Text Mining for the Extraction of Domain Relevant Terms and Term collocations

Daniela Kurz
XtraMind GmbH, Germany
Stuhlsatzenhausweg 3
D-66123 Saarbrücken
Germany
Email: kurz@xtramind.com

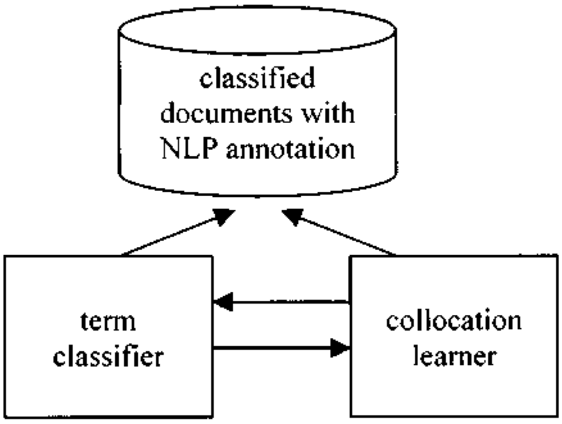
Feiyu Xu
DFKI GmbH
Language Technology Lab
Stuhlsatzenhausweg 3
D-66123 Saarbrücken
Germany
Email: feiyu@dfki.de

1. Introduction

The domain adaptation capability of information extraction (IE) systems relies on automatic acquisition of domain specific knowledge. The domain specific knowledge contains domain relevant terms, semantic relations for ontology building, or lexico-syntactic patterns for template filling [Riloff & Jones 1999 and Yangarber et al 2000]. Recently, an ever-growing interest in automatic term extraction methods in NLP [Church & Hanks 1989, Smadja 1994, Daille 1996 and Evert & Krenn 2001] has been observed. In this paper, we present an approach to automatic acquisition of single-word terms, multi-word terms and collocations by taking classified documents as input. A word in our approach corresponds to a token unit after the text tokenization. A single-word term is a term consisting of a single word, whereas a multi-word term normally consists of more than one word. By collocations, we consider combinations of words that are not only lexically determined but also semantically related words. Our method is based on the integration of term classification methods and statistical measures for word association. It exhibits that very good results may be achieved on training corpora of different sizes. In particular, we can handle free word-order languages like German using special term collocation techniques. Thus all combinations of elements in a collocation candidate are allowed instead of using a window of predefined size.

2. The System

Our system contains two main components: a specific TFIDF-based term classification tool and a collocation learner [Xu et al 2002]:



In our approach we pursue a less linguistics-driven candidate selection (for candidate selection see section 4), but a more statistics-driven selection. In the literature, approaches can be found, that take into account morphosyntactic and syntactic properties to different degrees, while frequency-based approaches exist [Krenn, 1999]. Besides the frequency of the candidates, we take term classification into consideration. The combination of the term extraction and learning of multi-word terms and collocations can be done in several ways.

The term classifier identifies the domain relevant terms by taking classified documents as input. We use the term classifier as our initial component for extracting the single-word terms. However, in our system, the term classifier and the collocation learner can work on the one hand independent of each other, on the other hand, either of them can take the output of another component as their input. In our work, the collocation learner takes the output of the term classifier as its input in the initial run. Then the learned multi-word terms will be further input of the term classifier. Our bootstrapping algorithm works as follows:

Input: classified documents enriched with linguisticinformation

Step 1: term classification

a. Initial loop: extraction of single-word terms
b. After the first loop: classification of the multi-word terms

Step 2: learning multi-word terms and term collocations

Step 3: enrich the training corpus with multi-word terms and collocations from step 2 and go to Step 1 b.

There are several possibilities for the initial step. Instead of taking the results of the term classifier as input, the collocation learner takes the linguistically annotated data as input. For the linguistic annotation (stemming, pos-tagging, named-entity, phrase recognition) of the corpus, we use SPPC [Piskorski & Neumann, 2000]. In parallel relevant terms are identified by the term classifier and collocations are learned by the collocation learner. In the next step the learned multi-terms and collocations are matched against the extracted terms. Instead of using the frequency of the candidates as threshold a weighted combination of both (frequency and term weight) can be used to determine the right collocation candidates. The training corpus will be enriched by the learned data and the following steps will be same as described above.

