Information Extraction from Text Corpora: Using Filters on Collocation Sets

Gerhard Heyer, Uwe Quasthoff, Christian Wolff

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

Abstract

This paper describes the application of filtering techniques to collocation sets calculated for very large text corpora. Additional information like patterns, grammatical information, subject areas and numerical values associated with the collocations are used to identify collocations with given semantic structure. Various examples and different techniques for applying such filters are described. We also give several examples of practical applications for this type of information extraction.

1. Introduction

Within he last years we have been developing an infrastructure for the analysis of large text corpora of different Indo-European languages, following an integrated approach which combines corpus data with lexicographic information like linguistic categories or semantic attributes. We have studied various approaches towards information extraction from text corpora to be used for academic as well as industrial applications in domains like knowledge management, ontology engineering or information retrieval. Among them, one of the most powerful is the calculation and further analysis of collocation sets for all concepts identified in a text corpus. While the general setup of our corpus analysis infrastructure has already been presented, in this paper we want to concentrate on information extraction using different types of filters.

For tests and examples we used the German and English corpora of more than 10 million of sentences each. Both the corpora and the collocations can be found at *http://wortschatz.uni-leipzig.de*.

2. Collocation sets as unstructured information pools

In this paper, the term collocation means a pair of words (or multi word terms) which appear significantly often within a given text window. The two kinds of windows we are exploring are the sentence and immediate (left or right) neighborhood. The collocations found depend on the statistics for the calculation of the significance. For the calculations in this paper, we use a Poisson measure (see Chung 2000) which gives similar results to the well known Log-likelihood measure (Krenn 00) but can be shown to have some nice mathematical properties justifying the name measure (cf. Heyer et al. 01). The filtering methods described below do not depend on the method used for calculating the collocations, though. Usually, collocations are given together with a significance value.

Given a large text corpus, collocations for all concepts may be calculated suing highly efficient software tools (see Heyer, Quasthoff, Wolff 00). The resulting collocation sets may easily be interpreted by humans, but they are hardly apt for further automatic processing, as collocations represent different types of relations between concepts: They may be interpreted using syntactic as well as semantic categories and with respect to the direction of analysis, different sets of relation types may be applied. In the following example the set of sentence-based collocations for the German word *Zylinder* is given:

Frack (217), Kolben (113), Gehrock (71), Ventile (58), Hubraum (48), Motor (43), Kaninchen (42), PS (39), Kegel (32), Kolbens (28), Luft (27), schwarzen (27), Frischgase (26), Kubus (25), schwarzem (25), Einspritzung (22), Quader (22), Ventil (22), Dampf (21), Kugel (20), Liter (18), Motorblock (18), Altgase (17), Auslaßventile (17), Würfel (17), Zylinders (17), bewegt (17), strömt (17), Spazierstock (16), rotierenden (16), Ansaugen (15), Auslasskanal (15), Durchmesser (15), Luft-Kraftstoff-Gemisch (15), Ventilen (15), Grundformen (14), Konus (14), Kraftstoff (14), [...]

It can easily be seen that the basic fact of significant co-occurrence within a sentence can be interpreted in various ways: Concepts occurring in the collocation set identifying different meanings of "*Zylinder*" like "*Frack*" (tail-coat) or "*Gehrock*"(cutaway) hinting at the meaning of "*Zylinder*" as a type of hat worn on formal occasions; "*Quader*" (cuboid), "*Kugel*" (sphere), and "*Würfel* (cube) as in interpretation of "*Zylinder*" as a geometrical form; "*Ventile*" (valves), "*Motor*", "*Einspritzung*" (injection) giving the interpretation of "*Zylinder*" as part of an engine.

1. Typical *attributes* of "*Zylinder*" like "*schwarzem*" (black) or "*rotierendem*" (rotating).
2. Typical *verbs* (i. e. activities or operations) associated with the concept ("*bewegt*" (moves). "*strömt*" (streams).
3. Part-of-relationships: "*Zylinder*" being part of the "*Motorblock*" (engine block) as well as the "*Motor*"
4. Co-Hyponymy-relationships for concepts having a common broader term (i. e. concepts appearing as neighbours at the same level in a classification tree).

