Learning relational nouns from corpora

Berthold Crysmann

Explorations in syntactic government and subcategorisation, Cambridge

September, 2 2011

Outline

Motivation

- Substantial minority of (German) nouns feature internal arguments expressible as syntactic complements
- Mining relational nouns provides empirical basis for studies in lexicography and derivational morphology
- Identification of relational nouns also important for computational linguistic tasks
  - accurate deep parsing:
    assignment of correct semantics (predicate–argument structure)
  - Semantic Role Labeling:
    treebank-based semantic role annotations recently extended to nouns (Meyers et al., 2004)
  - Machine Translation:
    separate semantic task of translating modifiers from the syntactic task of translating complements

Syntactic classes of relational nouns in German

- nouns taking genitival complements
  e.g., *Beginn der Vorlesung* 'beginning of the lecture', *Zerstörung der Stadt* 'destruction of the city'
- nouns taking propositional complements
  - complementiser-introduced finite clauses
    der Glaube, daß die Erde flach ist ' the belief that Earth is flat'
  - infinitival complements
    der Versuch, das Publikum zu überzeugen 'the attempt to con vince the audience'
  - both
    die Erwartung, im Lotto zu gewinnen 'the expection to win the lottery' / *die Erwartung, daß er im Lotto gewinnt* 'the expectation that noone will wiull the lottery'
- nouns taking PP complements

Properties of German PP-taking nouns

- Prepositions used with relational nouns form a small circumscribed set
- Choice of preposition
  - relatively fixed (compared to modifiers)
  - arbitrary
    Interesse für/an 'interest in' (lit.: interest for/at) vs.
    *sich interessieren für/*an* 'to be interested in' vs.
    *interessiert an/*für* 'interested in'
  - Lack of alternation implies semantic vacuousness
- Complements of nouns almost exclusively optional
- PP-complements syntactically almost indistinguishable from PP-modifiers
  - grammar-based learning techniques (Cholakov et al., 2008) unapplicable
- similarity to multi-word expression suggests collocation-extraction approach

Data preparation

- Primary data: 1.6 billion word deWaC corpus (Baroni and Kilgariff, 2006), POS-tagged and lemmatised by TreeTagger (Schmid, 1995)
- Extraction of noun-preposition bigram and unigram counts
  - Using strict adjacency (non-adjacent complements highly marked)
  - Counts are lemma-based: motivated by acquisition task (lemma-based HPSG lexicon)
  - Removal of counts with noun frequency < 10
- Extraction of bigram frequency best-lists, a standard heuristic in collocation extraction (Krenn and Evert, 2001)
  - Frequency-based ranking highly suitable to the task
  - Ensures availability of sufficient positive training data

Rank	Abs. frequency	Bigram
1	99773	Umgang mit
2	96612	Institut für
3	86835	Höhe von
4	85879	Zusammenhang mit
5	84148	**Mensch in**
6	77836	Suche nach
7	77740	**Jahr in**
8	76426	Blick auf
9	75215	Zusammenarbeit mit
10	73510	Voraussetzung für
11	71589	Hinblick auf
12	70744	Anspruch auf
13	68652	Bezug auf
14	60617	Form von
15	60612	Reihe von

Table: Top 15 noun–preposition bigrams

Annotation

- Manual annotation of frequency best list
  - Initial annotation by 2 human annotators with basic training in linguistics (A1, A2): 2500 items
  - Second annotation by third-year student (A3): 8500 items
  - Interannotator agreement (top 2500) at .82 (A1/A3) and .84 (A2/A3)
  - Final accommodation step
- Annotation guidelines:
  - deverbal noun?
  - affectedness of preposition's complement?
  - paradigmatic interchangeability of preposition?
  - only possessor reading?
- 36% of annotated data classified as relational (3029/8268): clear bias for non-relational nouns

Features I

- Linguistic (string) features
  - Preposition
  - Noun suffix
    common derivational suffixes, like *-tion*, *-ung* etc.
  - Noun prefix
    common *verbal* prefixes, hinting at deverbal nature
- Association measures
  - Mutual information (MI; Church and Hanks, 1990)
  - MI^2 (variant of MI that does not overestimate bigrams with low marginal probabilities; Daille, 1994)
  - Fisher's t-score (Krenn, 2000; Krenn and Evert, 2001; Evert and Krenn, 2001)
  - Association strength (Smadja, 1993)
  - Likelihood ratio (Dunning, 1993)
  - Best/best ratio: most frequent preposition given noun

