Experiments on Candidate Data for Collocation Extraction

Stefan Evert and Hannah Kermes
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
`{evert,kermes}@ims.uni-stuttgart.de`

Abstract

The paper describes ongoing work on the evaluation of methods for extracting collocation candidates from large text corpora. Our research is based on a German treebank corpus used as gold standard. Results are available for adjective+noun pairs, which proved to be a comparatively easy extraction task. We plan to extend the evaluation to other types of collocations (e.g., PP+verb pairs).

1   Introduction

While a mostly British tradition based on the ideas of J. R. Firth defines collocations as (significantly) frequent combinations of words cooccurring within a given text span, applications in terminology, lexicography, and natural language processing prefer a more restricted view. Collocations are understood as unpredictable combinations of words *in a particular (morpho-)syntactic relation* (adjectives modifying nouns, direct objects of verbs, or English noun-noun compounds). The extraction of such collocations from text corpora is usually performed in a three-stage process (cf. Krenn (2000, 28–32) and references therein):

1. The source corpus is annotated with varying amounts of linguistic information (ranging from part-of-speech tags to full parse trees), depending on the tools available. Then a list of word pairs satisfying the required (morpho-)syntactic constraints is extracted (typically based on part-of-speech patterns). This first candidate list will contain both collocational and non-collocational pairs.

2. Linguistic and/or heuristic filters may be applied to reduce the size of the candidate set. For instance, certain "generic" adjectives as well as those derived from verb participles are rarely found in adj+noun collocations.

3. The remaining candidates are ranked by statistical measures based on their frequency "profiles". Usually, word pairs are considered likely to be collocations if their cooccurrence frequency is much higher than expected by chance.

Authors typically evaluate the performance of a single collocation extraction system as a whole (e.g. Smadja (1993)). A small number of in-depth comparative evaluations (mostly Daille (1994) and Krenn (2000)) concentrate on the quality of the statistical measures and the corresponding ranking of the candidates, and to a lesser extent on the performance of linguistic filters. Although Evert and Krenn (2001) are aware of the influence that the first extraction step has on their results, they fail to give a quantitative evaluation of different pre-processing and extraction methods.

Our research aims to fill this gap. Currently, we are evaluating methods for the extraction of adjective+noun pairs from German newspaper text. It is planned to extend our work to other types of collocations, including PP+verb and noun+verb pairs.

2 Evaluation procedure

With a collocation definition that is not based purely on observed frequencies, the statistical ranking of candidates has to be evaluated against a manually confirmed list of true positives (cf. Daille (1994) and Krenn (2000)). This methodology is of little use for the evaluation of the candidate extraction step, though, for several reasons:

- The accuracy of the extraction step influences the final results in two quite different ways: (a) by changing the set of candidate types; (b) by changing their frequency profiles.

- The influence of changes in the frequency profiles depends crucially on the particular statistical measure applied in the third stage.

- In many cases, different extraction methods will produce only minor changes in the set of candidates, especially when frequency thresholds are applied. These subtle effects will be masked by the much greater impact of the statistical ranking.

- Simple extraction methods may find many spurious candidates which do not satisfy the required (morpho-)syntactic constraints. Even though some of those might be true positives per se (i.e. they are accepted as collocations by a human annotator), they are not a part of the source corpus and thus should not be included in the list of candidates.

Hence, it is necessary to evaluate the extraction step independently, and to find an appropriate definition of the expected goal *of the first processing stage*, i.e. what results should ideally be produced.

Clearly, one cannot expect the extraction step to distinguish collocations from non-collocations without access to frequency information. The frequency profiles of candidates should accurately report the number of co-occurrences in the source corpus, and spurious matches should be avoided. This leads to the following evaluation goal:

> Find all instances of word pairs that occur in a specific (morpho-)syntactic relationship in the source corpus.