While these examples for the interpretation of collocations are taking into account linguistic information as well as knowledge of the world, it is obvious that this kind of further analysis of collocation sets requires additional information in order to be performed automatically.

Moreover, the structure of a collocation set of a word strongly depends both on semantic and syntactic properties of this word. The following chapter introduces differ-

ent filtering methods as well as the analysis results to be obtained by their application.

3. Filtering Methods

The general approach to analysing collocation sets is that of applying well-defined filters to a set of collocations aiming at the extraction of a pre-defined type of information. According to the type of filter applied to the collocation set, the outcome may be interpreted in various ways. The following examples for filters only show some of the possibilities given with this approach:

1. Analysis of the *numeric value* of our collocation measure: The comparison of relative collocation strengths as well as its additivity property both may be used as filters.
2. *Positional information* and typical *patterns* for well-defined types of information like person names and titles or company and product names can be applied as a filter.
3. *Additional knowledge* given by *categorical* and *part-of-speech* information: Given an additional dictionary in which information on possible syntactic categories for each concept in the corpus is stored, the typical adjectives (features) or verbs (operations, activities) associated with a concept may be filtered from the set of its collocations, given that the concept is a noun. Additionally, part-of-speech (POS) information generated with state-of-the-art pos taggers can be used as filter as well.
4. *Additional knowledge* about *categories* of named entities or given by *subject area codes*: For generic features and functions, a list of typical representatives can be extracted. Good examples are names for professional functions; among their right neighbor collocations typical representatives appear which can be filtered out, if information on what may be a name is given.
5. Domain specific collocation sets from another corpus: The various interpretations of *Zylinder* in the example mentioned above result from the analysis from a very large general language corpus. If the same analysis is run on a more specific corpus (e. g. containing technical texts on automobiles) a more specific set of collocations will be calculated which can be used as a filter for the original, polysemic collocation set.
6. Applying set operations on collocation sets: Calculating the intersection of collocation sets for two different concepts typically yields a secondary set of concepts containing concepts that have something in common with both of the starting concepts. This operation can ideally be of direct use for question answering:

- Using groups of sentence collocations

Depending on the structure of the collocation set these filters can be applied in different scenarios. The following problems will be addressed:

- Extraction of multiwords and phrases;
- Inference of semantic categories and subject areas;
- Identifying the type of relation between words;
- Dealing with polysemy (see also ch. 3.6 below).

While it is clear that there are many practical problems to be solved with this approach (How are concepts defined? How are phrases or multi term concepts handled? In what way is generality or domain-specificity of a corpus defined?), we strongly believe that this general approach towards interpreting and filtering of collocation sets can be of great value in various fields of application In the following subsections, the above-motioned types of filtering are presented in more detail. It should be noted that all examples below are directly derived from the analysis of our reference corpora, which are available online at *http://www.wortschatz.uni-leipzig.de*.

3.1. Filtering by Analyzing Numeric Values of Collocation Measures

The first two types of filtering operations make use of the values of the collocations measure itself, either for detection of multiwords, or for the analysis of polysemous words.

3.1.1. Relative Collocation Strength: Extracting Multiwords and Phrases

If for a word *A* there is a (left or right) neighbor collocation *B* with

- large collocation value compared to all other collocations of *A* and
- collocation value near to the maximum possible for *A*,

then there is a good chance that *BA* (or *AB*) form a multiword term or phrase. Examples found this way are *20th Century-Fox, Agents provocateurs, abdominale Hernie,….*

Note that the above relation is not symmetric in *A* and *B*. For example, the terms *Corpus iuris, Corpus ventriculi, Corpus mandibulae, Corpus juris, Corpus uteri* can only be identified starting from the right component.

3.1.2. Quantitative Analysis of Polysemy

Given a polysemous word, usually one knows the different meanings, but there is no quantitative information assigned to the different meanings. Such information is interesting on its own, but also useful for Machine Translation or Information Retrieval.

