HOW TO AVOID BURNING DUCKS

EIN KOMBINIERTER ANSATZ ZUR VERARBEITUNG VON KOMPOSITA IM DEUTSCHEN

nach Fritzinger & Fraser

GLIEDERUNG

- Hintergrund
- SMOR
- Korpusbasierter Ansatz
- Kombinierter Ansatz
- Evaluation
- Fazit

HINTERGRUND

- Viele Komposita im Deutschen (Kindergarten, Haustürschlüssel etc.)
- Beliebig lang
  - Donaudampfschifffahrtselektrizitätenhauptbetriebswerkbauunterbeamtengesellschaft
- Splitting von Komposita zur automatischen Übersetzung
- Herausforderung: Finden der richtigen Splittpunkte (Notation: ‚|')
  - Aktionsplan → Aktion | Plan
    → *Akt | Ion | Plan

HINTERGRUND - GRUNDLEGENDE ANSÄTZE

Linguistische Analysen

- SMOR
- Morphologische Analyse von Wörtern

Korpusbasierte Ansätze

- Koehn & Knight, 2003
- Splittung von Komposita
- Basis: Worthäufigkeiten

GLIEDERUNG

- Hintergrund
- SMOR
- Korpusbasierter Ansatz
- Kombinierter Ansatz
- Evaluation
- Fazit

SMOR

- Stuttgarter Morphologisches Analysewerkzeug
- Wortbildung = Konkatenation von Morphemen (Affixe, Wortstämme)
- Berücksichtigung deutscher Wortbildungsprozesse
  - Wortbeugung → ein schönes Kind, eine schöne Frau, ein schöner Mann
  - Derivation → Frei-heit, mach-bar
  - Komposita
- Basis: Lexikon

Lexikon-Eintrag	Anzahl
Flexionsstammformen	47.671
Kompositionsstammformen	528
Derivationsstammformen	1.691
Präfixe	323
Infixe	0
Suffixe	208

Quelle: http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/SMOR/index.html

SMOR - BEISPIEL

Output der Analyse zu „Durchschnittsauto“:

```
analyze> Durchschnittsauto
Durchschnitt<NN>Auto<+NN><Neut><Nom><Sg>
durch<VPART>schneiden<V><NN><SUFF>Auto<+NN><Neut><Nom><Sg>
```

GLIEDERUNG

- Hintergrund
- SMOR
- Korpusbasierter Ansatz
- Kombinierter Ansatz
- Evaluation
- Fazit

KORPUSBASIERTER ANSATZ - DIE IDEE

- Koehn & Knight (2003)
- Methode zur Teilung von Komposita ohne morphologische Analyse
- Teilung der Komposita in Parts bzw. bekannte Wörter
  - Mittels monolingualem Trainingskorpus
  - Kommt ein Wort im Korpus vor? Wie oft?

KORPUSBASIERTER ANSATZ - DIE IDEE

- Beachtung verschiedener Aspekte
  - Fugenelemente (Inflation**s**rate, Freund**es**kreis)
  - Löschelemente (Kirchturm → Kirch**e**|Turm)
- Vermeidung falscher Splits
  - Entfernung von Wörtern, die seltener als dreimal im Korpus vorkommen
  - Stoppliste (Funktionswörter, Namen + erscheinen oft als falschen Splittings, wie z.B. den, der, eng, che….)

KORPUSBASIERTER ANSATZ - AUSWAHL DER SPLITTPUNKTE

- Wie wird der richtige Splittpunkt ausgewählt?
  - Basierend auf Worthäufigkeiten im Trainingskorpus
  - Auswahl der Splittmöglichkeit mit größtem geometrischen Mittel

$$argmax\ S\left(\prod_{p_i \in S} count(p_i)\right)^{\frac{1}{n}}$$

S = Split, p_i = part, n = number of parts

KORPUSBASIERTER ANSATZ - BEISPIEL I

$$argmax\ S(\prod_{p_i \in S} count(p_i))^{\frac{1}{n}}$$

S = Split, p_i = part, n = number of parts

- Zerteilung des Kompositums „Gummiboot“
- Mögliche Splittoptionen:

gummiboot(332) → 332

gummi(456) | boot(305) → $(456 * 305)^{\frac{1}{2}}$ = **372,93**

Richtiges Splitting gewählt

KORPUSBASIERTER ANSATZ - BEISPIEL 3

$$argmax\ S(\prod_{p_i \in S} count(p_i))^{\frac{1}{n}}$$

S = Split, p_i = part, n = number of parts

- Zerteilung des Kompositums „Vielfliegerbonus“
- Mögliche Splittoptionen:

vielfliegerbonus (5) → 5

viel(689) | flieger(567) | bonus(348) → 514,2

viel(689) | flieg(247) | er(4078) | bonus(348) → 701,03

vielflieger(56) | bonus(348) → 139,6

Richtiges Splitting erkannt, aber nicht gewählt, da „er“ hochfrequentiert

KORPUSBASIERTER ANSATZ - FEHLERANFÄLLIGKEIT

- Frequenzbasierte Fehler
  - lebenstreuen → Leben|streuen
  - traumatisch → Trauma|Tisch
- Fehler durch Lösch- und Fugenelemente
  - entbrannte → Ente|brannte
  - Belangen → Bela|Gen

