The Devil is in the Details: Parsing Unknown German Words

Daniel Dakota

Indiana University

September 13th, 2017

Parsing MRLs & German

- Strategies for English not directly transferable
- Harder to parse
- Increase data sparsity due to seldom seen words
- German posesses more morphology than English
- Possesses more rigid order than other MRLs
- morphologically rich-less configurational language (MR&LC)

Handling UNK Words

- *simple lexicon* model
- Rare words to estimate UNK words
  - Causes poor performance for rare words
- *sophisticated* model (Berkeley parser)
  - Uses suffix and capitalization information
  - Bias towards English lexicon

Brown Clustering

- Brown Clustering (Brown et al. 1992)
  - unsupervised clustering algorithm
  - obtains pre-specified numbers of clusters (C)
  - assigns most frequent words to C clusters
  - creates C+1 cluster and merges with defined clusters
  - minimize loss in likelihood determined via bigram model from corpus
- Results in bit-strings of variegated length
- Arranged heirachicaly

Data & Parsing

- Clustering
  - 175 million words Wikipedia data (Versley and Panchenko 2012)
  - Tagged with Mate Toolchains (Björkelund et al. 2010)
  - Brown clustering (Liang 2005)
- TiGer treebank (Brants 1997) from 2014 SPMRL (Seddah et al. 2014)
  - Crossing Branches
  - Flat internal phrase structure
- PCFG-LA Parsers
  - Berkeley (Petrov and Klein 2007)
  - Lorg (Attia et al. 2010)
- Average of grammars from 4 different seeds
- Train/Test on 5k

Methdology

- Experiments
  - Clustered: raw word & lemma_POS
  - UNK Words: suffix length, capitalization, open/closed class
- Evaluation
  - SPMRL Evalb
  - GFs are not scored
  - Penalty for unparsed sentences
  - Delete virtual roots
  - Score punctuation

Baselines

Parser	Terminal Type	Parsed	UNK TH.	F-score	POS Acc.	Lex. Red.
Lorg	orig	tokens	1	71.80	90.81	N/A
	orig	tagged	5	77.94	99.54	N/A
	lemma	tokens	1	71.54	90.87	27.83
	lemma	tagged	5	77.25	99.53	27.83
	lemma_pos	tokens	1	73.15	93.70	18.95
	lemma_pos	tagged	5	77.30	99.54	18.95
Berkeley	orig	tokens	5	75.10	94.04	N/A
	orig	tagged	5	76.69	99.87	N/A
	lemma	tokens	5	73.56	92.89	27.83
	lemma	tagged	5	75.91	99.83	27.83
	lemma_pos	tokens	10	75.21	95.97	18.95
	lemma_pos	tagged	10	76.01	99.93	18.95

- blue = gold POS tags
- red = parser-internal tags
- Lexicon Reduction (Lex. Red.) is defined as the proportional decrease in the vocabulary size of the word types from the original Tiger dev set to the dev set replaced with clusters and UNK types

Suffix Length

Token Type	Parsed	UNK TH.	F-score	POS Acc.	Lex. Red.
raw+orig	tokens	1	75.90	93.45	59.24
raw+orig	tagged	1	78.16	99.52	59.24
raw+unk_suffix0	tokens	1	75.88	93.26	**93.86**
raw+unk_suffix0	tagged	1	78.26	99.45	**93.86**
raw+unk_suffix1	tokens	5	76.14	94.05	93.45
raw+unk_suffix1	tagged	5	78.05	99.53	93.45
raw+unk_suffix2	tokens	5	76.27	94.23	91.09
raw+unk_suffix2	tagged	10	78.20	99.40	91.09
raw+unk_suffix3	tokens	1	76.05	93.86	86.61
raw+unk_suffix3	tagged	5	78.10	99.40	86.61
raw+unk_suffix4	tokens	1	76.03	93.92	80.63
raw+unk_suffix4	tagged	5	78.34	99.49	80.63

Table: Suffix Length for UNK Words for Lorg

Suffix Length

- No clear optimal suffix length
- Choose length 2 for three reasons
  - best results for parser-internal tagging
  - balances lexicon reduction
  - linguistic motiviation

Results for Lorg on raw words clusters

Token Type	Parsed	UNK TH.	F-score	POS Acc.	Lex. Red.
Craw	tokens	1	76.47	94.22	**93.38**
Craw	tagged	5	78.34	99.52	**93.38**
raw_suffix2	tokens	5	76.27	94.23	91.09
raw_suffix2	tagged	10	78.10	99.40	91.09
Craw_suffix2	tokens	1	76.50	94.57	89.98
Craw_suffix2	tagged	1	78.17	99.40	89.98
raw_noCC	tokens	1	76.00	93.68	92.73
raw_noCC	tagged	1	78.10	99.54	92.73
Craw_suffix2_noCC	tokens	1	76.57	94.93	88.86
Craw_suffix2_noCC	tagged	5	78.20	99.54	88.86

- C = capitalization on both clusters and UNK tokens
- suffix2 = UNK words have suffix of length 2
- noCC = closed class words were not replaced

Results for Lorg on lemma_pos clusters

Token Type	Parsed	UNK TH.	F-score	POS Acc.	Lex. Red.
Clemma_pos	tokens	1	76.86	96.54	93.32
Clemma_pos	tagged	1	77.44	99.51	93.32
lemma_pos_suffix2	tokens	1	76.78	96.69	91.63
lemma_pos_suffix2	tagged	1	77.67	99.52	91.63
Clemma_pos_suffix2	tokens	5	76.77	96.63	90.54
Clemma_pos_suffix2	tagged	5	77.46	99.54	90.54
lemma_pos_noCC	tokens	1	73.67	94.08	**94.04**
lemma_pos_noCC	tagged	10	77.48	99.53	**94.04**
Clemma_pos_suffix2_noCC	tokens	1	76.08	95.61	90.53
Clemma_pos_suffix2_noCC	tagged	5	77.45	99.53	90.53

- C = capitalization on both clusters and UNK tokens
- suffix2 = UNK words have suffix of length 2
- noCC = closed class words were not replaced

Lorg Clusters

- lemma_POS better without noCC
- raw words better with noCC
- Ambiguity?