Assume a given polysemous word *A* has *n* senses. We are interested in the probabilities p_i, $i = 1, \ldots, n$, for the different senses. Additivity of our collocation measure (see Heyer et al. 01) allows us to estimate the probabilities p_i as follows: First, assign the appropriate sense of *A* to each sentence collocations of *A*, if possible. Ignore collocations which did not get a sense in this step. Second, add the collocation measures for each sense. Then normalize to get p_i.

EXAMPLE: There are basically three meanings of *space* to be found in ordinary texts like newspaper articles. Most of the collocates of *space* belong to exactly one of these senses. The top collocations of space together with their measures, ordered by sense, are

Sense 1: *Outer space*. Total measure 11044: shuttle (2618), station (991), NASA (920), Space (602), launch (505), astronauts (473), Challenger (420), manned (406), mission (385), Discovery (341), Mir (335), rocket (329), orbit (326), NASA's (297), flight (293), Atlantis (291), cosmonauts (275), Earth (239), satellite (238), satellites (203), outer (193), orbiting (188), telescope (176),

Sense 2: *Computer*. Total measure 4910: disk (2629), memory (718), storage (479), hard (336), RAM (307), files (261), bytes (180),

Sense 3: *Real estate*. Total measure 3715: square (1163), feet (822), leased (567), lessor (390), office (382), lessee (201), heating (190),

Without assignment: address (653), represented (412), program (308), free (300), amount (230), requires (223), Cornish (209), virtual (198), desk (182).

Hence, we get the following approximate weights for the three senses:

Sense 1: Outer space; p_1=0.56
Sense 2: Computer; p_2=0.25
Sense 3: Real estate; p_3=0.19

These probabilities depend on the corpus and the additivity assumes independence of the collocates, which is never strictly fulfilled in language.

For this type of filter, a visualization strategy may be applied: Again we consider the collocates of a polysemous word *A*. We can expect that for a given sense, some of the collocations *B* and *C* belonging to this sense, are also collocations of each other. Hence, we have a triple (*A*, *B*, *C*) of collocations. To visualize the polysemy of *A*, we take the collocation set of A and remove all collocates which are not element of a triple as above. Next we use simulated annealing for a planar representation (see Davidson & Harel 96).

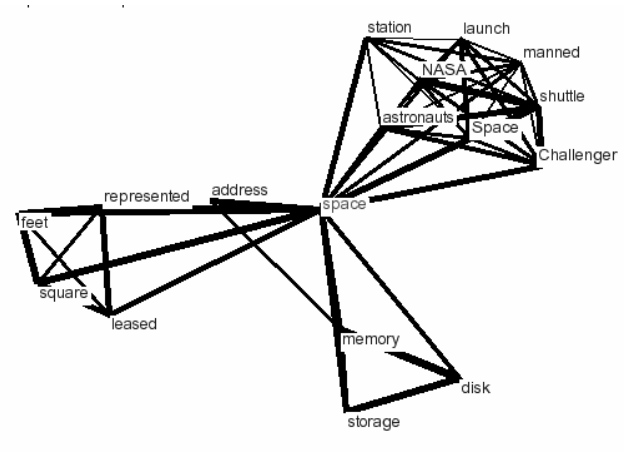

Figure 1: Visualization of the collocations of space

3.2. Filtering Using Positional or Pattern Information

Beyond relative collocation strength, the introduction of simple positional or pattern-based information is a next step in refining collocation analysis. Positional information is already taken into account in the calculation of immediate neighborhood collocation sets. Thus, it is used in many other filtering techniques described in subsequent chapters. A more complex example is given in ch. 3.2.1 for multiword extraction.

3.2.1. Pattern-based Multiwords

A person usually is referenced by an optional title, first name(s) and a surname. From this, we can find the following rules:

- Surnames are typical *right next neighbor collocations* of first names
- Both titles and first names are *left neighbor collocations* of surnames
- Only titles are left neighbor collocations of first names.
- In many languages, determiners are left neighbor collocations for titles, not for surnames.

These rules allow the extraction of large lists of persons together with their titles. Similar rules apply to company names. Company names, in turn, allow for the extraction of product names.