Features II

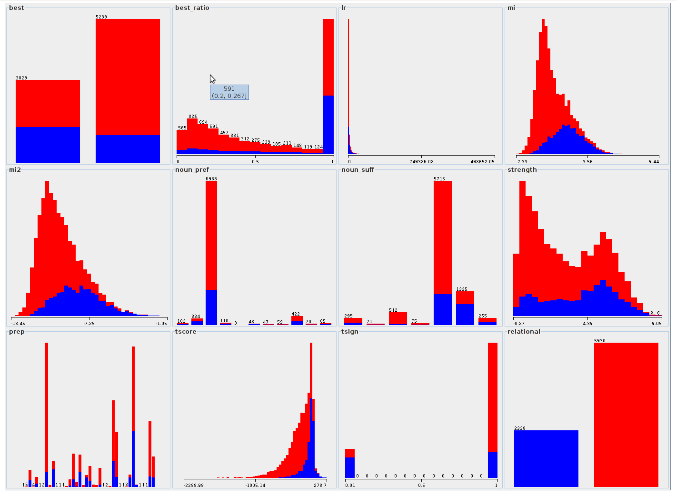

Figure: Distribution of relational nouns across features

Evaluation

- Experiments carried out over a set of 8268 annotated noun–preposition pairs (bigrams)
- All test runs performed using WEKA machine learning platform (Bouckaert et al., 2010)
  - decision trees
  - Bayesian classifiers
  - support vector machines
  - logistic regression
- Evaluation using 10-fold cross-validation

Performance of different learners

	Prec.	Rec.	F-meas.
ADTree	73	63.2	67.8
BFTree	79.7	55.9	65.7
DecisionStump	57.6	75.7	65.4
FT	75.8	62.4	68.5
J48	75.9	62.4	68.5
J48graft	76.1	62.6	68.7
LADTree	74.8	60.0	66.6
LMT	75.7	63.0	68.8
NBTree	75.2	64.2	**69.2**
RandomForest	70.0	66.7	68.3
RandomTree	64.4	64.7	64.5
REPTree	74.7	64.0	69.0
Naive Bayes	67.6	61.4	64.3
Bayes Net	61.8	70.0	65.7
SMO	76.9	63.6	69.6
Logistic	76.0	64.8	**69.9**
Bagging (RepTree)	77.0	64.4	70.2
Voting (maj)	75.5	67.1	71.0
Voting (av)	74.3	67.3	70.6

Individual association measures (AM)

- Mutual information and t-score show good individual performance, confirming results from collocation extraction
- Association strength and best feature useless on their own

	NBTree			Logistic		
	Prec.	Rec.	F-meas.	Prec.	Rec.	F-meas.
All (+form)	75.2	64.2	69.2	76	64.8	**69.9**
MI	63.1	65.6	**64.3**	68.6	47.3	56.0
MI2	65.0	46.3	54.1	67.4	40.8	50.8
LR	69.1	15.9	25.8	71.9	11.5	19.8
T-score	64.6	57.4	60.8	65.8	58.3	**61.8**
Strength	0	0	0	49.4	3.7	6.8
Best	0	0	0	0	0	0
Best-Ratio	0	0	0	0	0	0
All AM (−form)	67.9	48.2	56.4	68.1	50.3	57.9

Table: Classification by a single association metric

Sampling by preposition/noun type

- Addition of form features substantially increases performance of all association measures
- MI and t-score get close to their maximal values

	NBTree			Logistic		
	Prec.	Rec.	F-meas.	Prec.	Rec.	F-meas.
All	75.2	64.2	69.2	76	64.8	69.9
MI	78.5	60.4	68.3	76.3	64.0	69.6
MI2	75.2	58.5	65.8	75.2	60.2	66.9
LR	75.2	53.3	62.4	71.5	52.6	60.6
T-score	76.5	62.2	68.6	75.9	62.0	68.3
Strength	75.5	54.8	63.5	74.8	53.1	62.1
Best	73.1	51.5	60.4	75.2	48.8	59.2
Best-Ratio	75.6	55.3	63.9	76.2	51.9	61.7
No AM	67.7	49.1	56.9	0.703	46.8	56.2

