostí o hindštině
  - S málo jazykovými zdroji (jen malý paralelní korpus)
  - Ve velmi krátkém čase (pár týdnů v listopadu)

- Ale
  - Jistá protekce při vylaďování Mosese
  - Je to BLUE score, subjektivní verdikt může být jiný
    - Závěr: vytahujme se, dokud je čas ☺

Co dál?

- Dodělat s daty to, co jsme minule nestihli (párování EMILLE, Wiki, Bible, STRAND)
- Slovník a transliterace na neznámá slova
- Zkontrolovat tokenizaci a normalizaci
- Přidat pořádnou morfologii / tagger
- Lepší práce se slovosledem (X ké Y)
- Hierarchický překlad (Joshua, Ašíš)
- Najmout Gaurava na analýzu chyb