3.3. Filtering by Using Additional Syntactic or Semantic Knowledge

One of the most important types of filtering operation sis the introduction of additional lexical (syntactic, semantic) knowledge for analysis of collocation sets. In the simplest case, collocation sets are reduced to the noun category, excluding all other word categories which may be used in the automatic generation of concept networks. More interesting filtering techniques using additional knowledge are described below.

3.3.1. Category Detection with Keywords

For some types of proper names, its category or class identifier is given as part of the name or given in the text as direct neighbor This applies to left neighbors if *Island* (English corpus) and right neighbors of *Insel* (German corpus), where more than 200 islands can be identified each. The same applies to other geographic identities like regions, cities, etc.

Other category-keywords with category-elements as neighbors are *trademark*, professions like *tenor* or *minister*, can be used analogously. But, there seems to be no comprehensive list of such category words in the literature.

3.3.2. Standard Relations and Meaning-Text Theory

Meaning-Text Theory [Steele 90] describes more then 70 semantic relations between words. Some of the more important relations are the so-called standard relations like

- typical properties of Objects, usually given by an adjective-noun-relation
- typical participants and objects of an action, usually given by verb-noun-relations

To distinguish between subject and object we can use word order. Even in German, where word order in a sentence is less restrictive than in English, we typically have the default word order SPO. Hence, left neighbors for a verb are good candidates for typical subjects, right neighbors are good candidates for typical objects.

If we have grammatical POS information for the collocations, we can translate this to semantic information as, for example, as follows: Given a noun, we search for adjectives as left next neighbor collocations. We interpret the result as typical properties of tat noun.

EXAMPLE: Adjectives as left neighbors for *spacecraft* describe standard properties: *unmanned, manned, robot, winged, planetary, orbiting, long-duration, low-orbiting, reusable, alien, earth-orbiting, nuclear-powered, exploratory, ...*

3.3.3. Subject Area Descriptors in the Collocation Set

In technical documents, terms of the same subject area appear together. Hence, collocations often have the same subject field. This can be used to infer the previously unknown subject area of a word: If many of its collocates belong to a certain subject area, the starting word might belong to the same. Additional part-of-speech informa-

tion is helpful, because this rule works best for nouns. For verbs and adjectives it may simply indicate a special usage of this word.

EXAMPLE: The top noun collocates for *orchestra* belong to the subject area music: *conductor, music, symphony, Symphony, chorus, musicians, concert, Orchestra, concerts, opera, soloist, …*

3.4. Filtering by Applying set Operations on Collocation Sets

The last type of filtering operation is the application of set operations to collocation sets for different words.

3.4.1. Intersection of Collocation Sets

Given two words, or concepts, A and B, the intersection of their collocation sets is likely to reveal typical relations holding between A and B, if any. Thus, the intersection of collocation sets may be used for a simple type of question-answering: Given a binary relation of the type *A*(*B*, *C*), the intersection of the collocation sets of either *A* and *B*, *A* and *C*, or *B* and *C* will yield possible candidates for the missing third variable as in the following example for the relation *Oberbürgermeister(Name, Stadt)* (i. e. mayor(name, city) for which the correct "answer" in the case of *Munich* is *Oberbürgermeister(Christian Ude, München)* :

A(Christian Ude, München)	*Oberbürgermeister(B, München)*	*Oberbürgermeister(Christian Ude, C)*
Oberbürgermeister (3496), OB (2139), Bayern (1185), bayerische (534), Oberbayern (489), Landeshauptstadt (476), CSU (449), Stadt (446), Bayerns (376), bei (373)	in (8345), **Christian Ude** (3496), Frankfurt (1075), nach (1024), Stuttgart (853), Bayerischen (832), am (820), bayerischen (744), Köln (732), Stadt (552)	OB (2139), SPD (887), Stadt (552), Peter Menacher (263), Rathaus (240), CSU (206), **München** (200), hat (175), Münchner (173), Stadtrat (157)

Table 1: Intersection of Collocation Sets Used for Extracting Unknown Variables in Relations

As may easily be seen from this example, the correct variable instantiation does not necessarily appear in the first place, given the intersection of collocation sets ordered by strength, but may be inferred if additional knowledge about the type of information to be extracted is present (function name, proper name, geographical identifier).

