VERBESSERUNG DER STATISTISCHEN MASCHINELLEN ÜBERSETZUNG DURCH MORPHOLOGISCHE ANALYSE

Sharon Goldwater & David McClosky

Sarah Hartmann

13.01.2015

Advanced Topics in Statistical Machine Translation

AGENDA

- Einführung
- Modelle
- Experimente
- Diskussion

AGENDA

Einführung

- Das Problem
- Der Lösungsvorschlag

Modelle

Experimente

Diskussion

EINFÜHRUNG: DAS PROBLEM

Ich trage die Jacke meiner Schwester.

I wear the jacket of my sister.

trage	→	wear
trägst	→	wear
trägt	→	wears
tragen	→	wear
tragt	→	wear
tragen	→	wear
5		2

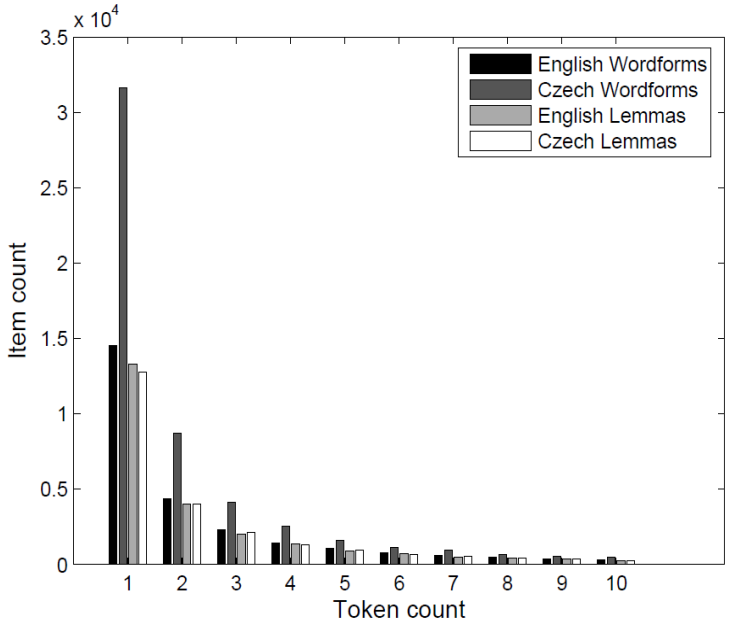
Pro-drop Sprachen:
Pronomen können weggelassen werden, da diese Information im Verb steckt

EINFÜHRUNG: DAS PROBLEM

- Ziel: wahrscheinlichste Übersetzung finden von Fremdsprache *f* (Tschechisch) in gewünschte Sprache *e* (Englisch)
- $\underset{e}{\operatorname{argmax}} P(e|f) = \underset{e}{\operatorname{argmax}} P(e)\, P(f|e)$

  Noisy-Channel-Modell — Sprachmodell & Übersetzungsmodell
- Das Übersetzungsmodell bestimmt die Qualität der Übersetzung
- Parallele Korpora für Training des Übersetzungsmodells häufig nicht verfügbar

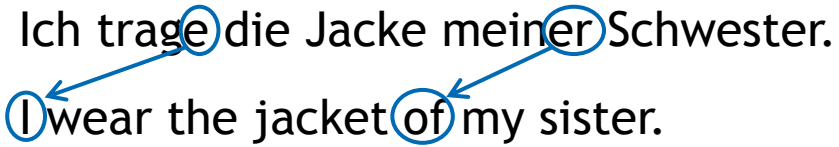
EINFÜHRUNG: DAS PROBLEM



Anzahl der Items y, die im Parallelen Korpus mit einem Token Count von x vorkommen (Goldwater & McClosky, 2005).

