Automatic Acquisition of Paradigmatic Relations using Iterated Co-occurrences

Chris Biemann, Stefan Bordag, Uwe Quasthoff

Leipzig University
Computer Science Institute, NLP Dept.
Augustusplatz 10/11
04109 Leipzig, Germany
{biem, sbordag, quasthoff}@informatik.uni-leipzig.de

Abstract

We introduce the notion of iterated co-occurrences, which can be obtained by performing the calculation of statistically significant co-occurrences not on sentence level, but on co-occurrence sets of previous calculations. The underlying mechanisms are explained in detail and we give reasons, why this iteration results in sets of semantically homogeneous words. These can be used for the automatic acquisition of paradigmatic relations in order to semi-automatically extend lexical-semantic word nets or thesauri, widening the acquisition bottleneck. A small evaluation for synset expansion for German language and some discussion conclude the work.

1. Introduction

A repeatedly addressed problem in computational linguistics is the so called 'acquisition bottleneck': A lot of time and money is being invested in building handcrafted lexical resources for the use in further processing. Well-known resources in this respect are lexical-semantic word nets, such as WordNet (Miller, 1990), EuroWordNet (Bloksma et al., 1996) or GermaNet (Kunze, 2000). These word nets are widely accepted and used, despite their coverage problems: None of the nets contains significantly more than 100'000 lexemes of one language, whereas millions of lexemes can be found in corpora of decent size. Another problem is that the hierarchy in these nets is defined once and for all by the according linguists and may or may not fit the domain in which it is going to be used. We present an statistic approach for the semi-automatic extension of word nets, using a large, nonannotated corpus.

2. Collocations and their iteration

Our major source for finding candidates is the notion of sentence-based statistical co-occurrence. The repeated occurrence of two or more words within a well-defined unit of information (sentence, document) is called a (statistical) co-occurrence. For the selection of meaningful and significant co-occurrences, an adequate co-occurrence measure has to be defined. We use a significance measure similar to the well known log-likelihood measure. Calculations are performed on very large corpora (>100 Million Tokens), using sentences or immediate neighboring words (sentence-based and neighborhood-based co-occurrences, cf. Heyer et al. (2001)) as units. From an intuitive point of view, significant co-occurrences of a word w contain all kinds of associated words, be it typical modifiers, synonyms, antonyms, hyperonyms, or members of the same semantic frame. Hence, the co-occurrence set of w contains words that are closely related to w. With the set of words comes a ranking for each word, based on the significance measure.

The calculation of the above (first-order) co-occurrences can be iterated in order to obtain co-occurrences of higher order. While in first order, a word co-occurs with another word when significantly often appearing in the same unit of information, i.e. sentences in this case, nth-order co-occurrences are words that significantly often occur together in co-occurrence sets of order n-1.

Hence, in the second step we construct an artificial corpus of "sentences" consisting of the co-occurrence sets of the original corpus. It is important to note, that this ignores the significances for each co-occurrence. Furthermore, a global threshold of a maximum number of co-occurrences is chosen and all other, weaker co-occurrences are ignored in each step. This artificial corpus is large enough for calculating co-occurrences because there are enough words having co-occurrences in the step n-1. In our experiments, there were co-occurrence sets with at least two elements for about 200.000 word forms.

While first-order co-occurrences of a word w usually consist of a biased mixture of mainly syntagmatically and sometimes paradigmatically related words (cf. de Saussure. 1916), the co-occurrence sets of higher orders seem to contain words that are mainly related in a paradigmatic way to w, such as hyponyms, hypernyms, antonyms and synonyms, in the following called 'X-Onyms'. Exactly those X-Onym-relations are used to build up structures like WordNet. But simply taking some high order co-occurrences of a given word is too vague yet to yield good candidates for word net extension: First - statistical approaches operating on word forms do not distinguish between word classes, second - the problem of word sense ambiguity has to be faced. And finally the errors due to statistics have to be taken account of.

3. Understanding iterated co-occurrences

Figure 1 shows how two words A and B become iterated co-occurrences in the first iteration step. The same procedure applies for further iteration steps.

We start with co-occurrences shown as solid lines in the figure. Here, both A and B are connected with all of the collocates C1, …, C5. Hence, A and B occur together in several of the co-occurrence sets and the co-occurrence of A and B gets a count of (at least) 5. This count will in the next step be transformed by the co-occurrence measure. If we assume the result being above the threshold, we get a new edge between A and B in the first iteration step.