3.4.2. Union of collocation sets

The union of two collocation sets of words *A* and *B* (with addition of the significance measure for collocates of both *A* and *B*) corresponds to the unification of the words *A* and *B* to a new, abstract concept *C*. If necessary, we unify the corresponding words in the collocation sets as well. As an obvious example, we can unify all first names to an abstract category [first_name] and all surnames to a category [surname] and get the result that [surname] is the strongest right neighbor of [first_name].

In the German corpus we introduced about 100 such abstract categories and found collocational relations like the following:

The strongest collocations of the concept [championship] are [medal], [win], [country_nom] and [country_adj].

In this example, the concepts are defined by the following German words, ordered by frequency:

[championship]: *Qualifikation, Meisterschaft, Olympia, Weltmeisterschaft, …*

[medal]: *Gold, Silber, Medaille, Bronze, Medaillen, Goldmedaille,*

[win]: *gewinnen, gewonnen, gewann, gewinnt, gewannen*

[country_nom]: *Deutschland, USA, Frankreich, Italien, China, Japan, …*

[country_adj]: *deutschen, deutsche, französischen, deutscher, französische, englischen, englische,*

In the above example we see how these concepts form relations as in the following sentence pattern:

The team (maybe of a country given by [country_nom] or [country_adj]) [won] a [medal] at a [championship]. Moreover, [country_nom] can also denote the place of the championship.

3.5. Goal-driven Combination of Filters

As should already have become clear from the discussion of the various types of filters, in most cases not a single type of filter will yield the desired result. Rather, the goal-driven combination of applicable filters will be satisfying. While subsections 3.1 – 3.4 discussed filtering methods from the perspective of filtering operation type, the following table is organized according to the goal of information extraction or corpus analysis. It should be noted, though, that the starting point for all filtering operations are collocation sets (sentence and left and right neighbor collocations) automatically calculated for a given corpus.

Type of Information Extracted	*Type of Filtering Operation(s) Applied*
Multiword Detection	Positional information, patterns, category information
Analysis of polysemy	Subject area information
Class – instance relationships; categorization of named entities	Positional information, knowledge about class names, category information
Naming of semantic relations	Positional and categorical information; set operations
Subject area detection for unclassified words	Subject area information, category information
Ontology and concept hierarchy generation	Set operations on collocations sets, categorical information

Table 2: Applicable Filtering Operations for Different Extraction Goals

4. Applications

Introducing different filtering techniques goes beyond the basic layer of statistical corpus analysis; filters may be applied to any given text corpus in a toolset like fashion. It

is obvious that various fields of application exist in areas as diverse as information retrieval or object-oriented reengineering using text mining methods. Two applications of this approach shall be mentioned; both of them have already been applied in industrial projects.

4.1. Knowledge Re-engineering

Building ontologies and semantic networks for very specialised domains can be a very time-consuming effort when being performed by intellectual analysis of the domain only. Various standards for ontology markup and description have been defined (DAML; Topic Maps; OIL, see Maedche & Stab 01). Still, a methodology for automatic ontology construction is an important academic as well as practical desire. Using our approach towards corpus analysis which calculates collocation sets for any given text corpus, the application of various filters results in a raw semantic net or ontology which represents knowledge found in the text corpus. Although manual refinement still has to be applied, the process of knowledge extraction is at least partially performed automatically. We employ the topic map standard (ISO/IEC 13250, see Biezunski & Newcomb 01) for representing these "raw semantic networks" (for a detailed discussion see Böhm et al. 02). An export tool generates Topic Maps in a XML format which is ready for postprocessing by Topic Map editors or knowledge management tools.

4.2. Object-oriented Reengineering

Categorical filters for the extraction of typical adjectives or verbs going along with a noun can be interpreted as typical properties and methods as applied in object-oriented analysis and design. The same holds for generic concepts and typical representatives when seen as a class - object (instantiation) relationship. This kind of analysis has been successfully applied to text corpora describing large software project. The analysis results can directly be used in OO-tools as a starting point for a new software model (see Heyer, Quasthoff, Wolff 02 for further details).

