Faktorované překladové modely

Základní informace

Faktorované překladové modely

- statistická metoda překladu

Faktorované překladové modely

- statistická metoda překladu
- založena na frázích (nikoliv slovo $\longrightarrow$ slovo)

Faktorované překladové modely

- statistická metoda překladu
- založena na frázích (nikoliv slovo ⟶ slovo)
- doplňková informace k tokenům (cílový i výchozí jazyk; o tom později)

Faktorované překladové modely

- statistická metoda překladu
- založena na frázích (nikoliv slovo ⟶ slovo)
- doplňková informace k tokenům (cílový i výchozí jazyk; o tom později)
  - ⟶ odtud faktorovaný SMT (multiple factors)

Faktorované překladové modely

- statistická metoda překladu
- založena na frázích (nikoliv slovo ⟶ slovo)
- doplňková informace k tokenům (cílový i výchozí jazyk; o tom později)
  - ⟶ odtud faktorovaný SMT (multiple factors)
- experimenty ukazují na podstatné zlepšení kvality překladu (podle BLEU)

Faktorované překladové modely

- statistické frázové překlady zatím nejlepší výsledky

Faktorované překladové modely

- statistické frázové překlady zatím nejlepší výsledky
- problémy při překladu *do* morf. bohatých jazyků

Faktorované překladové modely

- statistické frázové překlady zatím nejlepší výsledky
- problémy při překladu *do* morf. bohatých jazyků
  - čeština morf. velmi bohatý jazyk: podle Hajiče teoreticky **4000** tagů, reálně se užívá **2000**; angličtina běžně používá **50**

Faktorované překladové modely

- statistické frázové překlady zatím nejlepší výsledky
- problémy při překladu *do* morf. bohatých jazyků
  - čeština morf. velmi bohatý jazyk: podle Hajiče teoreticky **4000** tagů, reálně se užívá **2000**; angličtina běžně používá **50**
- problémy se řeší přidáním dodatečných informací (o tom později)

Faktorované překladové modely

- statistické frázové překlady zatím nejlepší výsledky
- problémy při překladu *do* morf. bohatých jazyků
  - čeština morf. velmi bohatý jazyk: podle Hajiče teoreticky **4000** tagů, reálně se užívá **2000**; angličtina běžně používá **50**
- problémy se řeší přidáním dodatečných informací (o tom později)
- zavedeno do Moses

Faktorované překladové modely

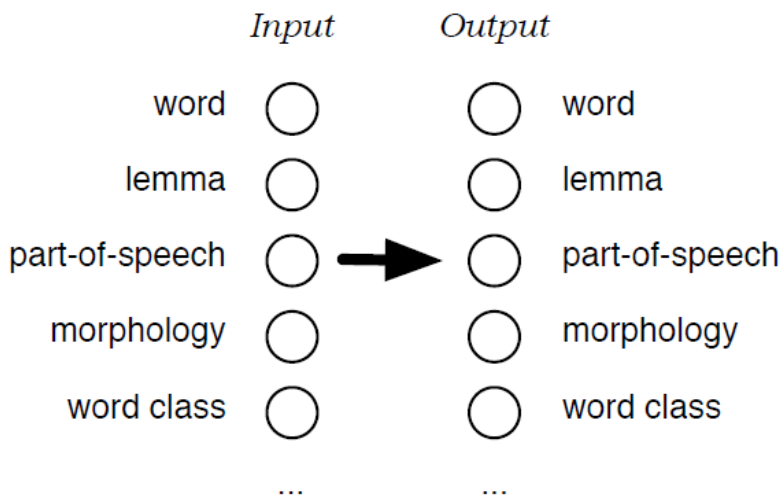
- klasické frázové překlady: překlad slovo za slovo (*house* je jiné než *houses*)

Faktorované překladové modely

- klasické frázové překlady: překlad slovo za slovo (*house* je jiné než *houses*)
- faktorované frázové překlady: přidávají dodatečnou informaci: morfologickou, syntaktickou nebo sémantickou