General Trends

- No direct correlation between lexicon reduction and performance
- Mixed results for German compared to other langauges
- Suggests more fine-tuned approach is needed

UNK Token Analysis Train

UNK Type	Count	Top 3 POS Categories		
CUNK_en	897	NN (836)	NE (36)	ADJA (15)
UNK_en	624	ADJA (279)	VVINF (134)	VVFIN (89)
CUNK_er	429	NN (332)	NE (72)	ADJA (22)
CUNK_ng	255	NN (231)	NE (23)	ADJA (1)
CUNK_te	127	NN (115)	ADJA (8)	NE (3)
CUNK_es	112	NN (86)	NE (18)	ADJA (7)
CUNK_rn	110	NN (110)		
UNK_er	108	ADJA (79)	ADJD (18)	NN (7)
CUNK_in	106	NN (69)	NE (37)	
CUNK_el	103	NN (74)	NE (27)	PITA (1)

- 8/10 top 10 UNK signatures are shared between train and dev (in red)

UNK Token Analysis Dev

UNK Type	Count	Top 3 POS Categories		
CUNK_en	884	NN (795)	NE (32)	VVPP (6)
CUNK_er	515	NN (351)	NE (123)	ADJA (34)
UNK_en	462	ADJA (185)	VVINF (122)	VVFIN (82)
CUNK_ng	265	NN (253)	NE (10)	FM/ADJD (1)
CUNK_te	174	NN (166)	NE (4)	ADJA (4)
CUNK_rn	108	NN (103)	NE (3)	ADV (2)
CUNK_ft	101	NN (95)	NE (6)	
UNK_er	94	ADJA (68)	ADJD (17)	NN (6)
CUNK_es	91	NN (74)	NE (11)	ADJA (6)
UNK_te	89	VVFIN (49)	ADJA (38)	ADV/NN (1)

- 8/10 top 10 UNK signatures are shared between train and dev (in red)

External POS Tagger

Token Type	System	F-Score	POS Acc.
Craw_suffix2_noCC	TnT	n/a	94.43
	Lorg w/ TnT Tags	74.62	94.28
	Berkeley w/ TnT Tags	**75.70**	**94.65**
Clemma_pos	TnT	n/a	**96.66**
	Lorg w/ TnT Tags	**76.03**	96.26
	Berkeley w/ TnT Tags	75.56	96.26

Table: TnT Results

- TnT uses smoothing in its probability model
- TnT tags with Berkeley are extremely similar to using Berkeley tags
- Lorg performance drops with TnT tags

Token Type	Cluster Size	F-score	POS Acc.
Craw_suffix2_noCC	500	76.48	94.06
	800	76.65	94.64
	1000	76.57	94.93
	1500	76.60	95.12
	2000	76.45	95.22
Clemma_pos	500	76.67	95.73
	800	76.78	96.37
	1000	76.86	96.54
	1500	76.81	96.72
	2000	76.66	96.87

Table: Different cluster sizes

- POS tagging increases as cluster size increases
- Optimal cluster size not obvious
- Needs to be redone if minimum threshold changed
- Default settings not necessarily optimal

Conclusion

- Intricate interaction between reducing data sparisity and the handling of UNK words
- Better noting the interaction allows better exploitation for improvements
- Smoothing
  - Helps create generalized models
  - Biased again rare and UNK words
  - Can hinder performance
- *Simple Lexicon*
  - Can be exploited better using clustering methods

Future Work

- Include morphological information
- Mininum frequency used during clustering
- Further *noCC* results investigation
- Further examine differences in suffix signatures

References I

Attia, M., J. Foster, D. Hogan, J. Le Roux, L. Tounsi, and J. van Genabith (2010). Handling Unknown Words in Statistical Latent-Variable Parsing Models for Arabic, English and French. In *Proceedings of SPRML 2010*.

Björkelund, A., B. Bohnet, L. Hafdell, and P. Nugues (2010, August). A High-Performance Syntactic and Semantic Dependency Parser. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, Beijing, China, pp. 33–36.

Brants, T. (1997). The NeGRA Export Format. (CLAUS Report No. 98). Department of Computational Linguistics, Saarland University. Saarbrücken, Germany.

References II

Brown, P., V. Della, P. Desouza, J. Lai, and R. Mercer (1992). Class-Based n-gram Models of Natural Language. *Computational Linguistics 19*(4), 467–479.

Liang, P. (2005). Supervised Learning for Natural Language. Master's thesis, Massachusetts Institute of Technology.

Petrov, S. and D. Klein (2007, July). Learning and Inference for Hierarchically Split PCFGs. In *Proceedings of the National Conference on Artificial Intelligence*, Vancouver, Canada, pp. 1663–1666.

References III

Seddah, D., S. Kübler, and R. Tsarfaty (2014, August). Introducing the SPMRL 2014 Shared Task on Parsing Morphologically-rich Languages. In *Proceedings of the First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages*, Dublin, Ireland, pp. 103–109. Dublin City University.

Versley, Y. and Y. Panchenko (2012). Not Just Bigger: Towards Better- Quality Web Corpora. In *Seventh Web as Corpus Workshop (WAC7)*, Lyon, France, pp. 44–52.