UdS-(retrain|distributional|surface): Improving POS Tagging for OOV Words in German CMC and Web Data

System Description for the EmpiriST Shared Task

Jakob Prange, Andrea Horbach, Stefan Thater
Saarland University

Overview

- Corpora and our previous retraining & distributional approach
- Adaptions to the Shared Task
- Results
- Analysis
  - Potential of system combination
  - Influence of additional training data

Previous Work: (Horbach & al. 2014+15)

Schreibgebrauch Corpus

- Corpora:
  - Users posts from www.chefkoch.de
  - Twitter
  - Dortmunder Chat Corpus
- Manual annotation of ~34k tokens

Previous System: (Horbach & al. 2014+15)

#1 – Re-Training

- **Basic idea**: Combine a standard training set (Tiger) with our in-domain training set (boosted 5 times)
- Accuracy: 85% ⇒ 91.5% (on chefkoch test data)
- Learn about frequent CMC specific words and constructions
- Many words still not in training data.

Previous System (Prange & al. 2015)

#2 – Learning a POS-dictionary

assumption: words have the same POS tag as their distributional neighbours

„aba“ → distributional similarity

Candidates

aber/KON
doch/KON
net/NE
se/FM
Aber/KON
…

Step 1: Candidate Generation

- 20 most similar IV words for each OOV word
- distributional model trained on chefkoch data
- Features: POS 5-grams (POS_{-2}, POS_{-1}, _, POS_{1}, POS_{2})
- Weights: PMI scores

#2 – Learning a POS-dictionary

„aba“

distributional similarity

Candidates

aber/KON
doch/KON
net/NE
se/FM
Aber/KON
…

Ranking

Ranker 1:
1. KON
2. NE
3. FM

Ranker 2:
1. KON
2. FM
3. NE

Ranker 3:
1. NE
2. KON
3. FM

Ranker 4:
1. KON
2. NN
3. NE

KON

Step 2: Ranking

- surface similarity
- frequency and position of POS tags in the candidate list
- combination of rankers

Previous System (Prange & al. 2015)

#2 – Learning a POS-dictionary

91.5 ⇒ 93%

„aba“ → distributional similarity → **Candidates** → **Ranking** → KON

Candidates

aber/KON
doch/KON
net/NE
se/FM
Aber/KON
…

Ranking

Ranker 1:
1. KON
2. NE
3. FM

Ranker 2:
1. KON
2. FM
3. NE

Ranker 3:
1. NE
2. KON
3. FM

Ranker 4:
1. KON
2. NN
3. NE

KON

Step 2: Ranking
- surface similarity
- frequency and position of POS tags in the candidate list
- combination of rankers

