

Hybrid Filtering for Extraction of Term Candidates from German Technical Texts

Munpyo Hong, Sisay Fissaha, Johann Haller

IAI, Martin-Luther-Str.14, D-66111, Saarbrücken, Germany

Tel +49-681-3895126, Fax +49-681-3895140

{munpyo, sisay, hans}@iai.uni-sb.de

<http://www.iai.uni-sb.de>

Keywords: term extraction, MULTILINT, log-likelihood, hybrid filtering, stop word list,
contextual cue

Abstract

Most of the current methodologies for automatic term extraction rely heavily on morpho-syntactic criterion in identification of term candidates, and are mainly developed for English. This in turn makes it difficult to apply these techniques for German texts directly due to the morpho-syntactic differences between the two languages. In this paper, we present an approach to automatic term extraction which takes into account the characteristics of German language, and attempts to combine different filtering techniques (linguistic and statistical). The current work is part of an on-going research at IAI on the development of multilingual document production/management tool MULTILINT which has already been successfully employed by various international enterprises.

1. Introduction

In the documentation of technical texts it goes without saying that the consistency at the lexical as well as stylistic level of the text must be kept in order to make them better understandable and translatable. An inconsistency at a lexical level, e.g. inconsistency among terms in a text does not only make it difficult for human translators to translate but also lowers the efficiency of employing translation tools such as Translation Memory Systems. Therefore it is necessary to check the term candidates with respect to their consistency before compiling them into a term database. The controlling task can be accomplished, on one hand, in a rather strict way by establishing a set of terms that are allowed to be used in a text of a certain domain. In this case the building of the set of terms is guided by a Controlled Language such as AECMA Simplified English. On the other hand, the consistency of the terms can be automatically checked by controlling tools like MULTILINT. In either case the terminological resources must be available in machine-readable forms, which often does not seem to be the case in reality. To our knowledge, the only commercially available systems for term extraction and controlling are the XEROX Terminology Suite, and MULTILINT developed at IAI.

This paper reports a part of our ongoing research at IAI in Saarbrücken to further develop a multilingual document production/management tool MULTILINT which has already been successfully employed by various international enterprises. The focus of the paper will be on automatic extraction of term candidates from German technical texts in the context of maintaining the MULTILINT system. In the MULTILINT system the term candidates are

extracted from a text to check their consistency and spelling etc., before they are compiled into the term database. We will not go into the details about the MULTILINT system in this paper. For more about MULTILINT, the readers are referred to Haller (1996).

Our methodology for automatic term extraction relies both on linguistic and on statistical knowledge. Our claim in this paper is that most of the current methodologies for automatic term extraction are developed for English so that they are not suitable for direct adaptation to German texts due to the morpho-syntactic differences between the two languages. We will present an approach where the morpho-syntactic characteristics of German are well reflected from the perspective of term extraction.

2. Some linguistic characteristics of German Terms from the perspective of automatic extraction

The notion of a term that constitutes the subject of our investigation can be defined in various ways. However, a term is defined in this work as words or phrases that denote specific concepts in a given subject domain and are categorised as a term by terminologists or experts for the domain following the view of (Jacquemin & Bourigault 2000). The term candidates below from automobile industry domain exemplify our view of terms:

- (1) *Doppelscheinwerfer, Wankstabilisierung, Aufbauregelung, Abrollkomfort, Passiver Gasdruck-Stossdämpfer, induktive Antenne, zwei-stufige Airbagaktivierung, Armauflage der Mittelkonsole, Automatikgetriebe mit fünf Fahrstufen, ...*

One of the most striking differences between German and English in technical documentation is that the use of single-word terms in German texts is much more frequent than in English texts in comparable size. This is in most cases due to the fact that compounding is much more productive in German than in English. To show the tendency in German texts that a compound is preferred to multi-word terms, some example term candidates extracted from parallel German-English corpus are presented below:

- (2) *Schwingungskomfort* : vibration comfort
Wankstabilisierung : anti-roll stabilization
Alu-Hohlspeichenfelgen: Hollow spoked aluminum wheel
Gas-Generator: gas generator

In most of the papers dealing with automatic term extraction the emphasis of the research is put on the extraction of multi-word terms. This can be explained by the fact that in English technical document the majority of the terms are multi-word terms. In three sorts of German technical document from different subject domains (automobile, electronics and machinery) we found out however that the proportion of single-word terms varies from about 57% to 94% depending on the subject domain:

Text	Total Number of Terms ¹	Total Number of single-word terms	Percentage of single-word terms
Text 1 (automobile)	295	276	93.56%
Text 2 (electronics)	123	70	56.9%
Text 3 (machinery)	162	137	84.6%

¹ We do not mean by this the total occurrence of the terms but only the number of term tokens.