EINFÜHRUNG: DER LÖSUNGSVORSCHLAG

- Morphologische Analyse
  - Spärliche Daten können reduziert werden
  - Ähnlichkeit zwischen den Sprachen kann erhöht werden
  - Übersetzungsqualität wird verbessert
- Input (hoch flektierend) → Output (geringe Flexion)



AGENDA

- Einführung
- **Modelle**
  - Lemmas
  - Pseudowörter
  - Modifizierte Lemmas
  - Morpheme
- Experimente
- Diskussion

MODELLE: LEMMA

```
Pro/pro/RR--4----------
někoho/někdo/PZM-4----------
by/být/Vc-X---3-------
její/jeho/PSZS1FS3-------
provedení/provedení/NNNS4-----A----
mělo/mít/VpNS---XR-AA---
smysl/smysl/NNIS4-----A----
./Z:-------------
```

Ein Satz vom PCEDT Korpus (Goldwater & McClosky, 2005).

MODELLE: LEMMA

- Sehr einfache Methode die Input Daten zu verändern, ist Wortform durch sein Lemma zu ersetzen: trag - ~~e~~
- Problem: Informationsverlust
- Darum zwei weitere Varianten:
  - Nur bestimmte Wortformen lemmatisieren (z. B. Substantive, Verben, und Pronomen tragen Flexionen im Englischen, darum alle anderen Wortarten lemmatisieren)
  - Nur die am wenigsten vorkommenden Wörter werden lemmatisiert (nur noch die Wortforminformationen von den häufiger vorkommenden Wörtern)

MODELLE: PSEUDOWÖRTER

- Tschechisch ist eine **pro-drop Sprache**, daher Personentags gut als Pseudowörter geeignet
- Negationen sind im Tschechischen oft im Verb enthalten darum gute Kandidaten für Pseudowörter, ebenso Kasusbezeichner

Words:	Pro ňekoho by její provedení mělo smysl .
Lemmas:	pro ňekdo být jeho provedení mít smysl .
Lemmas+Pseudowords:	pro ňekdo být PER_3 jeho provedení mít PER_X smysl .

MODELLE: MODIFIZIERTE LEMMAS

- Bei Vergangenheitsformen werden diese oft auch in englischer Flexion repräsentiert, darum könnte es hilfreich sein Pseudowörter an Lemmas anzuhängen, statt separat zu halten
- Vorteil: man hat die lemmatisierte Form + morphologische Informationen
- Es wird erwartet, das Numerus Marker am Substantiv und Zeitmarker an Verben am besten funktionieren

Words:	Pro ňekoho by její provedení mělo smysl .
Lemmas:	pro ňekdo být jeho provedení mít smysl .
Lemmas+Pseudowords:	pro ňekdo být PER_3 jeho provedení mít PER_X smysl .
Modified Lemmas:	pro ňekdo být+PER_3 jeho provedení mít+PER_X smysl .

MODELLE: MORPHEME

- Neues Übersetzungsmodell
- Kompositionelle Struktur für f_j, sodass $f_j = f_{j0} \ldots f_{jk}$ wobei f_{j0} das Lemma und der Rest Morpheme darstellen, die zu dem jeweiligen f_j gehören
- So werden Wörter in Morpheme zerlegt und Morpheme in f_j werden unabhängig generiert bedingt durch e_i

$$\mathrm{P}(f_j|e_i) = \prod_{k=0}^{K} \mathrm{P}(f_{jk}|e_i)$$

- Englisches Wort, das oft zu Tschechischem Wort mit bestimmtem Tag aligniert wurde, aligniert auch mit höherer Wahrscheinlichkeit zu einem anderen Wort mit diesem Tag

AGENDA

- Einführung
- Modelle
- **Experimente**
  - Korpora und Toolkits
  - Lemmas
  - Pseudowörter
  - Modifizierte Lemmas
  - Morpheme
  - Kombiniertes Modell
- Diskussion

EXPERIMENTE: KORPORA UND TOOLKITS

- **Prague Czech-Englisch Dependency Treebank (PCEDT)**
  - Tschechische Wörter bereits mit morphologischen Informationen annotiert
  - Paralleler Korpus mit Aufteilung in Trainings-, Entwicklungs-, und Testdaten
- Sprachmodell wurde mit dem **CMU Statistical Language Modelling Toolkit** trainiert
- Übersetzungsmodell wurde mit **GIZA++** trainiert
- **ISI ReWrite Decoder** für die Produktion der Übersetzungen

EXPERIMENTE: LEMMA

	dev	test
word-to-word	.311	.270
lemmatize all	.355	.299
except Pro	.350	
except Pro, V, N	.346	
lemmatize n < 50	.370	.306
truncate all	.353	.283

BLEU Scores für die word-to-word Baseline, Lemmatisierung, und Wort-Trunkierung (Goldwater & McClosky, 2005).