As a consequence, our evaluation is based on instances of candidate pairs, i.e. *tokens* rather than *types*. In our terminology, a *pair type* is a combination of two words, and the corresponding *pair tokens* are the individual occurrences of this word pair at specific positions in the corpus. Statistical ranking methods are usually applied to and evaluated on pair types.

The experiments reported here investigate the extraction of German adjective+noun pairs, where the noun is the head of a noun phrase (NP) and the adjective appears as a modifier in the NP.

3 A gold standard

It is theoretically easy to obtain a gold standard for our evaluation, since the purely syntactic relationships that have to be annotated are less ambiguous than the distinction between collocations and non-collocational candidates. However, the annotation of *tokens* rather than *types* is a prohibitively laborious task. Fortunately, a German treebank corpus is available, from which the gold standard data can be extracted by automatic means. The Negra corpus[1] (Skut et al., 1998) consists of 355 096 tokens of German newspaper text with manually corrected part-of-speech tagging, morpho-syntactic annotations, and parse trees.

We used XSLT stylesheets to extract a reference list of 19 771 instances of adjective+noun pairs from a version of the Negra corpus encoded in the TigerXML format (Mengel and Lezius, 2000). Unfortunately, the syntactic annotation scheme of the Negra treebank (Skut et al., 1997), which omits all projections that are not strictly necessary to determine the constituent structure of a sentence, is not very well suited for automatic extraction tasks.

So far, we have only been able to extract adjective+noun pairs. We plan to use the TIGERSearch tool[2] in combination with stylesheets to obtain reference data for PP+verb and noun+verb pairs.

4 Pre-processing and extraction methods

In addition to the hand-corrected part-of-speech tags in the Negra corpus, we used the IMS Tree-Tagger (Schmid, 1994) for automatic tagging.

---

[1] http://www.coli.uni-sb.de/sfb378/negra-corpus/
[2] http://www.ims.uni-stuttgart.de/projekte/TIGER/

With its standard training corpus, a tagging accuracy of 94.82% was achieved. A substantial part of the errors are due to proper nouns missing from the tagger lexicon.

In the next step YAC, a recursive symbolic chunk parser (Kermes and Evert, 2002), was applied to identify adjective phrases (APs), noun phrases (NPs), and prepositional phrases (PPs). An evaluation of YAC against NPs extracted from the Negra treebank shows a precision of $P = 88.78\%$ and a recall of $R = 90.80\%$ based on the hand-corrected tagging. With automatic tagging, $P = 82.33\%$ and $R = 86.15\%$ are achieved.

YAC was specifically designed for automatic extraction: all AP and NP projections are made explicit and annotated with head lemmas, which simplifies candidate extraction with XSLT stylesheets tremendously. We created two versions of the chunk annotations, based on the hand-corrected and the automatic tagging, respectively.

Finally, we used three common extraction methods to identify candidate pairs: (a) adjacent adjectives and nouns (based on part-of-speech tagging); (b) adjectives preceding nouns within a given window; (c) the lexical heads of APs and NPs in the chunk annotations, where the AP node is a child of the NP node.[3] In (b), only the adjective nearest to each noun was used, and no verbs, sentence-ending punctuation, or nouns were allowed in between. We arbitrarily chose a window size of 10 tokens for this experiment. Further tests confirmed that the evaluation results depend only minimally on the exact size of the extraction window.

We have evaluated all six combinations of pre-processing and extraction methods. In further experiments, we plan to study the quantitative effects of linguistic filters (excluding adjectives derived from verb participles and/or proper nouns) and lemmatisation (wrt. candidate *types*).

5 First results

The reference data extracted from Negra comprises 19 771 instances of adjective+noun pairs. The numbers for automatic extraction range from 17 694 (adjacent pairs based on automatic tagging) to 19 726 (YAC chunks on hand-corrected tagging). Table 1 lists *precision*[4] and *recall*[5] for all combinations of pre-processing and extraction methods. On the hand-corrected tagging, adjacent pairs yield the highest precision, but recall is much better for extraction from windows or YAC chunks. The 5% error rate of the automatic tagging reduces the extraction accuracy by approximately the same amount. The chunk-based extraction is slightly less sensitive to tagging errors and achieves both best precision and best recall in this realistic scenario.

