

Top a Splitter

Using Distributional Semantics for Improving Compound Splitting

Patrick Ziering[♣] Stefan Müller[♣] Lonneke van der Plas[◇]

[♣]Institute for NLP, University of Stuttgart, Germany

[◇]Institute of Linguistics, University of Malta, Malta

MWE 2016

Introduction

Compound splitting:

Decomposition of closed compounds

Hühnersuppe (chicken soup)

→ *Hühner* | *suppe* ⇒ *Huhn* + *Suppe*

Common approach of corpus-based splitters:

- Generate all possible splits
- Rank splits according to corpus frequency

Neglecting semantic compatibility:

⇒ Most frequent constituents are not always plausible

Eidotter (egg yolk) → *Eid* + *Otter* (oath otter)

Distributional semantics indicating compositionality:

- Recent work: [Weller et al., (2014)]
- ⇒ Splitting of only compositional compounds
- ⇒ **distributional similarity** (DS) between compound and its constituents
- ⇒ No improvements in extrinsic evaluation (SMT)

Distributional semantics for compound splitting:

We enrich a compound splitter's ranked output with DS

- $\cos(\vec{Eidotter}, \vec{Otter}) < \cos(\vec{Eidotter}, \vec{Dotter})$
- ⇒ *Ei* + *Dotter* (egg yolk)

Method

Ranking score	Candidate split	Correct?
14264	<i>Fisch</i> + <i>Zeugnis</i> 'fish certificate'	✗
9390	<i>Fisch</i> + <i>Erzeugnis</i> 'fish product'	✓
5387	<i>Fischer</i> + <i>Zeugnis</i> 'fisherman certificate'	✗

constituent-wise
cosine similarity

Compound Ω	Constituent ω	Cosine sim(Ω, ω)
<i>Fischerzeugnis</i>	<i>Fisch</i>	0.46
<i>Fischerzeugnis</i>	<i>Erzeugnis</i>	0.10
<i>Fischerzeugnis</i>	<i>Fischer</i>	0.03
<i>Fischerzeugnis</i>	<i>Zeugnis</i>	0.01

combination using geometric mean

Re-ranking score	Candidate split	Correct?
9390 · 0.22 ≈ 2034	<i>Fisch</i> + <i>Erzeugnis</i> 'fish product'	✓
14264 · 0.05 ≈ 709	<i>Fisch</i> + <i>Zeugnis</i> 'fish certificate'	✗
5387 · 0.01 ≈ 70	<i>Fischer</i> + <i>Zeugnis</i> 'fisherman certificate'	✗

re-ranking

Candidate split	GEO
<i>Fisch</i> + <i>Erzeugnis</i>	$\sqrt{0.46 \cdot 0.10} \approx 0.22$
<i>Fisch</i> + <i>Zeugnis</i>	$\sqrt{0.46 \cdot 0.01} \approx 0.05$
<i>Fischer</i> + <i>Zeugnis</i>	$\sqrt{0.03 \cdot 0.01} \approx 0.01$

Experiments

Tools and data:

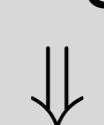
- German Wikipedia corpus (665M tokens)
- GermaNet test set (51K compounds)
- TreeTagger

Distributional model:

- window: 20 words to the left and right
- 20K most frequent nominal co-occurents

Three splitters:

knowledge-poor



knowledge-rich

Accuracy	SPAcc					NormAcc				
	MOD	HEAD	GEO	MULT	ADD	MOD	HEAD	GEO	MULT	ADD
[ZIERING AND VAN DER PLAS (2016)] Cov.: 99.9%										
INITIAL	97.5%					87.4%				
RR_{DS}	93.6%	94.6%	95.4%	92.7%	92.0%	75.9%	84.7%	77.8%	69.6%	61.2%
RR_{ALL}	97.5%	97.9%†	98.0%†	97.8%†	98.0%†	88.6%†	87.7%†	89.0%†	88.5%†	88.7%†
[WELLER AND HEID (2012)] Cov.: 97.6%										
INITIAL	98.1%					90.4%				
RR_{DS}	96.9%	97.0%	97.7%	96.9%	95.8%	86.5%	89.3%	87.1%	81.8%	75.3%
RR_{ALL}	98.2%†	98.2%†	98.3%†	98.2%†	98.3%†	91.3%†	90.5%†	91.1%†	90.9%†	90.9%†
[FRITZINGER AND FRASER (2010)] Cov.: 93.6%										
INITIAL	98.4%					94.9%				
RR_{DS}	97.9%	97.9%	98.4%	98.3%	98.2%	94.3%	94.3%	94.7%	94.5%	94.3%
RR_{ALL}	98.4%	98.3%	98.5%	98.4%	98.4%	94.8%	94.7%	95.0%	94.8%	94.7%

INITIAL: splitter without DS

RR_{DS}: DS-only baseline

RR_{ALL}: DS in addition

MOD: $\cos(\vec{compound}, \vec{modifier})$

HEAD: $\cos(\vec{compound}, \vec{head})$

GEO: geometric mean(MOD, HEAD)

MULT: $\cos(\vec{compound}, \vec{modifier} \cdot \vec{head})$

ADD: $\cos(\vec{compound}, \vec{modifier} + \vec{head})$

SPAcc: correct split point

NormAcc: correct split point and correctly normalized constituents

Observations

RR_{ALL} always outperforms **INITIAL** in at least one DS mode

Knowledge-poor methods benefit most from re-ranking

→ misleading normalizations are demoted

both corpus frequency and **DS** are crucial features

→ each feature in isolation lacks information

Examples

putzen + *Lappen* **RR_{ALL}** ✓ vs. *Putz* + *Lappen* **INITIAL** ✗

Wanderzirkus 'traveling circus'

wandern + *Zirkus* **RR_{ALL}** ✓ vs. *Wand(er)* + *Zirkus* **INITIAL** ✗

Blinddarmoperation 'appendix operation'

Blinddarm + *Operation* **RR_{ALL}** ✓ vs. *Blind* + *Darmoperation* **RR_{DS}** ✗

Arbeitsplatzmangel 'job scarcity'

Arbeitsplatz + *Mangel* **RR_{ALL}** ✓ vs. *Arbeits* + *Platzmangel* **INITIAL** ✗