Faktorované překladové modely

- klasické frázové překlady: překlad slovo za slovo (*house* je jiné než *houses*)
- faktorované frázové překlady: přidávají dodatečnou informaci: morfologickou, syntaktickou nebo sémantickou



Faktorované překladové modely

- principy:

Faktorované překladové modely

- principy:
  - lepší využití trénovacích dat (při použití lemmatu místo „word“)

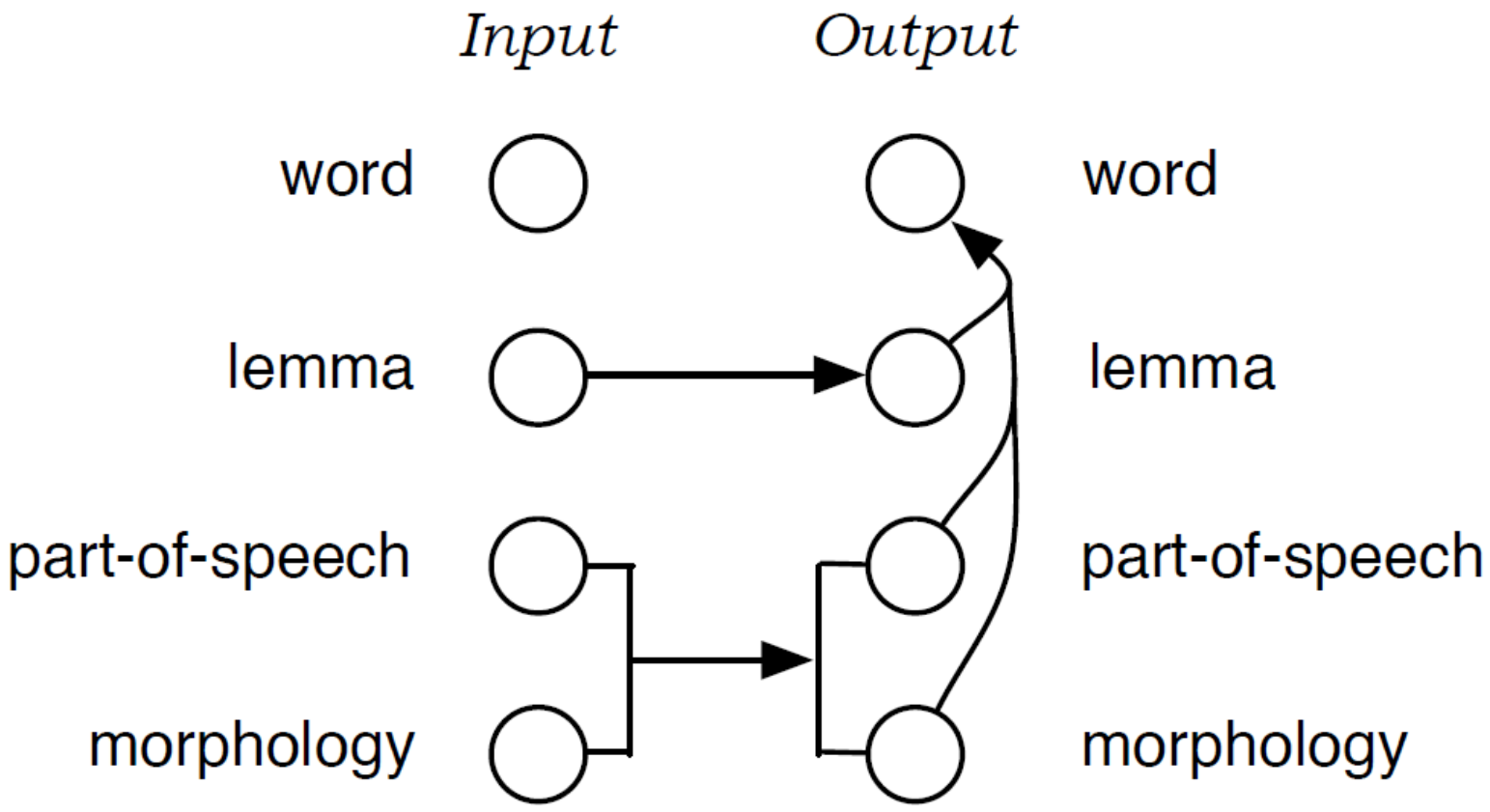
Faktorované překladové modely

- principy:
  - lepší využití trénovacích dat (při použití lemmatu místo „word“)
  - správný překlad většinou závisí právě na dodatečné informaci

Faktorované překladové modely

- principy:
  - lepší využití trénovacích dat (při použití lemmatu místo „word“)
  - správný překlad většinou závisí právě na dodatečné informaci
- v tomto pojetí tak „word“ není jen token, ale jakýsi vektor faktorů,
  kt. reprezentují různé úrovně anotace (viz násl. slajdy)

Faktorované překladové modely



Faktorované překladové modely

- překlad tak sestává ze tří částí:

Faktorované překladové modely

- překlad tak sestává ze tří částí:
  1. překlad výchozích lemmat

Faktorované překladové modely

- překlad tak sestává ze tří částí:
  1. překlad výchozích lemmat
  2. překlad výchozích morfologických charakteristik a POS

Faktorované překladové modely

- překlad tak sestává ze tří částí:
  1. překlad výchozích lemmat
  2. překlad výchozích morfologických charakteristik a POS
  3. generování cílových forem na základě 1 a 2

Faktorované překladové modely

- překlad tak sestává ze tří částí:
  1. překlad výchozích lemmat
  2. překlad výchozích morfologických charakteristik a POS