Table: Classification by a single association metric + form features (*preposition, noun prefix, noun suffix*)

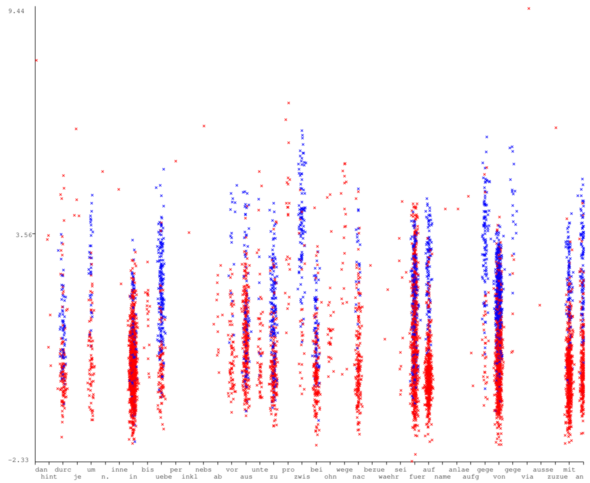

Figure: MI-values of relational nouns relative to preposition

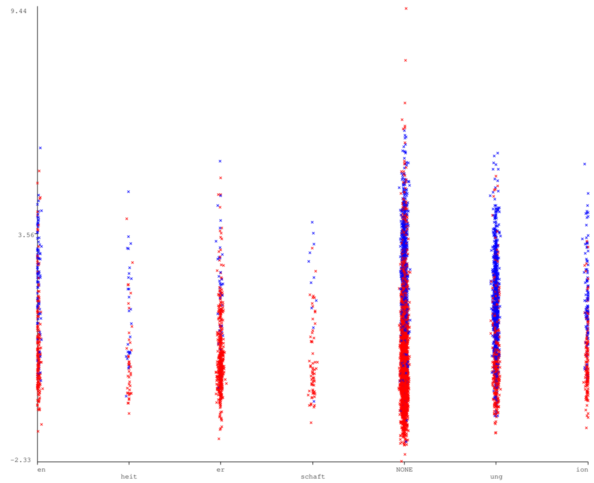

Figure: MI-values of relational nouns relative to noun suffix

Contribution of individual features

- Importance of suffix and preposition features evident in combined classifier: clear drop in precision and recall
- Omission of prefix heuristic displays a much weaker effect

	NBTree			Logistic		
	Prec.	Rec.	F-meas.	Prec.	Rec.	F-meas.
All	75.2	64.2	69.2	76	64.8	69.9
−T-score (signif.)	75.4	63.5	68.9	76.1	65.0	**70.1**
−T-score (abs)	75.6	62.3	68.3	**76.3**	63.3	69.2
−MI	75.9	63.7	**69.3**	75.1	64.8	69.6
−MI^2	74.4	63.7	68.6	76.0	64.2	69.6
−LR	74.9	63.9	68.9	75.8	**65.1**	**70.1**
−Strength	74.7	63.5	68.7	76.1	65.0	**70.1**
−Best	75.3	63.1	68.7	76.0	64.6	69.8
−Best-Ratio	75.1	63.9	69.1	76.0	64.8	69.9
−Prep	68.7	64.0	66.3	72.3	60.6	65.9
−Noun-Prefix	74.9	63.7	68.9	76.0	64.5	69.7
−Noun-Suffix	73.7	60.9	66.7	73.5	60.4	66.3

Table: Effects of leaving one feature out

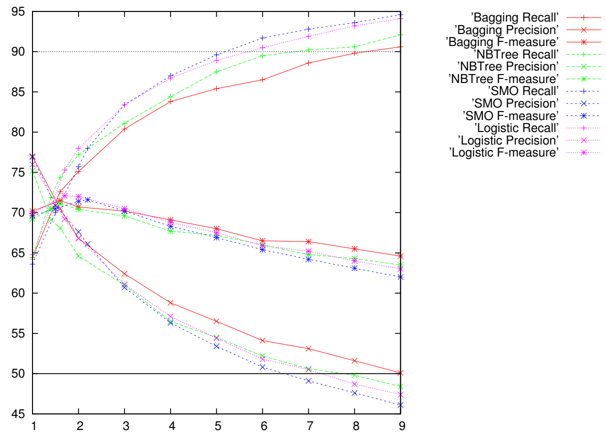

Figure: Effect of trading precision for recall

Conclusion

- Classifiers
  - Bayesian classifiers suboptimal
  - Best decision tree classifiers show competitive performance to support vector machines (SMO) and logistic regression
- Features
  - Mutual information and t-scores confirmed as best individual association measures
  - Corpus statistics on their own insufficient
  - Information about preposition and derivational noun suffixes crucially improves performance of all association metrics
  - Association measures with low predictive power still useful in combination
- Satisfactory overall performance
  - confirms suitability of collocation extraction approach
  - best learner can detect over 90% of relational nouns, with a precision above 50%, reducing the annotation effort by half