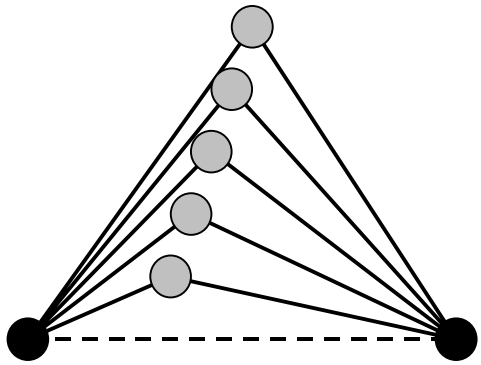


Figure 1: A single iteration step

Note that we need several joint collocates to connect A and B in an iteration step.

In any further iteration step, two words A and B are connected within two iteration steps if they have enough joint collocates. Hence, the iteration of the process corresponds to a cascaded picture of the above type. Details of two iteration steps, i.e. second order co-occurrences are shown in figure 2. Here, each solid line in figure 1 is replaced by a copy of the whole figure. All paths used to connect A and B are of length 4.

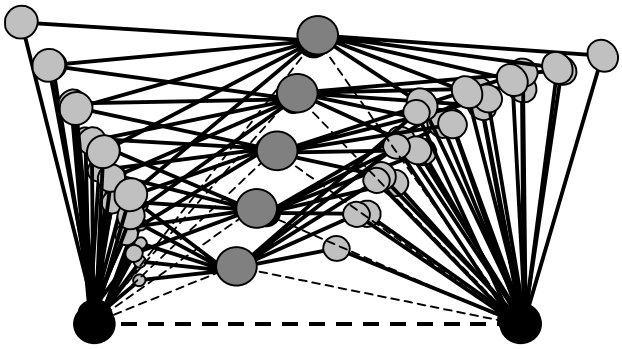


Figure 2: Two iteration steps

In the general case of n iteration steps we will have connecting paths that are of length 2^n. This number is only an upper bound because some words may be used repeatedly as joint collocates. Then the graph may collapse as shown in figure 3. In this situation, A and B will iterated co-occurrences in any further iteration step because there exist always enough joint collocates.

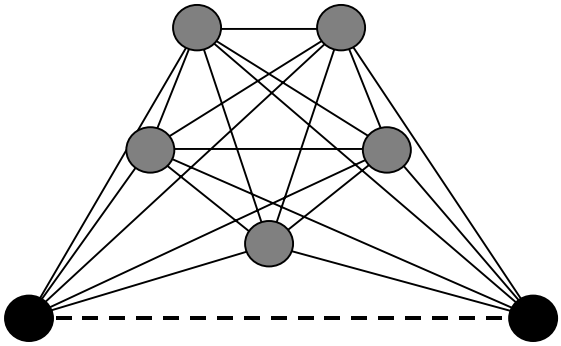


Figure 3: Collapsing bridging nodes

This situation can frequently be found in real data and is the reason for co-occurrence sets which are nearly stable under further iteration.

4. Finding Candidates for WordNet extension

X-Onyms for a given word w by definition belong to the same word class (noun, adjective, verb) as w. A possibility to exclude unwanted word classes is to filter the co-occurrence sets in that respect. This can be done by prior POS-tagging of the corpus and then performing the (iterated) co-occurrence calculation on pairs (word, tag). Alternatively, it is possible to train a string-based classifier on word suffixes of known words that are marked with their word class in WordNet already. The words in the co-occurrence set are then classified and filtered out if their word class differs from w.

The problem of lexical ambiguity (and statistical errors in the same time) is addressed by not simply using one input word to obtain only one co-occurrence set, but instead by taking several input words according to Yarowskys claim of one-sense-per-co-occurrence (cf Yarowsky, 1995) and disjointing the corresponding co-occurrence sets. While the co-occurrence set of an ambiguous word contains concepts related to two or more different readings, the disjunction set of two or more closely related words (e.g. members of the same synset in WordNet) does not suffer from this problem and reflects the appropriate singular reading unless all input words show the same ambiguity, which is rare.