  3. generování cílových forem na základě 1 a 2

новые	дома	строятся
↓	↓	↓
new	houses	are built

Faktorované překladové modely

- překlad tak sestává ze tří částí:
  1. překlad výchozích lemmat
  2. překlad výchozích morfologických charakteristik a POS
  3. generování cílových forem na základě 1 a 2

новые	дома	строятся
↓	↓	↓
new	houses	are built

1. překlad: mapování lemmat
   - *дом* ⟶ *house, home, building, shell*

Faktorované překladové modely

- překlad tak sestává ze tří částí:
  1. překlad výchozích lemmat
  2. překlad výchozích morfologických charakteristik a POS
  3. generování cílových forem na základě 1 a 2

новые	дома	строятся
↓	↓	↓
new	houses	are built

1. překlad: mapování lemmat
   - *дом* ⟶ *house, home, building, shell*
2. překlad: mapování morfologie
   - *NN|plural-nominative-masculine* ⟶ *NN|plural, NN|singular*

Faktorované překladové modely

- překlad tak sestává ze tří částí:
  1. překlad výchozích lemmat
  2. překlad výchozích morfologických charakteristik a POS
  3. generování cílových forem na základě 1 a 2

новые	дома	строятся
↓	↓	↓
new	houses	are built

1. překlad: mapování lemmat
   - *дом* ⟶ *house, home, building, shell*
2. překlad: mapování morfologie
   - *NN|plural-nominative-masculine* ⟶ *NN|plural, NN|singular*
3. generování „vnějších“ forem
   - *house|NN|plural* ⟶ *houses*
   - *house|NN|singular* ⟶ *house*
   - *home|NN|plural* ⟶ *homes*

Faktorované překladové modely

- každá fráze je tak expandována na seznam (množinu) možných překladů:
  { *houses|house|NN|plural,*
  homes|home|NN|plural,
  buildings|building|NN|plural,
  shells|shell|NN|plural,
  house|house|NN|singular, ... }

