Learning relational nouns from corpora

Berthold Crysmann, Universität Bonn

In this talk we present a collocation extraction approach to the acquisition of relational nouns in German. We annotated frequency-based best lists of noun-preposition bigrams extracted from the 1.6 billion token deWaC corpus (Baroni & Kilgariff, 2006) and subsequently trained different classifiers using (combinations of) association metrics, achieving a maximum F-measure for extracted relational nouns of 69.7 on a support vector machine (Platt, 1998). Trading precision for recall, we could achieve over 90% recall for relational noun extraction, while still halving the annotation effort, a factor we expect to go up considerably as we move down the Zipf distribution.

A substantial minority of German nouns are characterised by having an internal argument structure that can be expressed as syntactic complements. A non-negligible number of relational nouns are deverbal, inheriting the semantic argument structure of the verbs they derive from. In contrast to verbs, however, complements of nouns are almost exclusively optional.

The identification of relational nouns is of great importance for a variety of content-oriented applications: first, precise parsing cannot really be achieved, if a high number of noun complements is systematically analysed as modifiers. Second, recent extension of Semantic Role Labeling to the argument structure of nouns (Meyers et al., 2004) increases the interest in lexicographic methods for the extraction of noun subcategorisation information. Third, relational nouns are also a valuable resource for machine translation, separating the more semantic task of translating modifying prepositions from the more syntactic task of translating subcategorised for prepositions. Finally, the extraction fo relational nouns is also of central interest for the syntax of complementation, delivering a broad empirical basis for linguistic studies. Despite its relevance for accurate deep parsing, the German HPSG grammar developed at DFKI (Crysmann, 2005; Müller & Kasper, 2000) currently only includes 107 entries for proposition taking nouns, and lacks entries for PP-taking nouns entirely.

In terms of subcategorisation properties, relational nouns in German can be divided up into 3 classes: first, nouns taking genitival complements second, nouns taking propositional complements, either a complementiser-introduced finite clause, or an infinitival clause, or both, and, finally, nouns taking PP complements. We focus on nouns taking prepositional complements, although the method described here can also be easily applied to the case of complementiser-introduced propositional complements and genitival complements. The prepositions used with relational nouns all come from a small set of basic prepositions, mostly locative or directional. A characteristic of these prepositions when used as a noun's complement, is that their choice becomes relatively fixed, a property shared with multi word expressions (MWE) in general. Furthermore, choice of preposition is often arbitrary, sometimes differing between relational nouns and the verbs they derive from, e.g., *Interesse an* 'lit: interest at' vs. *interessieren für* 'lit: to interest for'. Owing to the lack of alternation, the preposition by itself does not compositionally contribute to sentence meaning, its only function being the encoding of a thematic property of the noun. Thus, in syntacto-semantic terms, we are again dealing with prototypical MWEs.

The fact that PP complements of nouns, like modifiers, are syntactically optional and that their surface form is indistinguishable from adjunct PPs makes the extraction task far from trivial. We therefore exploited the collocational properties of relational nouns, building on the expectation that the presence of a subcategorisation requirement towards a fixed, albeit optional, prepositional head should leave a trace in frequency distributions. Thus, building on previous work in MWE extraction, we pursued a data-driven approach that builds on a variety of association metrics combined in a probabilistic classifier.

As primary data for relational nonun extraction, we used the deWaC corpus by Baroni & Kilgariff (2006). From this corpus we extracted a best-list of noun–preposition bigrams, based on absolute frequency counts, a well-established heuristical measure for collocational status Krenn (2000). Using a frequency based best list not only minimises initial annotation effort, but also ensures the quickest improvement of the target resource, the grammar's lexicon. Finally, the use of ranked best lists also made sure that we always had enough positive items in our training data. The first 4333 items of the ranked best list were subsequently annotated by three human annotators. Among these, 27.2% were classified as relational nouns.

In addition to bigram frequency, we calculated several statistical association measures to be used as features in the learner: Mutual Information (MI, Church & Hanks, 1990), squared MI (MI^2; Daille, 1994), the scores of Fisher's t-test (e.g. Krenn, 2000), likelihood ratios (*LR*, Dunning, 1993), and association strength (Smadja, 1993). In addition, we used the form of the preposition as well as the noun's prefixes or suffixes as linguistic features.

All experiments reported here were carried out using WEKA, a platform for data exploration and experimentation (Hall et al., 2009). Testing different classifiers and different metrics, we found that optimal results were obtained using a support vector machine Platt (1998), including *MI*, MI^2, and *LR* as association measures, together with information about the identity of the preposition and the noun's prefix and suffix. The second best classifier, a hybrid decision tree with Naive Bayes classifiers at the leaves (Kohavi, 1996) produced highly competitive results. T-scores, while being a good predictor on their, however led to a slight decrease in performance, when a full feature set was used. Likewise, performance suffered when Association Strength Smadja (1993) was included. Performance of the best individual classifier figured at an F-score of 69.7 for the actual task of relational noun extraction, and at 82.6 for overall performance, including classification of both relational and non-relation nouns.

Baroni, Marco & Kilgariff, Adam, 2006. Large linguistically-processed web corpora for multiple languages. In *Proceedings of EACL 2006*.

Church, Kenneth & Hanks, Patrick, 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1): 22–29.

Crysmann, Berthold, 2005. Relative clause extraposition in German: An efficient and portable implementation. *Research on Language and Computation* 3(1): 61–82.

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

Dunning, Ted, 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19: 61–74.

Hall, Mark *et al.*, 2009. The WEKA data mining software: an update. *SIGKDD Explorations* 11(1): 10–18.

Kohavi, Ron, 1996. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Second International Conference on Knowledge Discovery and Data Mining*, 202–207.

Krenn, Brigitte, 2000. The usual suspects: Data-oriented models for the identification and representation of lexical collocations. Ph.D. thesis, Universität des Saarlandes.

Meyers, A. *et al.*, 2004. The nombank project: An interim report. In *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, A. Meyers (ed), 24–31. Boston, Massachusetts, USA: Association for Computational Linguistics.

Müller, Stefan & Kasper, Walter, 2000. HPSG analysis of German. In *Verbmobil: Foundations of Speech-to-Speech Translation*, Wolfgang Wahlster (ed), 238–253. Berlin: Springer.

Platt, J., 1998. Fast training of support vector machines using sequential minimal optimization. In *Advances in Kernel Methods - Support Vector Learning*, B. Schoelkopf, C. Burges & A. Smola (eds). MIT Press.

Smadja, Frank, 1993. Retrieving collocations from text: Xtract. *Computational Linguistics* 19(1): 143–177.