5. Outlook

While our approach towards filtering of collocation sets has already been applied in number of projects, three major lines of further development can clearly be identified:

- The development of a *general theory of information filtering and extraction* based on a multi-level approach comprehending statistical corpus analysis as a baseline process upon which knowledge-based filters as described in this paper may operate: Starting from a *generic idea of relatedness* which is represented by the calculation of collocations, additional knowledge or various kinds allows for a fine-grained in-depth analysis.
- The *collection of additional knowledge* to be used in filtering like syntactic or semantic information. The examples given are derived from our German corpus as for this language our knowledge base is already comparatively large.
- The *optimisation of software tools* for corpus analysis and comparison which shall allow for a straightforward application of this kind of analysis as well as its integration in ontology or knowledge management software (e. g. software for defining and handling topic maps).

6. References

[Armstrong 93]. Armstrong, S. (ed.); "Using Large Corpora"; Computational Linguistics 19, 1/2 (1993) [Special Issue on Corpus Processing, repr. MIT Press 1994].

[Biezunski & Newcomb 01] "XML Topic Maps: Finding Aids for the Web"; IEEE Multimedia 8, 2 (2001), 108.

[Böhm et al. 02]. Böhm, K.; Heyer, G.; Quasthoff, U.; Wolff, Ch. "Topic Map Generation Using Text Mining." In: Proc. IKNOW '02 – Intl. Conf. on Knowledge Management, Graz, to appear.

[Chung 00]. Chung, K. L. "A Course in Probability Theory", Academic Press 2000.

[Davidson & Harel 96]. Davidson, R., Harel, D. (1996). "Drawing Graphs Nicely Using Simulated Annealing." In: ACM Transactions on Graphics 15(4), 301-331.

[Heyer et al. 01] Heyer, G.; Läuter, M.; Quasthoff, U.; Wittig, Th.; Wolff, Ch.; "Learning Relations using Collocations"; In: Proc. IJCAI Workshop on Ontology Learning, Seattle/WA, August 2001, 19-24.

[Heyer, Quasthoff, Wolff 00] Heyer, G.; Quasthoff, U.; Wolff, Ch.; "Aiding Web Searches by Statistical Classification Tools. " Proc. Proc. 7. Intern. Symposium f. Informationswissenschaft ISI 2000, UVK, Konstanz (2000), 163-177.

[Heyer, Quasthoff, Wolff 02] Heyer, G.; Quasthoff, U.; Wolff, Ch.; "Automatic Analysis of Large Text Corpora - A Contribution to Structuring Web Communities. " In: Proc. I2CS 2002, Rostock, Germany, to appear in LNCS.

[Krenn 00] Krenn, B.; "Distributional and Linguistic Implications of Collocation Identification." Proc. Collocations Workshop, DGfS Conference, Marburg, March 2000.

[Lemnitzer 98] Lemnitzer, L.; "Komplexe lexikalische Einheiten in Text und Lexikon." In: Heyer, G.; Wolff, Ch. (edd.). Linguistik und neue Medien. Wiesbaden: Dt. Universitätsverlag, 1998, 85-91.

[Maedche & Staab 01] Maedche, A.; Staab, St.; „Ontology Learning fort he Semantic Web"; IEEE Intelligent Systems 16, 2 (2001), 72-79.

[Manning & Schütze 99]. Manning, Ch. D.; Schütze, H.; Foundations of Statistical Language Processing; Cambridge/MA, London: The MIT Press 1999.

[Quasthoff & Wolff 00] Quasthoff, U.; Wolff, Ch.; "An Infrastructure for Corpus-Based Monolingual Dictionaries." Proc. LREC-2000. Second International Conference on Language Resources and Evaluation. Athens, May/June 2000, Vol. I, 241-246.

[Schatz 02] Schatz, B.; "The Interspace: Concept Navigation across Distributed Communities"; IEEE Computer 35, 1 (2002), 54-62.

[Smadja 93] Smadja, F.; "Retrieving Collocations from Text: Xtract"; Computational Linguistics 19, 1 (1993), 143-177.

[Steele 90] Steele, J. (ed.): The Meaning-Text Theory of Language: Linguistics, Lexicography, and Practical Implications, University of Ottawa Press, 1990.