References I

Baroni, M. and A. Kilgariff: 2006, 'Large linguistically-processed Web corpora for multiple languages'. In: *Proceedings of EACL 2006*.

Bouckaert, R. R., E. Frank, M. A. Hall, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten: 2010, 'WEKA–Experiences with a Java Open-Source Project'. *Journal of Machine Learning Research* **11**, 2533–2541.

Cholakov, K., V. Kordoni, and Y. Zhang: 2008, 'Towards Domain-Independent Deep Linguistic Processing: Ensuring Portability and Re-Usability of Lexicalised Grammars'. In: *Coling 2008: Proceedings of the workshop on Grammar Engineering Across Frameworks*. Manchester, England, pp. 57–64, Coling 2008 Organizing Committee.

Church, K. and P. Hanks: 1990, 'Word Association Norms, Mutual Information, and Lexicography'. *Computational Linguistics* **16**(1), 22–29.

Daille, B.: 1994, 'Approche mixte pour l'extraction automatique de terminologie : statistique lexicale et filtres linguistiques'. Ph.D. thesis, Université Paris 7.

Dunning, T.: 1993, 'Accurate methods for the statistics of surprise and coincidence'. *Computational Linguistics* **19**, 61–74.

Evert, S. and B. Krenn: 2001, 'Methods for the qualitative evaluation of lexical association measures'. In: *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, Toulouse, France*. pp. 188–195.

References II

Krenn, B.: 2000, 'The Usual Suspects: Data-oriented Models for the Identification and Representation of Lexical Collocations'. Ph.D. thesis, Universität des Saarlandes.

Krenn, B. and S. Evert: 2001, 'Can we do better than frequency? A case study on extracting PP-verb collocations'. In: *Proceedings of the ACL Workshop on Collocations, Toulouse, France*. pp. 39–46.

Meyers, A., R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman: 2004, 'The NomBank Project: An Interim Report'. In: A. Meyers (ed.): *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*. Boston, Massachusetts, USA, pp. 24–31, Association for Computational Linguistics.

Schmid, H.: 1995, 'Improvements in Part-of-Speech Tagging with an Application to German'. In: *Proceedings of the ACL SIGDAT-Workshop*.

Smadja, F.: 1993, 'Retrieving Collocations from Text: Xtract'. *Computational Linguistics* **19**(1), 143–177.

Mutual information (MI) (Church and Hanks, 1990)

$$MI = \frac{p(noun, prep)}{p(noun) * p(prep)}$$

MI^2 (Daille, 1994)

$$MI^2 = \frac{(p(noun, prep))^2}{p(noun) * p(prep)}$$

t-score Fisher's t-test Krenn (2000); Krenn and Evert (2001); Evert and Krenn (2001).

$$tscore = \frac{p(noun, prep) - (p(noun) * p(prep))}{\sqrt{\frac{\sigma^2}{N}}}$$

Likelihood ratios (Dunning, 1993)

$$\begin{aligned} LR = \log L(p_i, k_1, n_1) + \log L(p_2, k_2, n_2) \\ - \log L(p, k_1, n_1) - \log L(p, k_2, n_2) \\ \text{where} \\ \log L(p, n, k) = k \log p + (n-k) \log(1-p) \\ \text{and} \\ p_1 = \frac{k_1}{n_1}, p_2 = \frac{k_2}{n_2}, p = \frac{k_1 + k_2}{n_1 + n_2} \end{aligned}$$

Association Strength (Smadja, 1993)

$$Strength = \frac{freq_i - \bar{f}}{\sigma}$$

Best Indicates whether a bigram is the most frequent one for the given noun or not.

Best-Ratio A relative version of the previous feature indicating the frequency ratio between the current noun–preposition bigram and the best bigram for the given noun.