GLIEDERUNG

- Hintergrund
- SMOR
- Korpusbasierter Ansatz
- Kombinierter Ansatz
- Evaluation
- Fazit

KOMBINIERTER ANSATZ - DIE IDEE

- Kombination des korpusbasierten Ansatzes und SMOR
  - Splittpunkte durch SMOR
  - Suche nach Worthäufigkeiten im Korpus
  - Bestimmung der besten Splittung

→ Linguistisch motivierter, korpusbasierter Ansatz

KOMBINIERTER ANSATZ - DIE IDEE

- keine Stoppliste mehr nötig
- keine minimale Länge eines Parts mehr nötig

durch SMOR bereits abgedeckt

linguistisches Wissen

- Beispiel:
  - Splittung nur in freie Morpheme (Wortstämme, trennbare Teile)
  - „aufgeben“ → auf | geben
    „begraben“ → *be | graben
  - „auf“ kann auch ohne Verb auftreten, „be“ jedoch nicht

KOMBINIERTER ANSATZ - ERWEITERUNGEN I

- SMOR lemmatisiert
- Zurücktransformierung der Lemmatisierung
- Information über mögliche Splittpunkte benötigt

- Beispiel:
  - SMOR: Beitritt | Land
  - Hier: Beitritt | Länder

KOMBINIERTER ANSATZ - ERWEITERUNGEN 2

- Rekombinationen bei mehr als zwei Teilen
- Beispiel: „Wortbildungsarten“
  1. SMOR-Analyse (ohne Lemmatisierung):

     Wort | Bildung | Arten
  1. Zusätzliche Suche im Korpus nach Rekombinationen:

     Wortbildung | Arten

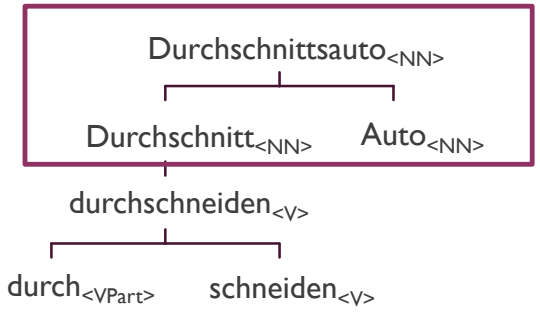
     Wort | Bildungsarten

KOMBINIERTER ANSATZ - ERWEITERUNGEN 3

- Für Zwecke der SMT sind keine tiefen, morphologischen Analysen nötig
- Ziel: für jeden Teil einen Korrespondent in Zielsprache finden

→ Splittung auf höchster Ebene genügt

- Beispiel:



KOMBINIERTER ANSATZ - EINSCHRÄNKUNGEN

1. Nur Splitts mit minimaler Anzahl an Teilen
   - Problem:
     - lexikalisierte Komposita
       - Im Lexikon von SMOR als freie Morpheme („Geländewagen“)
   - Lösung:
     - Beide Varianten werden beibehalten
2. Eine Variante, die nur Nomen splittet
   - Identifizierung der Nomen durch POS-Tagger

→ 3 Varianten

Notation:

- Uneingeschränkte Variante: **sm**
- Minimale Anzahl an Parts: **smc**
- Nur Nomen splitten: **@nn**

GLIEDERUNG

- Hintergrund
- SMOR
- Korpusbasierter Ansatz
- Kombinierter Ansatz
- Evaluation
- Fazit

EVALUATION

- Zwei Goldstandard-Evaluationen
  - Linguistisch motivierter Goldstandard
  - One-to-one Correspondence Goldstandard
- Trainingsdaten: MT Arbeit der ACL Machine Translation Workshop 2009

EVALUATION - TERMINOLOGIE

- **raw** = Baseline ohne Splitting
- **cd** = Korpusbasiertes Splitting
- **sm** = Kombinierter Ansatz mit allen durch SMOR ermittelten Optionen (uneingeschränkte Variante)
- **smc** = Kombinierter Ansatz mit minimaler Anzahl an Parts
- **@nn** = Nur Nomen splitten

- **correct split**: soll gesplittet werden und wurde richtig gesplittet
- **correct not**: soll nicht gesplittet werden und wurde nicht gesplittet
- **wrong split**: soll nicht gesplittet werden, aber wurde gesplittet
- **wrong not**: soll gesplittet werden, aber wurde nicht gesplittet
- **wrong faulty (fty)**: soll gesplittet werden, wurde auch gesplittet, aber falsch

- **precision**: $\frac{\text{correctsplit}}{\text{correctsplit+wrongfaulty+wrongsplit}}$ → werden nur die gesplittet, die es auch sollen?
- **recall**: $\frac{\text{correctsplit}}{\text{correctsplit+wrongfaulty+wrongnot}}$ → wie viel % der Komposita werden richtig gesplittet?
- **accuracy**: $\frac{\text{correct}}{\text{correct+wrong}}$ → wie viel % der Wörter wird richtig analysiert?