3. Term Extraction

As we work on classified documents, the extraction of relevant terms is done by the so-called KFIDF [Xu et al 2002] measure. KFIDF is a modification of the TFIDF measure [Salton 1991].

$$KFIDF(term, cat) = docs(term\ cat) \times LOG(\frac{n \times |cats|}{cats(term)} + 1)$$

docs(term, cat) = number of documents in
the category containing the *term*
n = smoothing factor
cats(term) = the number of categories in
which the term occurs

A term is regarded as relevant if it occurs more frequently than other terms in a certain category, but occasionally elsewhere. In the following, we will give some examples of the extraction of single-word terms. The term classifier deploys part-of-speech tagging in advance, since we considered only the part of speeches: adjective, noun and verb. The words are normalised to their lemma forms. In the case of nouns, we considered the fullforms and did not perform a decomposition analysis, since especially compositions are domain specific and decomposition would lead to opposite result we want to achieve. We could observe that depending on the domain, a particular kind of part of speech will play a dominant role for the domain. In the *drug* domain, the relevant terms are nouns whereas in the domain of *management succession* verbs clearly dominate.

As example of the top-scoring extracted terms in the *drug* and the *management succession* domain are given in 1) and 2).

1) *Haschisch 79.13055*
[hashish]
Droge 55.192017
[drug]
Marihuana 55.151592
[marihuana]
Rauschgift 53.61485
[drug]
Kilogramm 52.038185
[kilogram]
Marktwert 51.142445
[market price]
Heroin 48.095898
[heroin]
Kokain 44.153614
[cocaine]
Schwarzmarktwert 40.913956
[black-market price]
Konsument 32 390213
[consumer]
Ecstasy-Tabletten 28.774744
[ecstasypills]

2) *berufen 38.45143*
[appoint to]
wahlen 35 155594
[choose]
ubernehmen 32.95837
[accept]
bestellen 28.56392
[nominate]
verlassen 20.873634
[leave]
wechseln 19.77502
[change]
ausscheiden 17.577797
[resign]
nachfolgen 15.380572
[succeed]
zurucktreten 12.084735
[resign]
antreten 8.788898
[assume office]

A comparison between the KFIDF measure and the TFIDF measure clearly shows that KFIDF delivers better results. Moreover the scores KFIDF delivers are much more fine grained than the scores of TFIDF. This becomes evident from the results shown in 3). These are some top-scoring terms extracted using TFIDF for the drug domain.

3) *Paket-Kurier-Sendungen 1 0*
[courier parcel]
Entzugserscheinungen 1 0
[withdrawal symptoms]
Anfall 1.0
[fit]
Stengel 1 0
[stick]
Ostafrikaner 1.0
[East African]
Nutzern 1.0
[user]
Hilfssubstanz 1.0
[supply]
Feuerloescher 1 0
[fire extinguisher]
Gesundheitsministerium 1.0
[Health Department]
Rauschgiftbande 1.0
[narcoticsring]

4. Learning term collocations

The objective of term collocation is the identification of multi-word terms and learning collocations that are lexically determined, e.g. *zur Verfugung stehen.* We also took into account semantically related words. The kind of semantic relation is not restricted to a certain type, since in this application we do not make use of lexico-syntactic patterns indicating a certain semantic relation.

Applying our system to the drug domain we learned synonyms such as

haschisch and *marihuana* or *kilo* and *kilogram* and hyponyms such as *drug* and *heroin.* Due to the free word-order characteristic of German, we considered all possible pairs in a sentence ignoring the linear order. Therefore we regarded the sentence as span. All main verbs and adjectives are reduced to their stem forms, nouns are kept in their fullforms for reasons already explained in section 2.
We used following association measures: Mutual Information [Church & Hanks, 1989], Log-Likelihood Measures [Daille, 1996], and T-test [Manning & Schutze, 1999]. Let us give a short explanation of the different measures we used:

Assocation measures

Mutual Information is defined as follows:

$$\mathrm{I(x,y)} = \frac{\log_2 P(x,y)}{P(x)P(y)}$$

where *P(x,y)* denotes the joint probability and *P(x)* and *P(y)* denote the probability of *x* and *y* separately. This association measure assumes that the occurrence of one word predicts the occurrence of another one. If there is an interesting relationship between *x* and *y,* the mutual information is expected to increase. We observed as mentioned in (Manning & Schutze, 1999) that mutual information is not practical when dealing with sparse data.