Faktorované překladové modely

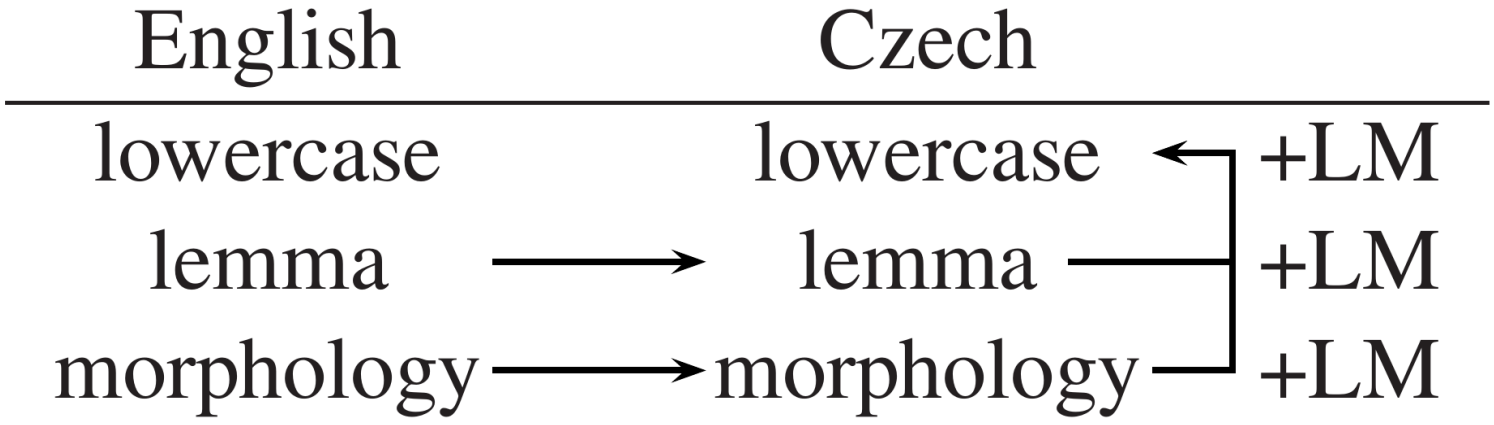
Implementace pro anglicko-český překlad
Realizoval RNDr. Ondřej Bojar, Ph.D.
v rámci workshopu SMT

Data

- Využití systému Moses
- Trénování konfigurováno volbami [3]
- Zdrojem dat je News Commentary corpus (NC)
- Cca 55 tis. párů vět [1]
- Sekce pro ladění a vyhodnocení (cca po 1000)
- Zarovnání na slova pomocí nástroje GIZA++ [2]
- Anglický text byl převeden na ”lowercase” a český lemmatizován

Scénáře faktorového překladu

- Frázový překlad
- Dekompozice a rozšíření
- 3-gramový jazykový model přes tvary slov a lemmata
- 7-gramový jazykový model přes morfologické značky
- Východisko - ”T” scénář ”Single factored” (faktor: slovní tvar)
- Úspěšnost multifaktorových scénářů



Obrázek: ”T+T+G” scénář: tři jazykové modely

Strojové učení

- Jak zvolit rysy?
- Úplné tagy
- Sloveso ”vykonat”, tvar ”vykoná”, značka: VB-S—3P-AA— [1]
- Pouze POS
- CNG03: optimalizovaný tagset
- V případě větší trénovací množiny jsou úplné tagy úspěšnější

Problémové jevy

- Příslovečná určení rozvíjející slovesa
- Lokální shoda versus chybný pád u jmenného doplnění

Translation of	Verb	Modifier
. . . preserves meaning	56%	79%
. . . is disrupted	14%	12%
. . . is missing	27%	1%
. . . is unknown (not translated)	0%	5%

Obrázek: Analýza 77 příslovečných určení rozvíjejících slovesa v 15 větách

- Výsledky parsování závislostní syntaxe
- Valence?

Zdroje

- BOJAR, O.: English-to-Czech Factored Machine Translation. In: *Proceedings of the Second Workshop on Statistical Machine Translation.* 2007, p. 232–239.
- KOEHN, P. – HOANG, H.: Factored Translation Models. In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.* 2007, p. 868–V876.