In any case the percentage of single-word terms is much higher than that of English. Most of the German single-word terms are compounds.² From this we could conclude that a **compound word** is a very likely term candidate in German technical documentation. A sophisticated morphological analyzer is needed for the detection of compounding structure.

Many of the other single-word terms are such words as are directly adapted from English. In German technical texts we encountered some single English terms that denote specific functions or parts of automobiles or machines. In most of the cases they were invariably used in German texts. Our investigation showed that **English words** in German texts are also very likely term candidates, as is shown in (3):

(3) DISTRONIC, Economy-Program, Power-Program ...

In comparison with English, the frequency of multi-word terms is much lower in German technical documentation. However in one case we found out that the percentage of multi-word terms reach about 45 % of the whole terms. In the following the most frequent syntactic patterns of the multi-word terms are listed with the examples:

(4) ADJ + N : *induktive Antenne, passiver Gasdruck-Stoßdämpfer, dreischalige A-Säule*

N + NP_{GEN} : *Armauflage der Mittelkonsole, Wippbewegung der Wippspitze*

N + PP : *Schiebedach mit Memoryfunktion, Automatikgetriebe mit fünf Fahrstufen*

Among the three types, the "ADJ + N" structure was the most frequent one, though the exact percentage still remains to be calculated. Interestingly, in the case of "N + PP" structure we found out that the most of the terms corresponding to this pattern is built with the preposition "*mit (with)*". This may be explained by the fact that the PPs with "*mit*" usually denote an important functionality or subparts of a modified noun that can differentiate the concept denoted by the term from other concepts, as is the case in "*Schiebedach mit Memoryfunktion* (sliding sunroof with memory function)" and "*Automatikgetriebe mit fünf Fahrstufen* (automatic five-speed transmission)". The other expressions matching the "N + PP" pattern are either with the preposition '*aus*' or '*bei*'.

3. Linguistic Filtering

As is often the case in other research into automatic term extraction, our starting point for the extraction is to find out the syntactic contexts of German term candidates. Arppe (1995) claims that NPs constitute about 80-99 % of whole terms in an English text with the varying percentage depending on the text types, so that NPs can be a good starting point for a term extraction. Heid (1999) also investigates the following syntactic patterns for term extraction from German technical texts:

(5) N + N (Genitiv)

N + Prp (+ Det) + N

ADJ + N

Although we also found some adjectives and verbs which can be regarded as terms in a certain domain, we are mainly concerned with the NPs in this paper. For this we concentrated on the following regular expression:

² In this paper we mean by the term '*compound*' only such words as are composed of more than one morphemes and do not contain any blanks between the morphemes.

(6) (ADJ*) NP+ (PP*)

The regular expression corresponds for example to the following term candidates:

- (7) *Passiver Gasdruck-Stoßdämpfer* (passive gas-pressurized shock absorber)
Verstellmechanismus (adjusting mechanism)
Active Body Control (Active Body Control)
Schiebedach mit Memory-Funktion (sliding sunroof with memory function)

To extract the expressions complying with the above syntactic patterns, we employed a German syntactic parser equipped with a powerful morphological analyzer MPRO developed at our institute.³ The morphological analyzer MPRO does not only lemmatize the input words but also provides rich information such as derivational, inflectional and compounding structures of words so that this information can be usefully employed for term extraction as well as for the treatment of term variation.⁴ For detail, let us take a look at the result of the morphological analysis of "*Zylinderabschaltung* (cylinder cutout)" represented in a feature structure⁵:

(8)

```
{ori=Zylinderabschaltung,c=noun,lu=zylinderabschaltung,s=massnahme,t=zylinder#abschaltung,cs=n#n,ts=zylinder#abschaltung,ds=zylinder#ab_ $schalten-ung,ls=zylinder#ab_ $schalten,ss=form#massnahme,lng=germ,lngs=germ#germ,w=2,wnrr=5,case=dat,nb=sg,g=f}
```

The MPRO morphological analysis delivers not merely the lemma (*lu*) of the input words but also the compounding structure (*cs*) and compounding parts (*t*), as well as the part of speech information (*c*) etc. We can also see in the feature structure that the language to which the word belongs (*lng*) is specified. It can play an important role in identifying a single-word term in German technical documentation.

The MPRO parser works on the basis of the morphological analysis of MPRO.⁶ There are several reasons for employing a morphological analyzer and a syntactic parser for term extraction. Firstly, it enables the NPs to be identified with a high accuracy. Secondly, it helps to identify the strings which are treated as terms in a particular domain and are already stored in the dictionary. Thirdly, it allows for the uniform treatment of such typographical variations as inconsistent use of hyphens, e.g. "*zwölfzylindermotor*" and "*zwölfzylinder-motor*" which will be analysed as "*zwölfzylinder#motor*". For illustration the parse result of an example NP "*der elektronisch gesteuerten Luftfederung*" is presented below:

(9)

```
{c=np,r=217b,tr=ehead_up1,snr=1735,ehead={nb=sg,g=f,case=dat;gen,flex=weak},case=dat;gen,nb=sg,g=f}
  {c=np,r=203,ehead={nb=sg,g=f,case=dat;gen,flex=weak},snr=1735}
```

```
{ori=der,wnra=11935,snr=1735,gra=small,pctr=no,pctl=no,last=no,oldsgml=no,c=w,sc
```

³ cf. Maas (1996).

⁴ cf. Schmidt-Wigger (1999) for the treatment of term variation based on MPRO.

⁵ For the better illustration, some irrelevant AV pairs are removed from the original feature structure.

⁶ We will call this parser simply MPRO parser. In this paper the term 'MPRO' is used to refer to the morphological analyzer and 'MPRO parser' to refer to the syntactic parser.

```
=art,fu=det,np,s_flex=weak,spec=def,lu=d_art,ds=d_art,ls=d_art,wnrr=18,case=dat;gen,nb=sg,g=f}  
{c=adj,r=124,endung=en,snr=1735,case=dat;gen,nb=sg,g=f}
```

```
{ori=elektronisch,wnra=11936,snr=1735,gra=small,pctr=no,pctl=no,last=no,oldsgml=no,o,c=adv,lu=elektronisch,deg=base,t=elektronisch,cs=a,ts=elektronisch,ds=elektronik~isch,ls=elektronik,ss=a,w=1,wnrr=19}
```

```
{ori=gesteuerten,wnra=11937,snr=1735,gra=small,pctr=no,pctl=no,last=no,oldsgml=no,c=adj,lu=gesteuert,endung=en,deg=base,ptc=2,t=gesteuert,cs=v,ts=gesteuert,ds=steuern,ls=steuern,ss=v,w=1,wnrr=20}
```

```
{ori=Luftfederung,wnra=11938,snr=1735,gra=cap,pctr=no,pctl=no,last=no,oldsgml=no,c=noun,lu=luftfederung,s=ation,t=luft#federung,cs=n#n,ts=luft#federung,ds=luft#federn~ung,ls=luft#federn,ss=mat#ation,lng=germ,lngs=germ#germ,rss=ok#ok,w=2,wnrr=21,case=dat;gen,nb=sg,g=f}
```

Based on the morphological analysis and syntactic parsing, the NPs corresponding to the above regular expressions are extracted. However a deficiency of the approach based solely on the syntactic patterns is that any NPs that match the patterns will be selected as term candidates, as is shown in the following examples:

- (10) a. *die Spitze der deutschen Automobilproduktion* - ‘the pinnacle of German automobile production’
b. *Technologieführer in allen Disziplinen* - ‘technological leader in all disciplines’,
c. *flache Motorhaube* - ‘flat hood’

In (10.a) the NP structure with a nominative NP and a genitive NP is wrongly extracted, though in this case it is concerned with a trivial compositional building of an NP structure rather than a multi-word term. In (10.b) the NP ‘*Technologieführer in allen Disziplinen*’ cannot be treated as a term, because it is rather unlikely that this expression will be used frequently in the same text which would imply the termhood of the expression. Also in (10.c) the relationship between the adjective ‘*flach*’ and the noun ‘*Motorhaube*’ cannot be regarded as that of a term.