The definition of **Log-Likelihood** is given bellow:

$$\begin{aligned}&\mathrm{LogLike(x,\ y)} = \\ &a\log a + b\log b + c\log c + d\log d \\ &-(a+b)\log(\mathrm{o}+b)-(a+c)\log(\mathrm{a}+c) \\ &-(b+d)\log(b+d)-(c+d)\log(\mathrm{c}+d) \\ &+(a+b+c+d)\log(a+b+c+d)\end{aligned}$$

with *a, b, c* and *d* being elements of the contingency table of words *x* and *y* occurring with each other or not, e.g. *a* stands for the frequency of pairs involving both *x andy* etc. This measure tells us how much more likely the occurrence of one pair is than the occurrence of another one.

T-test is defined as:

$$\mathrm{T} = \frac{x-\mu}{\sqrt{\frac{s^2}{N}}}$$

where *x* denotes the sample mean, μ the mean of the distribution, s^2 the sample variance, *N* the sample size. This test tells how probable or improbable it is that a certain constellation occurs. The null hypothesis assumes that the occurrence of the two terms is independent. The T-test value tells us, if this hypothesis can be rejected or not.

Results

We focused on the extraction of noun-noun, verb-noun and adjective-noun combinations. By looking at the precision values of the statistical measures, we can confirm the results from other studies (Krenn & Evert, 2001) suggesting that LogLike delivers the best precision values for low-frequency data. Moreover, they could show that the ranking of the association measure depends on the kind of collocation to be identified: the T-test delivers better results for preposition-noun-verb combinations, whereas the Log-Likelihood measure leads to significantly better results for Adjective-Noun combinations.

Since we worked on corpora of extremely small size, it can be expected that LogLike works best. It turned out that our method performs reasonably well. We evaluated four corpora of different size and different domains: drugs, stock market, running amok and management succession. The data were chosen from German news reports from DPA (1999 and 2000). The smallest corpus contains 6361 tokens, the biggest one contains 84747. The main observations we could conclude are the following:

1) There is a correlation between corpus size and precision. The bigger the corpus the more collocations could be correctly identified.
   Table 1 shows the precision values for the 200 highest-ranked words in a corpus applying Log-Likelihood for computing Noun-Verb collocations.

2) For both combinations Noun-Noun collocations and Noun-Verb collocations LogLike compared to Mutual Information and T-Test delivers the best results. A comparison between LogLike and Mutual Information for Noun-Verb collocations is shown in Table 1.

3) We could not observe a dominance of a certain collocation type depending on a certain domain. In each domain Noun-Verb collocation were most prominent and delivered best results. In the drugs domain we obtained a precision of 56% for Noun-Verb collocations. The precision for Noun-Noun collocations is only 41%.

Size of corpus	LogLike (Noun-Verb)	Mutual Information (Noun-Verb)
6361 tokens	52%	34%
29143 tokens	56%	42%
59134 tokens	63%	36%
84747 tokens	61%	49%

Table 1 Precision values for corpora of different size

The extracted collocations can be used as lexico-syntactic patterns for the identification of terms indicating a certain semantic relation. The noun-noun combination in *7)* helps to find more hyponyms of 'Droge' (drug).

2) Kilogramm <NP_drug>

The extracted noun-noun combination in 3) and 4) indicate semantic relations, which can be used to build up a domain specific ontology. .

***3)* Hyponymy**

a Arzneimittel, Medizinprodukte
b Reparatur, Wartung

Additionally, they are often multi-word terms:

4)

a Frankfurter, Flughafen
b Industrie, Handelskammer
c Volksrepublik, China

Further, the verb-noun combinations can be used to enhance existing subcategorization lexicons and may also constitute candidates for template filling rules.

a sitzen, Untersuchungshaft
b treten, Ruhestand
c Leitung ubernehmen

5. Conclusion

In this paper we presented an unsupervised and domain adaptive approach to automatic extraction of domain relevant terms and multi-word terms and collocations. The KF-IDF based term extraction has proved to be very promising for the extraction of single word terms. We combined this method with the learning of collocations in using the extracted terms as input for the collocation learner. The next step will be using the term classification after the collocation learning in order to determine appropriate collocations. The approach proves to be suitable for handling free-word order languages like German. The extracted terms and collocations can be used on several ways and in several applications. They serve to build up domain specific resources for the IE application, e.g. lexicons, ontologies. In addition, the deployment of collocations can improve text mining tasks such as text categorization or text clustering.