EVALUATION - LINGUISTISCH MOTIVIERTER GOLDSTANDARD

- Linguistisch motivierte Splittpunkte
- Testdaten beinhalten 6187 verschiedene Typen
- Annotieren der plausibelsten Splittpunkte durch Muttersprachler auf höchster Hierarchieebene

- Erlaubte Splitts:
  - Wortstämme, trennbare Partikel
- Unerlaubte Splitts:
  - Gebundene Morpheme (Präfixe, Suffixe)

EVALUATION - LINGUISTISCH MOTIVIERTER GOLDSTANDARD

Ergebnisse:

	Correct		Wrong			Metrics		
	split	not	split	not	fty	Precision	Recall	Accuracy
raw	0	5073	0	1114	0	-	0,00%	81,99%
cd	679	4192	883	120	313	36,21%	61,06%	78,73%
sm	912	4534	541	35	165	56,37%	82,01%	88,02%
sm@nn	628	4845	230	337	147	62,49%	56,73%	88,46%
smc	884	4826	249	135	93	72,10%	79,50%	92,29%
smc@nn	648	4981	94	380	84	78,45%	58,27%	90,98%

EVALUATION - ONE-TO-ONE CORRESPONDENCE GOLDSTANDARD

- Manuelle Übersetzung der Komposita → kompositionelle Übersetzung (5000 Tokens)
- Übersetzung muss nicht konsistent sein
  - Verschiedene Kontexte → verschiedene Übersetzungen → unterschiedliche Splitting
  - Beispiel: Zauberei

- Folge: linguisitsch richtiger Splitt muss nicht als korrekt angesehen werden

EVALUATION - ONE-TO-ONE CORRESPONDENCE GOLDSTANDARD

Ergebnisse:

	Correct		Wrong			Metrics		
	split	not	split	not	fty	Precision	Recall	Accuracy
raw	0	4845	0	155	0	-	0,00%	96,90%
cd	81	4435	404	14	59	14,89%	52,60%	90,32%
sm	112	4563	283	8	34	26,11%	72,73%	93,50%
sm@nn	107	4677	169	15	32	34,74%	69,48%	95,68%
smc	128	4666	180	12	14	39,75%	83,12%	95,88%
smc@nn	123	4744	102	18	13	51,68%	79,87%	97,34%

EVALUATION - FEHLERANFÄLLIGKEIT

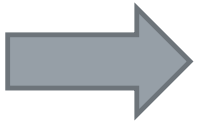
- Abhängigkeit von SMOR's lexikalischer Abdeckung
- Abhängigkeit von der Qualität des genutzten Korpus

- Ca. 97% der Fehler sind auf die Worthäufigkeiten zurückzuführen
- Fehler durch falsche SMOR Analysen nur sehr gering

EVALUATION - ÜBERSETZUNGSPERFORMANCE

- Phrasenbasiertes Übersetzungsmodell
- 227 Millionen Wörter für die Frequenzberechnung
- Daten aud EACL 2009 Workshop

System	Test BLEU
raw	15,72
cd	16,17
sm	16,59
sm@nn	16,76
smc	16,63
smc@nn	16,40



Alle sm*-Varianten sind leistungsstärker als cd (bis zu +0,59 BLEU)

GLIEDERUNG

- Hintergrund
- SMOR
- Korpusbasierter Ansatz
- Kombinierter Ansatz
- Evaluation
- Fazit

FAZIT

- Kombination aus linguistischem Wissen und korpusbasiertem Ansatz…
  - …kann die Anzahl korrekter Splitts erhöhen
  - … kann Oversplitting reduzieren
- Die kombinierten Ansätze erzielen bei beiden Evaluationen die besten Ergebnisse

→ Hilfe zur Verbesserung von Ansätzen, die mit Komposita umgehen müssen (Sprachverarbeitung, Information Retrieval)

VIELEN DANK!!!

REFERENZEN

- Fritzinger, F. & Fraser, A. (2010). *How to Avoid Burning Ducks: Combining Linguistic Analysis and Corpus Statistics for German Compound Processing*, WMT 2010.
- Koehn, P. & Knight, K. (2003). Empirical methods for compound splitting. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, 187–193, Morristown, NJ, USA.
- Stuttgarter Morphologisches Analysewerkzeug (SMOR): Dokumentation, retrieved from http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/SMOR/index.html.