Another problem of the approach is that it is difficult to extract substrings which themselves also show the termhood as in the following example:

- (11) **Original string:** *Zwölfzylindermotor mit automatischer Zylinderabschaltung* (12-cylinder engine with automatic cylinder cutout)

Substring: *automatische Zylinderabschaltung* (automatic cylinder cut-out)

From this it is clear that a post filtering mechanism is needed after linguistic analysis. As a post filtering mechanism we adopt an algorithm for calculating the associativity between words, which is often used to identify collocations and technical terms.

4. Statistical Filtering

In the field of term extraction, several statistical techniques have been proposed. Among the most commonly applied statistical measures in the identification of collocations and technical terms are Simple matching coefficient (SMC), Kulczynsky coefficient (KUC),

Ochiai coefficient (OCH), Fager and McGowan coefficient (FAG), Yule coefficient (YUL), Mutual Information and Log-Likelihood coefficient.⁷ These coefficients measure the strength of association of the words in the technical terms or collocations. The frequency data, which is required for computing all of the above statistics, comes from the following contingency table.

(12)

	a	b
1	A & B	(not A) B
2	A (not B)	(not A) (not B)

Another statistic, which makes use of a relatively different type of frequency data, is C-value statistic (for details on C-value statistic see Franzi & Ananiadou (1997)).

For the current work, Log-likelihood coefficient has been employed as it is shown to perform well for sparse data and to possess other important characteristics such as ease of interpretation (cf. Dunning (1993); Daille et al. (1994) and Manning et al. (2000)). This statistic has also been employed in other similar problems like discovery of subcategorization frames (cf. Sarkar et al. (2000)). Assuming a binomial distribution, the log-likelihood statistic is given by

$$-2\log \lambda = 2 [\log L(p_1, k_1, n_1) + \log L(p_2, k_2, n_2) - \log L(p, k_1, n_1) - \log L(p, k_2, n_2)]$$

where

$$\log L(x, y, z) = z \log x + (z-y) \log(1-x)$$

The parameters required for computing the Likelihood Statistic are

$$k_1 = a_1, k_2 = a_2, n_1 = a_1 + b_1, n_2 = a_2 + b_2, p_1 = k_1/n_1, p_2 = k_2/n_2, p = (k_1 + k_2)/(n_1 + n_2),$$

where a_1, a_2, b_1 and b_2 are the values taken from the contingency table shown in (12).

The larger the value of $-2\log\lambda$ is, the stronger is the association between the two pair of substrings. This information may serve in turn as an important clue in identifying term candidates.

Before proceeding to the statistical analysis, however, one more step is required which is generating the bigrams and the corresponding frequency counts. In generating the bigrams, the program starts first by extracting strings using the pattern shown in the previous section. However, an initial run of the program showed that some of the strings extracted are long ones containing substrings which are likely to function as term candidates. In order to avoid potential term candidates being included in other longer strings possibly containing noise, the program also generates substrings automatically using the following patterns:

(13) (ADJ)* N*
N*

While extracting strings, the program also splits the strings into two parts at word level and generates the necessary bigrams. We illustrate the above procedure with the following two example term candidates, "*Mercedes Benz CL mit Active Body Control*" and "*Zwölfzylindermotor mit automatischer Zylinderabschaltung*". Each of these term candidates consists of two substrings, "*mercedes-benz CL*" and "*Active body control*", and

⁷ cf. Oakes (1998) for the mentioned statistical measures.