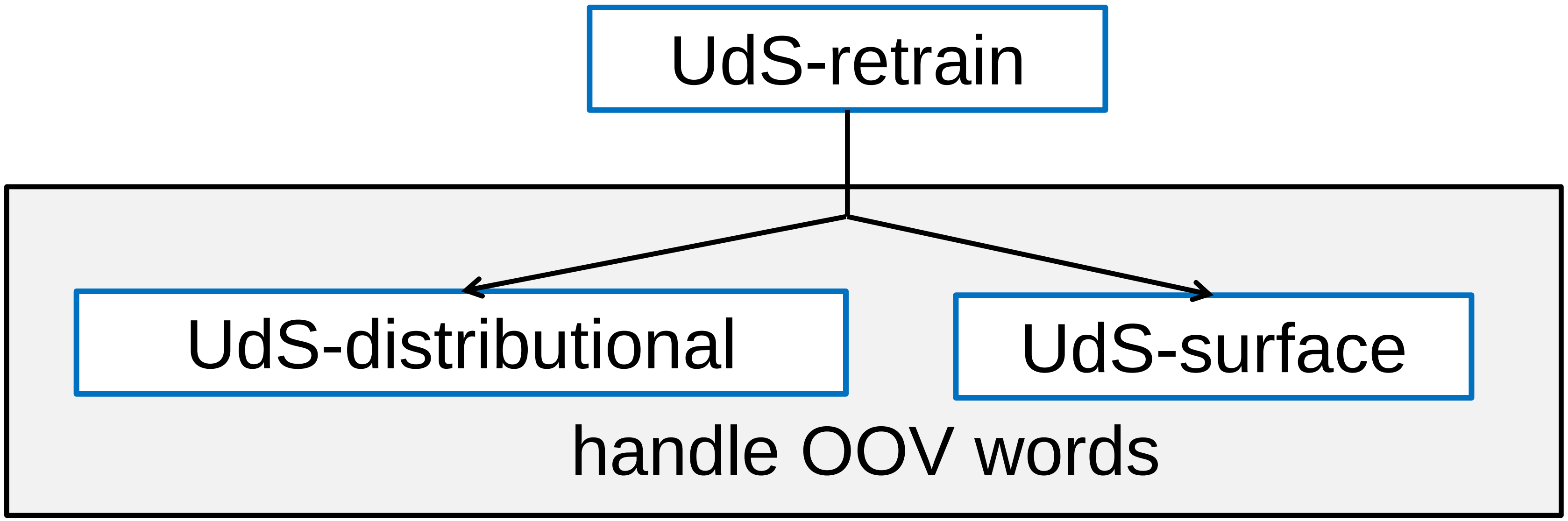
Shared Task – Our Objectives

- #1 – **Generalize the Approach**: allow for more than just one tag to be predicted (distributional)
- #2 – **Consider Alternatives**: use a language model to normalize input prior to tagging (surface)

Summary: Training Data

Dataset	#tokens	Domain	Tagset
TIGER	900 000	Newspaper	STTS 1.0
EmpiriST-Train CMC	5 000	chat, Twitter, Wikipedia talk, blog comments, whatsapp	STTS 2.0
EmpiriST-train Web	5 000	monologic Internet texts	STTS 2.0
Schreibgebrauch	34 000	forum, chat, Twitter	STTS 2.0* & STTS 2.0

Our Systems:


UdS-retrain
UdS-distributional
UdS-surface
handle OOV words

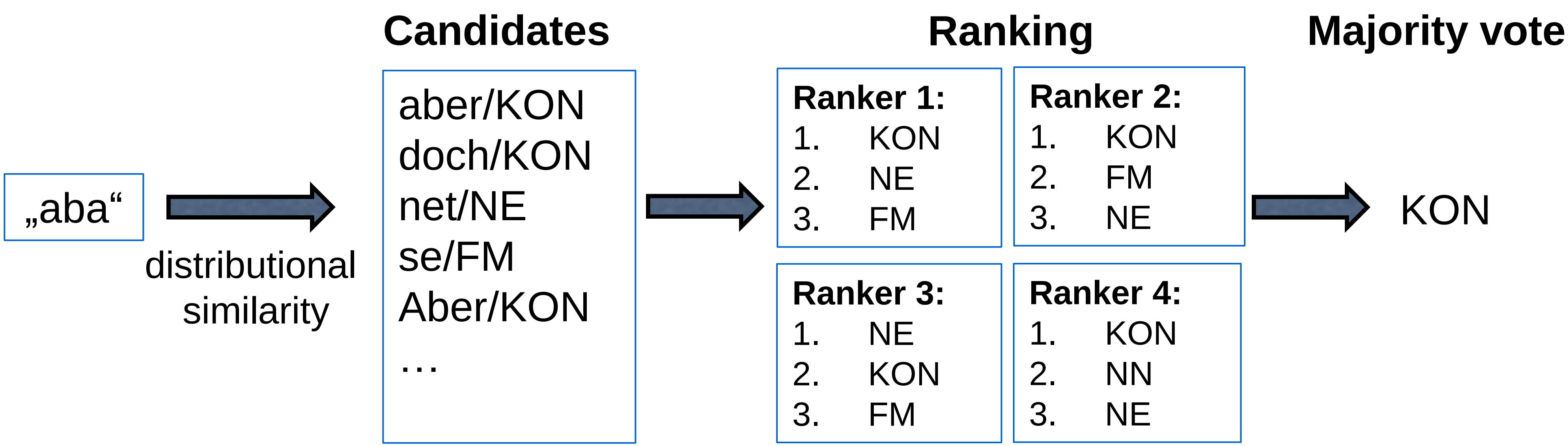
Our Systems: UdS-retrain

- **UdS-retrain**: baseline system, add additional annotated training data to TIGER corpus

Corpus	Run 1	Run 2	Run 3
TIGER	✓	✓	✓
EmpiriST – same domain	✓	✓	✓
EmpiriST – other domain			✓
Schreibgebrauch – original	✓		
Schreibgebrauch – adapted		✓	✓

Our Systems: UdS-distributional

- **UdS-distributional** – assumption: words have the same POS tag as their distributional neighbours