The result set now contains X-Onyms at a high rate. Parameters are the order of the co-occurrence iteration and the fuzziness of the disjunction (especially when using large n, full disjunction tends to result in empty sets. The fuzziness can be defined with respect to co-occurrence significance).

So far we have performed experiments on the extension of GermaNet synsets, using the Wortschatz-Corpus for the calculation of co-occurrences up to third order. These preliminary experiments showed three interesting points: First, co-occurrences of higher than first order increase the number of X-Onyms in the result set. Second, the fraction of X-Onyms is higher amongst the higher ranked elements of the result set – that justifies the ranking via the significance measure. Third, the rate of X-Onyms is at about 40%-50%, which means that this method provides a fast and efficient candidate search for the semi-automatic extension of word nets. The extension itself cannot be done unsupervised, but is the task of the lexicographer.

5. Determining the appropriate relation

So far we have determined candidates for X-Onyms but did not further classify them into the appropriate relation, which is crucial for a higher automatization level. To differentiate between co-hyponyms and hypernyms we again can use iterated co-occurrences. From the observations, that co-hyponyms tend to occur together in sentences (e.g. in enumerations) and have similar contexts, the co-occurrence significance between two cohyponyms should be high for first-order (co-occurrence in sentences) as well as for second order (context similarity) co-occurrences. Hyperonyms, opposed to that, have similar contexts but seldom occur in the same

sentence; so two words in hyponymy relation bear high co-occurrence significance for second order co-occurrences and low significance for the first order. First tests showed good results; more thorough evaluation is under way. The following figure depicts the idea graphically.

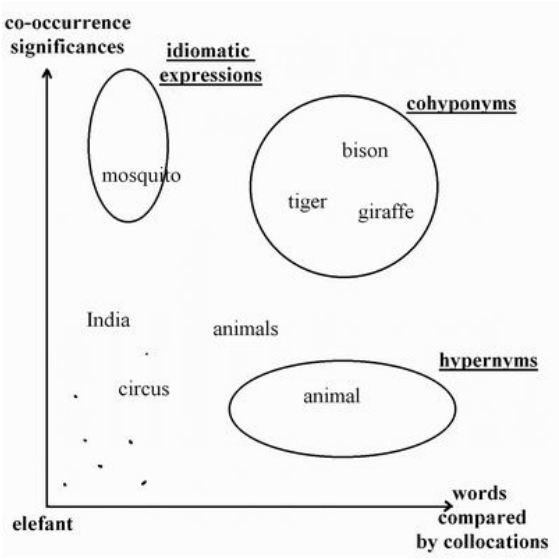


Figure 4: sorting co-hyponyms and hypernyms

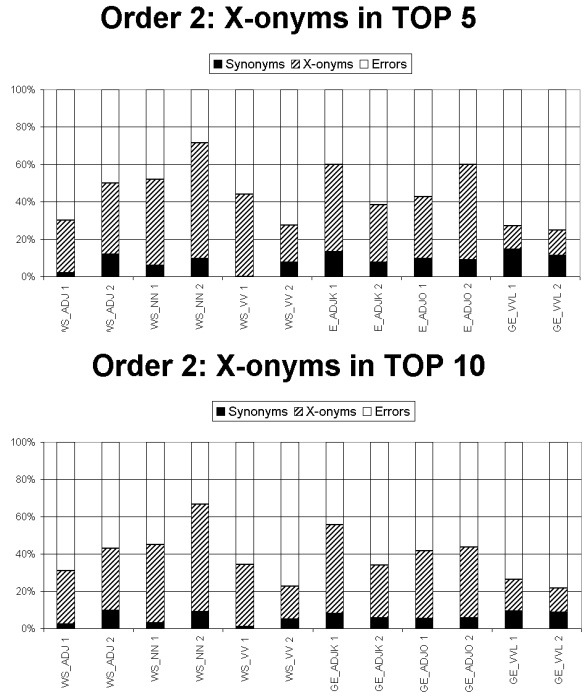
6. Evaluation of the algorithms

An evaluation of a prototypical implementation of the algorithms described in part 4 is presented here. Neighborhood co-occurrences of the second and third order along with a POS-tagger (TNT, cf. Brants, 2000) for word class filtering have been used in order to obtain X-Onym sets as pure as possible. Intuitively the neighborhood co-occurrences contain word forms, which have similar bigram contexts. We have used synsets of three different lexical word fields of GermaNet as well (manually obtained) synonyms out of our own corpus in order to measure the consistency of the word sets the algorithms generated. There are two measures – first, how many direct synonyms and second how many more general X-Onyms the algorithm found. All other word forms found are considered as wrong. Since the co-occurrence measure gives a significance value and thus an ordering according to this significance, the sets are ranked and it is possible to compare top-5 against top-10 rankings of these sets of words. Table 1 shows some samples of synsets and the sets the algorithm gave:

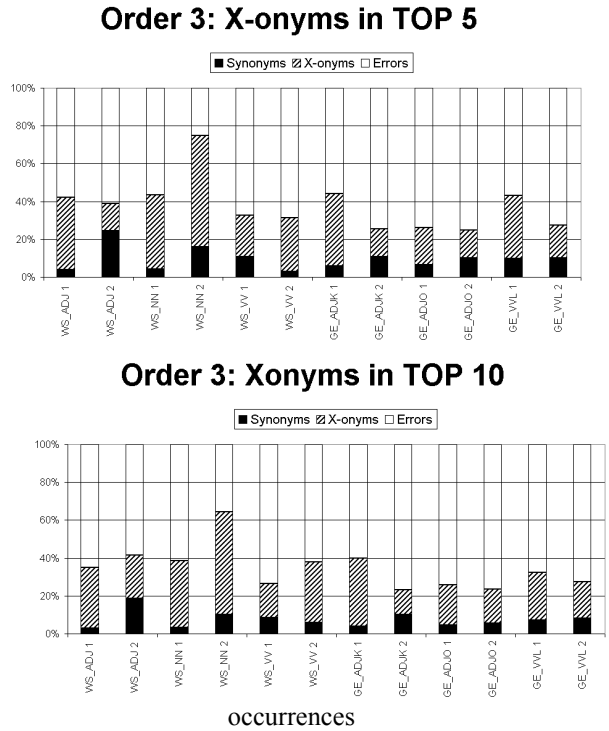
Table 1: Synonyms are marked with S, X-Onyms with X, the wrong words are not marked.

Synset	**result set of algorithm**
Mörder, Killer	Täter(X), Verbrecher(X), Attentäter(X), Kriegsverbrecher(X), Straftäter(X), Räuber(X), Mann, Einbrecher(X), Terrorist(X), Brandstifter(X)
verrückt	gerne, enttäuscht, bekannt, albern(S), begeistert, bescheuert(S), erschöpft, frustriert, dünn
fertigen, stabilisieren	sichern(S), sorgen, auswirken, beitragen(X), gefährden, stützen(S), erwarten, fördern(X), profitieren, bringen, führen
dunkelbraun	braun(X), hellbraun(X), rotbraun(X), orange(X), rot(X), rosa(X), violett(X), religiös, blutrot(X), graubraun(X)

The figures 5 to 8 show graphically how many Synonyms and X-Onyms against wrong words were found in a small statistical evaluation:



Figures 5 and 6: evaluation for second order co-occurrences



Figures 7 and 8: evaluation for third order co-occurrences

Table 2: Explanations of the abbreviations used in the images 1 to 4

abbrev.	meaning
WS_ADJ	adjectives in the Wortschatz corpus
WS_NN	nouns in the Wortschatz corpus
WS_VV	verbs in the Wortschatz corpus
GER_ADJK	adjectives in GermaNet/body
GER_ADJO	adjectives in GermaNet/location
GER_VVL	verbs in GermaNet/location

The ratio of synonyms and X-Onyms in the top-5 rankings being higher then in the top-10 rankings shows clearly that the ranking is a useful means of restricting the amount of wrong candidated.

The expectations that larger input sets (i.e. disjoining instead of using just single words) should bring better results were only partly fulfilled. The quality in the two-items sets from the Wortschatz corpus is significantly better as compared to the single words. But when comparing GermaNet sets this observation cannot be made.

It also becomes obvious that the class of verbs works worst with these algorithms, what is not surprising because of their complex argument structure and due to their long-range dependencies.

With these results it becomes possible to generate good candidate sets but it is not yet possible to fully automatically generate nearly 100%-pure candidate sets of synonyms or other paradigmatic relations.

7. Further work

We described a method that extracts paradigmatically related words for input words. The method, combined with the determination of the corresponding relation between input word and extracted word, gives rise to a workbench for the semi-automatic extension of lexical-semantic word nets. Because the mechanism is of statistical nature, it can be applied to any natural language. Before these methods can find their way into applications, however, more thorough evaluation has to be undertaken.