Because of the small size of the Negra corpus, the observed cooccurrence frequencies of pair *types* rarely differ from the reference values by more than 1. The few substantial differences are mostly due to systematic errors in the automatic tagging, e.g. *Joe Cocker* as a false positive (*Joe* is wrongly tagged as an adjective) and *Rotes Kreuz* ("Red Cross") as a false negative (*Rote(s)* is wrongly tagged as a noun).

Unsurprisingly – considering the large number of hapaxes among the candidates – there are still considerable differences between the automatically extracted sets of pair types and the gold standard. Our gold standard contains 16 112 different pair types, whereas numbers for automatic extraction range from 14 782 to 16 056. The best results for YAC chunks on perfect tagging include 660 pair types that are not found in the reference data, while 716 pair types were missed by the automatic extraction. These differences are of little practical importance, though, since they mostly affect low-frequency types for which statistical association measures are not reliable. Most applications will set a frequency threshold to exclude such types.[6]

6 Conclusion

The extraction of German adjective+noun pairs has proven to be a comparatively easy task. Depending on tagging quality, almost perfect results can be obtained. Moreover, even with a straight-

---

[3]These candidates were extracted from the XML output format of the chunker with a simple XSLT stylesheet.

[4]*precision* = proportion of correct pair tokens among the automatically extracted data

[5]*recall* = proportion of pair tokens in the reference data that were correctly identified by the automatic extraction

[6]Interestingly, the popular t-score measure (Church and Hanks, 1990) effectively sets a frequency cut-off threshold when only the *n* highest-ranking candidates are extracted.

candidates from	perfect tagging		TreeTagger tagging	
	precision	recall	precision	recall
adjacent pairs	98.47%	90.58%	94.81%	84.85%
window-based	97.14%	96.74%	93.85%	90.44%
YAC chunks	98.16%	97.94%	95.51%	91.67%

Table 1: Results for Adj+N extraction task

forward stochastic tagger and naive window-based extraction precision and recall values above 90% provide an excellent starting point for statistical ranking.

The considerable differences between the sets of pair types primarily affect hapaxes and have little impact on statistical methods for collocation identification (where hapaxes are rarely found among the higher-ranking candidates). Likewise, small changes in the frequency profiles of more frequent pairs have little impact on the association scores and the precise ranking of the candidates. It will be interesting to see how these results translate to more demanding extraction tasks.

References

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

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

Stefan Evert and Brigitte 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.

Hannah Kermes and Stefan Evert. 2002. YAC – a recursive chunker for unrestricted german text. In Manuel Gonzalez Rodriguez and Carmen PazSuarez Araujo, editors, *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC)*, volume V, pages 1805–1812, Las Palmas, Spain.

Brigitte Krenn. 2000. *The Usual Suspects: Data-Oriented Models for the Identification and Representation of Lexical Collocations.* DFKI & Universität des Saarlandes, Saarbrücken.

Andreas Mengel and Wolfgang Lezius. 2000. An XML-based representation format for syntactically annotated corpora. In *Proceedings of the Second International Conference on Language Resources and Engineering (LREC)*, volume 1, pages 121–126, Athens, Greece.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Wojciech Skut, Brigitte Krenn, Thorsten Brants, and Hans Uszkoreit. 1997. An annotation scheme for free word order languages. In *Proceedings of the Fifth Conference on Applied Natural Language Processing ANLP-97*, Washington, DC.

W. Skut, T. Brants, B. Krenn, and H. Uszkoreit. 1998. A linguistically interpreted corpus of german newspaper texts. In *Proceedings of the ESSLLI Workshop on Recent Advances in Corpus Annotation*, Saarbrücken, Germany.

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