Zdroje

- Morphological Analysis of Czech Word Forms. LINDAT/CLARIN [online]. Praha: Institute of Formal and Applied Linguistics, 2015, 2012 [cit. 2015-11-09]. Dostupné z: http://quest.ms.mff.cuni.cz/morph/index.html
- GIZA++. Statistical Machine Translation [online]. Baltimore: Johns Hopkins University, 2001, leden 2001 [cit. 2015-11-09]. Dostupné z: http://www.statmt.org/moses/giza/GIZA++.html
- Factored Training. Moses [online]. Edinburgh: University of Edinburgh, 2015, červenec 2013 [cit. 2015-11-09]. Dostupné z: http://www.statmt.org/moses/?n=FactoredTraining.FactoredTraining

Experiment a jeho výsledná evaluace

- Moses se základním nastavením
- využití evaluačního algoritmu BLEU

1) Syntakticky obohacený výstup

- Přeložení "surface forms of words", přidání lexikálních faktorů
- Implementování morfologické a mělké syntaktické analýzy
- Získáme sekvenční model podobný n-gramům
- Podpora syntaktické koherence na výstupu
- Použité modely: Eng-Ger, Eng-Sp, Eng-Cz, Eng-Chin

English-German

- Europarl korpus, 750 tis. vět
- Přidání sl. druhu a morf. analýzy na výstupu a využití 7-gramů přineslo zlepšení (0,18%)
- "Baseline systém" se neuměl vypořádat s určitými i neurčitými členy ve větě

English-German

Model	**BLEU**
Best published result	18,15%
Baseline (surface)	18,04%
Surface+ POS	18,15%
Surface+ POS+ Morph	18,22%

English-Spanish

- Europarl korpus, 40 tis. vět
- Použití sl. druhu a morfologické analýzy na výstupu a 7-gramového sekvenčního modelu přineslo zlepšení o 1,25% (morph) a 0,84% (morph+ kat)

English-Spanish

Model	**BLEU**
Baseline (surface)	23,41%
Surface+ morph	24,66%
Surface+ morph+ kat	24,25English-Spanish%

English-Czech

- Wall Street journal, 20 tis. vět
- Využití morfologické analýzy a 7-gramového jazykového modelu
- Potřeba zvážit, které morfologické rysy využít
- Všechny modely předčily základní variantu

English-Czech

Model	**BLEU**
Baseline (surface)	25,82%
Surface+ all morph	27,04%
Surface+ case/ number/ gender	27,45%
Surface+ CNG/ verb/ preposition	27,62%

2) Morfologická analýza a generování

- Místo surface přeložíme lemma a morfologii a vytvoříme surface na výstupu
- Experiment proveden na English-German
- News commentary korpus, 52 tis. vět
- Německá morfologická a slovnědruhová analýza: LoPar Schmitd and Schulte im Walde (2000)
- Anglická slovnědruhová analýza: Brill's tagger (Brill, 1995)

2) Morfologická analýza a generování II

- Při použití slovnědruhové analýzy- zlepšení o 0,86%
- Lemma+ morfologická analýza- propad
- Vytvoření vybíracího modelu
- Pokud není v trénovacich datech výskyt surface, pak užijeme generování

German-English

Model	**BLEU**
Baseline (surface)	18,19%
Surface+ POS	19,05%
Lemma/ morph	14,46%
Vybírací model	19,47%

3) Použití automatických slovních druhů

- Automaticky trénované rozdělení do sl. druhů shlukováním kontextové podobnosti
- Zlepšení o 1,25%

English-Chinese

Baseline (surface)	19,54%
Surface+ word class	21,10%

4) Integrovaný recasing

- Recasing= různá podoba zápisu: the, The, THE
- V SMT- minuskulní písmo, potřeba přidat krok k navrácení původní podoby
- Lze integrovat do modelu

Chinese-English

Standart two-pass SMT+ recase	20,65%
Integrated factored model (optimized)	21,08%