8. References

Armstrong, S. (ed.) (1993): "Using Large Corpora"; Computational Linguistics 19, 1/2 (1993). Special Issue on Corpus Processing, repr. MIT Press 1994.

Biezunski, M., and Newcomb, S.R. (2001): XML Topic Maps: Finding Aids for the Web, IEEE Multimedia 8, 2 , p. 108.

Bloksma L., Diez-Orzas, Vossen, P. (1996): User requirements and functional specification of the EuroWordNet project. Deliverable D001, EuroWordNet, LE2-4003, Computer Centrum Letteren, University of Amsterdam. Amsterdam.

Böhm, K., Heyer, G., Quasthoff, U., Wolff, Ch. (2002): Topic Map Generation Using Text Mining. In: Proc. IKNOW '02 – Intl. Conf. on Knowledge Management, Graz.

Brants, T. (2000): TnT - A Statistical Part-of-Speech Tagger. Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000, Seattle, WA.

Chung, K. L. (2000): A Course in Probability Theory, Academic Press.

Davidson, R., Harel, D. (1996): Drawing Graphs Nicely Using Simulated Annealing. In: ACM Transactions on Graphics 15(4), 301-331.

Saussure, F de. (1916): Cours de Linguistique Générale, Paris, Payot.

Heyer, G., Läuter, M., Quasthoff, U., Wittig, Th., Wolff, Ch.(2001): Learning Relations using Collocations; In: Proc. IJCAI Workshop on Ontology Learning, Seattle/WA, August 2001, pp. 19-24.

Heyer, G., Quasthoff, U., Wolff, Ch.(2000): Aiding Web Searches by Statistical Classification Tools. Proc. Proc. 7. Intern. Symposium f. Informationswissenschaft ISI 2000, UVK, Konstanz , pp. 163-177.

Heyer, G., Quasthoff, U., Wolff, Ch.: (2002):; Automatic Analysis of Large Text Corpora - A Contribution to Structuring Web Communities. In: Unger, Herwig; Böhme, Thoams,; Mikler, Armin (edd.) (2002). Innovative Internet Computing Systems. Proc. Second Internation al Workshop, Kühlungsborn, Juni 2002. Berlin et al.: Springer, 15-36 LNCS Vol. 2346.

Krenn, B. (2000): Distributional and Linguistic Implications of Collocation Identification. Proc. Collocations Workshop, DGfS Conference, Marburg.

Kunze, C. (2000): Extension and Use of GermaNet, a Lexical-Semantic Database. In: Proceedings of the Second International Conference on Language Resources and Evaluation, Vol. II, S. 999-1002. Athens, Greece.

Lemnitzer, L. (1998): Komplexe lexikalische Einheiten in Text und Lexikon. In: Heyer, G.; Wolff, Ch. (edd.). Linguistik und neue Medien. Wiesbaden: Dt. Universitätsverlag, pp. 85-91.

Maedche, A., Staab, St. (2001): Ontology Learning for the Semantic Web; IEEE Intelligent Systems 16, 2 , pp.72-79.

Manning, Ch. D., Schütze, H. (1999): Foundations of Statistical Language Processing; Cambridge/MA, London: The MIT Press.

Miller, G. A. (1990): Wordnet - an on-line lexical database. International Journal of Lexicography 3(4):235-312

Quasthoff, U., Wolff, Ch. (2000): An Infrastructure for Corpus-Based Monolingual Dictionaries. Proc. LREC-2000. Second International Conference on Language Resources and Evaluation. Athens, Vol. I, pp.241-246.

Schatz, B. (2002): The Interspace: Concept Navigation across Distributed Communities; IEEE Computer 35, 1 (2002), pp.54-62.

Smadja, F. (1993): Retrieving Collocations from Text: Xtract; Computational Linguistics 19, 1 pp143-177.

Steele, J. (ed.) (1990): The Meaning-Text Theory of Language: Linguistics, Lexicography, and Practical Implications, University of Ottawa Press, 1990.

Yarowsky, D. (1995): Unsupervised Word Sense Disambiguation Rivaling Supervised Methods, in Meeting of the Association for Computational Linguistics, pp. 189-196.