EXPERIMENTE: PSEUDOWÖRTER

Tag Typ	Pseudowörter
PER	.365
TEN	.365
PER,TEN	.355
NUM	.354
CASE	.353
NEG	.357

Scores stammen vom Entwicklungsdaten Set.
Unterschiede von .009 sind signifikant ($p < .05$).

BLEU Scores für die Einbeziehung unterschiedlicher Informationen von morphologischen Tagklassen (Goldwater & McClosky, 2005).

EXPERIMENTE: PSEUDOWÖRTER

Tag class	Count	Avg/sentence
PER	49700	2.35
TEN	47744	2.26
past	22544	1.07
pres	20291	0.96
fut	1707	0.08
‘any’	3202	0.15
NUM	151646	7.17
CASE	151646	7.17
NEG	3326	0.16

Auftretens Häufigkeit für jede Tagklasse in den Tschechischen Trainingsdaten (Goldwater & McClosky, 2005).

EXPERIMENTE: MODIFIZIERTE LEMMA

Tag Typ	Pseudowörter	Modifiziertes-Lemma
PER	.365	.356
TEN	.365	.361
PER,TEN	.355	.362
NUM	.354	.367
CASE	.353	.340
NEG	.357	.356

Scores stammen vom Entwicklungsdaten Set. Unterschiede von .009 sind signifikant ($p < .05$).

BLEU Scores für die Einbeziehung unterschiedlicher Informationen von morphologischen Tagklassen (Goldwater & McClosky, 2005).

EXPERIMENTE: MORPHEME

Tag Typ	Pseudowörter	Modifiziertes-Lemma	Morpheme
PER	.365	.356	.356
TEN	.365	.361	.364
PER,TEN	.355	.362	.355
NUM	.354	.367	.361
CASE	.353	.340	.337
NEG	.357	.356	.353

Scores stammen vom Entwicklungsdaten Set. Unterschiede von .009 sind signifikant ($p < .05$).

BLEU Scores für die Einbeziehung unterschiedlicher Informationen von morphologischen Tagklassen (Goldwater & McClosky, 2005).

EXPERIMENTE: KOMBINIERTES MODELL

- Vorschlag: Kombination des Pseudowortansatzes (Person und Negation) mit dem modifizierten Lemma Ansatz (Zeit und Geschlecht)
- BLEU Score von .390 (development) and .333 (test)
- Schneidet damit besser ab als alle anderen Modelle in den vorherigen Experimenten

AGENDA

- Einführung
- Modelle
- Experimente
- **Diskussion**
  - Vorschläge der Autoren
  - Kritik

DISKUSSION: VORSCHLÄGE DER AUTOREN

- Ob Pseudowörter oder Modifiziertes Lemma besser funktionieren, hängt davon ab wie die jeweilige Tagklasse im Englischen ausgedrückt wird (Funktionswort oder Flexion)
- **Ziel der Morphologischen Analyse ist es Tschechische Daten dem Englischen ähnlicher zu machen**
- Einbindung syntaktischer Informationen könnte hilfreich sein
- Vorschlag: Erweiterung des Morphembasierten Übersetzungsmodells (wenn beide Sprachen hoch flektierend sind)

$$\mathrm{P}(f_j|e_i) = \prod_{k=0}^{K} \mathrm{P}(f_{jk}|e_{ik})$$

DISKUSSION: KRITIK

- Korpus enthält bereits morphologische Informationen
- Korpusgröße entscheidend für Ergebnisse
- Methoden nur sinnvoll bei hoher Flexion
- Tschechisch ist eine der Sprachen mit stärkster Flexion

Vielen Dank für die Aufmerksamkeit! ☺

QUELLE

Goldwater, S. & McClosky, D. (2005). Verbesserung der Statistischen Maschinellen Übersetzung mittels Morphologischer Analyse. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language*, (pp. 676–683). Vancouver: Association for Computational Linguistics.