References

[Brill 92] E. Brill. *A Simple Rule-Based Part-of-Speech Tagger.* In Proceedings of the Third Conference on Applied Computational Linguistics (ACL), Trento, Italy, 1992.

[Church & Hanks 1989] *Kenneth Ward Church and Patrick Hanks Word association norms, mutual information and lexicography.* In Proceedings of the 27th Annual Conference of the Association of Computational Linguistics, pages 76-82, 1989

[Daille 1996] Beatrice DAILLE. *Study and Implementation of Combined Techniques of Automatic Extraction of Terminology* In J.L. Klavans and P. Resnik, editors, The Balancing Act: Combining Symbolic and Statistical Approaches to Language, pages 49-66, MIT Press, Cambridge, MA..

[Evert & Krenn 2001] Stefan Evert and Brigitte Krenn. *Methods for the Qualitative Evaluation of Lexical Association Measures* In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics Toulouse, France.

[Finkelstein-Landau & Morin 1999] Michal Finkelstem-Landau and Emmanuel Morin. *Extracting Semantic Relationships between Terms Supervised vs Unsupervised Methods.* In Proceedings of International Workshop on Ontological Engineering on the Global Information Infrastructure, pages 71-80, Dagstuhl Castle, Germany, May 1999.

[Hamp & Feldweg 1997] Birgit Hamp and Helmut Feldweg. *GermaNet - a Lexical-Semantic Net for German.* In: Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications". Madrid, 1997.

[Hearst 1992] Marti A. Hearst. *Automatic Acquisition of Hyponyms from Large Text Corpora.* In: Proceedings of the Fourteenth International Conference on Computational Linguistics, Nantes, France, July 1992.

[Inkpen & Hirst 2001] Diana Zaiu Inkpen and Graeme Hirst. *Building a Lexical Knowledge-Base of Near-Synonym Differences,* In Proceedings of Workshop on WordNet and Other Lexical Resources (NAACL 2001), Pittsburgh, pages 47-52, June 2001.

[Krenn 2000] Brigitte Krenn. *The Usual Suspects Data-Oriented Models for Identification and Representation of Lexical Collocations,* German Research Center for Artificial Intelligence and Saarland University Dissertations in Computational Linguistics and Language Technology, Volume 7. Saarbrücken, Germany, 2000.

[Manning & Schütze 1999] Chris Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing.* MIT Press, Cambridge, MA..

[Piskorski & Neumann 2000] Jakub Piskorski and Günter Neumann. *An Intelligent Text Extraction and Navigation System,* In proceedings of 6th International Conference on Computer-Assisted Information Retrieval (RIAO-2000), Paris, 2000.

[Riloff & Jones 1999] Ellen Riloff and Rosie Jones. *Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping,* Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99), 1999, pp. 474-479.

[Salton 1991] Gerald Salton. *Developments in Automatic Text Retrieval,* Science, Vol 253, pages 974-979, 1991.

[Smadja 1994] G. Smadja. *Retrieving Collocations from Text Xtract,* Computational Linguistics 19(1): 143-177, 1994.

[Yangarber et. al 2000] Roman Yangarber, Ralph Grishman, Pasi Tapanainen and Silja Huttunen. *Automatic Acquisition of Domain Knowledge for Information Extraction,* In Proceedings of COLING 2000: The 18th International Conference on Computational Linguistics, (August 2000) Saarbrücken, Germany.

[Xu et al 2002] Feiyu Xu, Daniela Kurz, Jakub Piskorski and Sven Schmeier. *A domain adaptive approach to automatic acquisition of domain relevant terms and their relations with bootstrapping.* In Proceedings of LREC 2002: Third international conference on language resources and evaluation, Las Palmas, Canary Island, Spain, 2002.