  - Run-1: majority vote between all rankers, use top-ranked POS tag
  - Run-2: use up to top-three POS tags
  - Run-3: linear combination of two best rankers (Prange et al. 2015)

Our Systems: UdS-surface

- **UdS-surface –** assumption: OOV words are often misspellings and similar to their intended forms

Candidates **Ranking in context**

„Das ist **aba** doof.“

surface similarity

aber
abend
ABBA
…

Language Model

1. Das ist **aber** doof.
2. Das ist aba doof.
3. Das ist abend doof.
4. Das ist ABBA doof.
5. …

select best sentence as tagger input

- Run-1: Jaro Winkler similarity above threshold of 0.8
- Run-2: threshold of 0.95
- Run-3: candidate(s) with highest similarity score

Results

Run	CMC	Web
TIGER baseline	71.2	91.2
UdS-retrain 1	85.5	92.7
UdS-retrain 2	86.4	**92.8**
UdS-retrain 3	**86.4**	92.7
UdS-distributional 1	87.3	93.5
UdS-distributional 2	**87.3**	**93.6**
UdS-distributional 3	87.3	93.
UdS-surface 1	84.6	91.2
UdS surface 2	**86.5**	**92.4**
UdS surface 3	85.4	92.0

Results

Run	CMC	CMC – OOV	Web	Web – OOV
TIGER baseline	71.2	29.0	91.2	71.1
UdS-retrain 1	85.5	74.0	92.7	77.9
UdS-retrain 2	86.4	**74.8**	**92.8**	78.1
UdS-retrain 3	**86.4**	74.7	92.7	**78.2**
UdS-distributional 1	87.3	78.9	93.5	82.9
UdS-distributional 2	**87.3**	**79.2**	**93.6**	**83.1**
UdS-distributional 3	87.3	78.8	93.0	79.4
UdS-surface 1	84.6	70.8	91.2	68.6
UdS-surface 2	**86.5**	**76.5**	**92.4**	**76.2**
UdS-surface 3	85.4	74.2	92.0	74.0

Oracle Experiment

- Estimate upper bound of classifier combination:

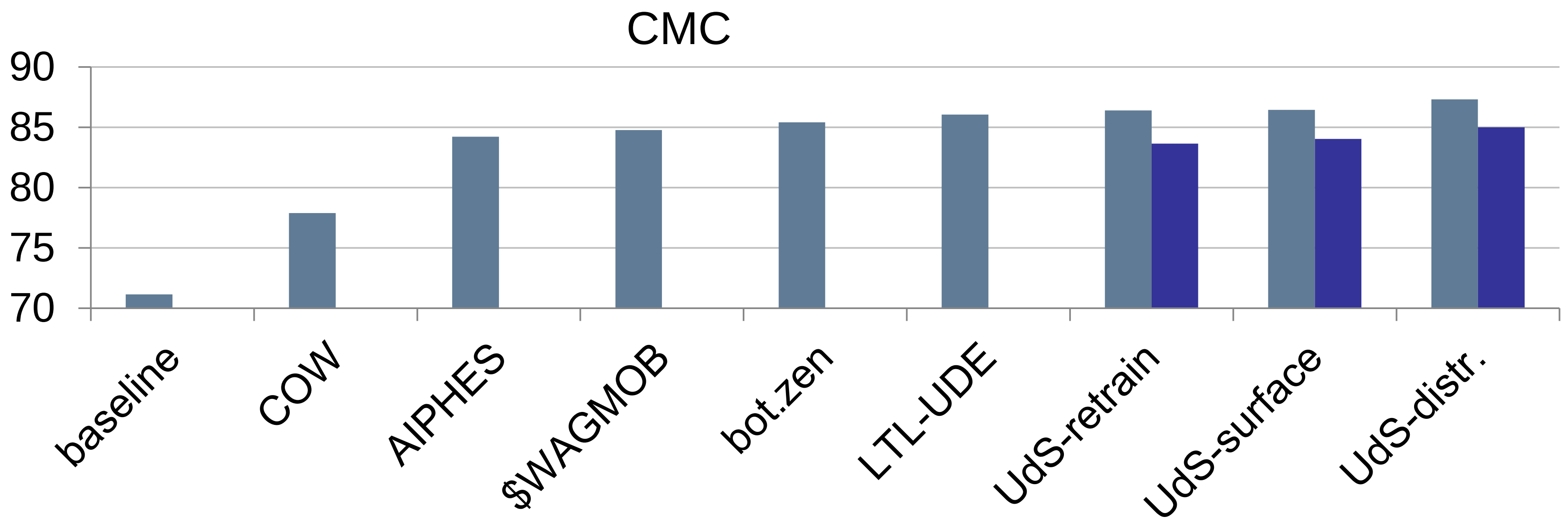
Gold	Run1	Run 2	Run 3	Oracle
ADV	ADV	ADV	ADV	✓
ART	ART	XY	ART	✓
NE	NN	NN	VVINF	✗
VVINF	VVFIN	VVINF	VVFIN	✓