"zwölfzylindermotor" and "automatischer zylinderabschaltung" which can themselves be term candidates:

- (14) ((Mercedes-Benz CL) mit (Active body control))
 ((zwölfzylindermotor) mit (automatischer zylinderabschaltung))

Hence, from the above example strings, the program generates the following substrings:

- (15) mercedes-benz CL
 Active body control
 Active body
 body control
 automatischer zylinderabschaltung
 zwölfzylindermotor
 zylinderabschaltung

The above operation seems to overgenerate unnecessary data. Although "Active Body Control" is one possible term candidate, "Active body" is less likely to function as a term in the current context. Therefore, only those substrings which occur (at least once) independently of the main string are retained. In the above example, only the first two and the last three substrings will be considered for analysis, since these are the only substrings which occur at least once independently of the main string. Of course, the original strings will also be included in the analysis.

As mentioned in the previous section, most of the single words which are likely to function as terms are compound words which are composed of two or more morphemes. Hence, compound words will also be split at morpheme boundaries and counted as strings. This in turn, enables the identification of compound words which occur frequently and are likely to function as terms, and enables the uniform treatment of multiword and single-word term candidates. Then, the final data needed for statistical analysis looks like the following:

- (16)

A	B
mercedes-benz CL	mit Active body control
mercedes-benz	CL
mercedes	benz
Active	body control
Active body	control
automatischer	zylinderabschaltung
zylinder	abschaltung
zwölfzylinder	motor

The program will then generate the frequency counts for each bigram and compute the likelihood ratio.

As mentioned previously, for the extraction of the single-word terms we employed the same technique as for the extraction of multi-word terms. This was based on our assumption that most of the single-word terms in German are compound words. However there are certainly other single-word terms that are not compound words. We employed for the extraction of such words certain syntactic contexts in which a term is typically defined. Such contexts are in German e. g. '... als *TERM* bezeichnet sein (be called *TERM*)' or '... mit

TERM ausgestattet sein (be equipped with TERM)'. This kind of a contextual cue is applied at the last stage after the linguistic and statistic filtering have been applied.

A preliminary test has been carried out with 12,000 words corpus in the field of automobile. The table below shows the top, middle and last 5 term candidates. As can be seen from this table most of the top most ranking term candidates are compound words:

(17)

1.	mercedes-benz	462.14	
2.	s-klasse	288.28	
3.	Active body control	205.35	
4.	neuer CL	203.63	
5.	CL 600	179.60	
...			
881.	drehstab-stabilisator		15.58
882.	elektronisch geregelten luftfederung AIRmatic		15.58
883.	elektronisch gesteuerten luftfederung AIRmatic		15.58
884.	fahrwerkstechnischen zielkonflikt		15.58
885.	federkern-prinzip		15.58
...			
2248.	neue qualität	1.05	
2249.	neue coupe	0.66	
2250.	fahrweise	0.53	
2251.	CL modell	0.17	
2252.	neue modell	0.15	

Furthermore, we found out that among the term candidates matching the pattern 'ADJ + N', many of them are not relevant to the termhood in contrast to what the likelihood ratio actually suggests. These phrases were given relatively high scores because they were repeatedly used in the text. Some example of such phrases are presented below:

(18) *neuer Integralsitz, neues V8-Triebwerk, neuer zwölfzylinder-Motor, innovative Karosserietechnik, bisherige Doppelverglasung, herkömmliche Schweißverbindung ...*

One possible explanation to such a problem is that the sample text used in the identification of term candidates is relatively small and hence may be not representative enough to allow accurate recognition of term candidates. Depending on the text sorts (manuals, booklets, etc.) one can expect that such phrases can occur frequently. For example, in our sample text which is written for the presentation of a new model, it is quite easy to see why such phrases as "*neuer zwölfzylinder Motor* (new 12-valve engine)" or "*bisherige Doppelverglasung* (previous dual-pane)" are used repeatedly. This in turn makes it necessary to use a stopword list. For this reason, we collected manually such adjectives that cannot really have anything to do with termhood and enriched our system with the stop word list.

(19) Stop Word List (Adjektive)

ander, außergewöhnlich, beachtlich, beide, besonder, bisherig, deutlich, echt, eigen, einfach, einzig, erforderlich, erst, ganz, gemeinsam, genau, gering, gesamt, gleich, herkömmlich, hochwertig, jeweilig, komplett, konkret, konventionell, kostenlos, lebenslang, luxuriös, maßgeblich, möglich, neu, notwendig, perfekt, richtig, schließlich, sinnvoll, sogenannte, solch, speziell, spezifisch, tatsächlich, teilweise, traditionsreich, typisch, üblich, übrig, unterschiedlich, ursprünglich, viel, vorbildlich, weiter, wenig, wichtig, zahlreich, ...

The resulting system has been tested again with a different sample text consisting of 75,000 words using the above stop word list as an additional filter and the following result was obtained:

(20)

Total number of words	74,676
Total number of extracted term candidates	3,176
Total number of correctly extracted term candidates	2,372
Precision	75%

Another important feature of the system is that it enables the user to modify the stop word list interactively. Using this feature, we could improve the precision rate to 80%.

5. Conclusion

In this paper we presented our approach for the extraction of term candidates from German technical texts. We showed that a compound word is a very likely candidate for the extraction because of the productive compounding phenomenon in German. The termhood of a compound word is calculated based on the associativity of the base morphemes. If they occur together frequently, they are regarded as a term candidate. As for the multi-word terms, the associativity is calculated between the words composing the term candidates. In case of the 'ADJ + N' construction, only such phrases where the adjective and the noun occurs frequently together are extracted. We also showed that such a phrase as "*neuer Zwölfzylindermotor* (new 12-cylinder engine)" is assigned relatively high scores which is due to the characteristics of the text type (advertising text). For this we presented a stop word list where all the irrelevant adjectives are listed. The same method has been applied for the 'NP + PP' and 'NP + NP' structures. In our experiment with 'NP + PP' structure we also encountered a similar problem to that of 'ADJ + NP'. Because of the collocational nature of such phrases as "*im Vergleich zu*" and "*in Gegensatz zu*" the N and P pairs like "*Vergleich zu*" and "*Gegensatz zu*" are wrongly extracted. We are currently working on an algorithm to exclude such errors. The next step of our research will be about the interface between our term extractor and a term database.

References

- ARPPE, A. (1995). Term Extraction from Unrestricted Text, in *NODALIDA-95*, URL:<http://www.lingsoft.fi/doc/nptool/term-extraction.html>
- DAILLE, B.; GAUSSIÉ, E. & LANGE, J. M. (1994). Towards Automatic Extraction of Monolingual and Bilingual Terminology, in *Proceedings of COLING 94*, pp.515-521.
- DUNNING, T. (1993). Accurate methods for the statistics of surprise and coincidence, in *Computational Linguistics*, Vol. 19(1), pp. 61-74.
- FRANTZI, K. & ANANIADOU, S. (1997). Automatic Term Recognition using Contextual Cues, in *Proceedings of MulSaic 97, IJCAI*

- HALLER, J. (1996). MULTILINT - Multilingual Intelligence for Technical Documentation, in *ASLIB-Proceedings*, pp.1-13
- HEID, U. (1999A). Extracting terminologically relevant collocations from German technical texts, in *5th International Congress on Terminology and Knowledge Engineering (TKE'99)*
- HEID, U. (1999B). A linguistic bootstrapping approach to the extraction of term candidates from German text, in *Terminology*
- JACQUEMIN, C. & BOURIGAULT, D. (2000). Chapter 19 Term Extraction and Automatic Indexing, in *Handbook of Computational Linguistics (R. Mitkov (ed.))*, Oxford University Press, Oxford
- MAAS, D. (1996). MPRO - Ein System zur Analyse und Synthese deutscher Wörter, in *Linguistische Verifikation (R. Hausser (ed))*, *Sprache und Information*, Tübingen: Max Niemeyer Verlag
- MANNING, C. & SCHÜTZE, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.
- OAKES, P. M. (1998). *Statistics for Corpus Linguistics*. University Press, Cambridge
- SARKAR, A. & ZEMAN, D. (2000). Automatic Extraction of Subcategorization Frames for Czech, in *Proceedings of COLING 2000*, 691-697.
- SCHMIDT-WIGGER, A. (1999). Term Checking through Term Variation, in *5th International Congress on Terminology and Knowledge Engineering (TKE'99)*