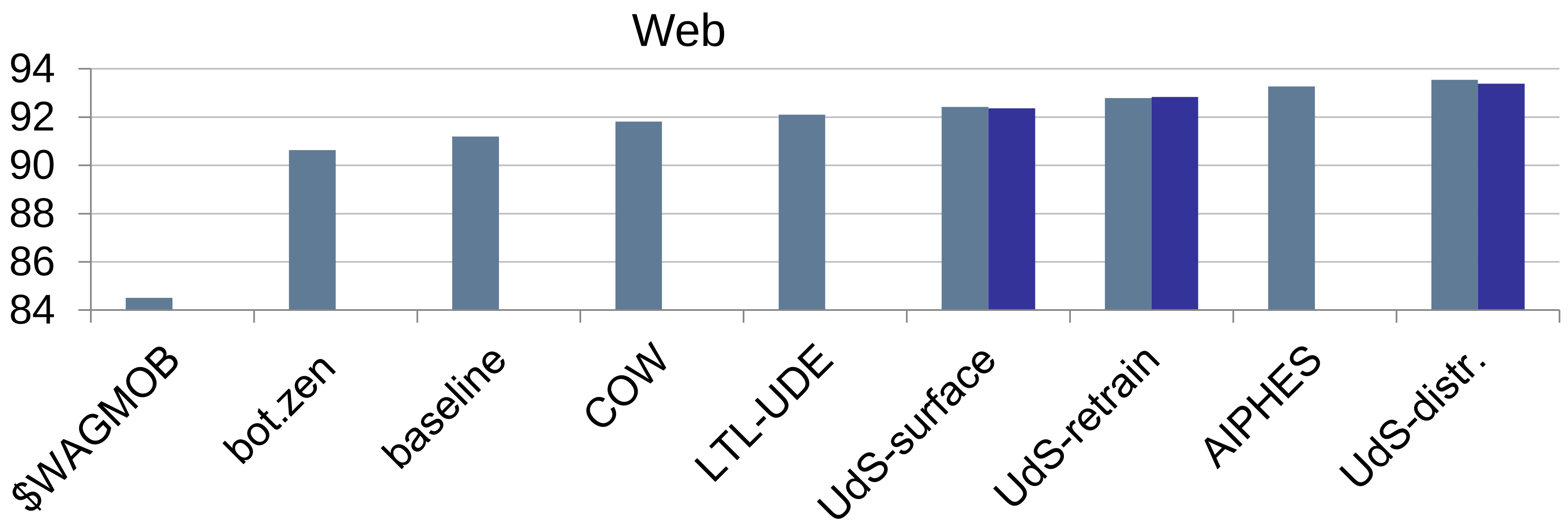
	CMC	Web
oracle – retrain	87.0 (+0.6%)	93.1 (+0.3%)
oracle – distributional	87.6 (+0.3%)	93.7 (+0.2%)
oracle – surface	87.5 (+1.1%)	93.6 (+1.2%)
oracle – all	89.8 (+2.5%)	94.9 (+1.6%)

Remaining Problems

- Most frequent mistaggings that all of our systems got wrong:
  - New adverb classes: PTKIFG, PTKMA, PTKMWL
  - ADR vs NE/NN
  - Common confusions such as NN vs NE, VVFIN vs VVINF, ADJD vs ADV, ADJD vs ADJA, ADJD vs VVPP
  - Punctuation: S(vs S: vs XY

Impact of our manually annotated training data





Conclusions

- Distributional models work better than surface-based normalization.
- No significant improvement, if we allow for several POS tags.
- Differences between datasets: CMC profits much more from our methods
- Oracle experiment indicates potential for future work.

Conclusions

- Distributional models work better than surface-based normalization.
- No significant improvement, if we allow for several POS tags.
- Differences between datasets: CMC profits much more from our methods
- Oracle experiment indicates potential for future work.

Thank you!