Evaluating noise reduction strategies for terminology extraction

Johannes Schäfer[1], **Ina Rösiger**[1], **Ulrich Heid**[2], **Michael Dorna**[3]
[1]Institute for Natural Language Processing, University of Stuttgart, Germany
[2]University of Hildesheim, Germany, [3]Robert Bosch GmbH, Germany
{schaefjs|roesigia}@ims.uni-stuttgart.de
heid@uni-hildesheim.de, michael.dorna@de.bosch.com

Abstract

We present work on the task of reducing noise in nominal terminology extraction. Based on a comparative evaluation of statistical measures aimed at capturing domain specificity, we propose strategies to increase the typically quite low accuracy of classical hybrid nominal multi-word term extraction. Our experiments on a set of German do-it-yourself instruction texts show that using linguistic filters that determine the right span of the MWE before applying a suitable combination of statistical measures improves results.

1 Introduction

The automatic extraction of terminology from domain-specific text is a task that has gained interest in the research community over the last twenty years. It is an important prerequisite for applications such as ontology creation or knowledge extraction from texts.

The work presented here is part of a project that deals with knowledge extraction and ontology creation from German texts from the do-it-yourself domain. As a first step, we aim at high quality terminology extraction of nominal candidates, as these describe the objects of the domain, followed by the extraction of verbal items and verb+complement patterns before we bring them together to build up partial ontologies of the domain. This paper describes strategies to reduce noise in nominal term extraction.

We consider single-word terms and multi-word terms, but focus on the latter because the extraction of multi-word terms (MWTs) is more difficult. As they are of variable length it is in many cases nontrivial to ensure the correct span of the term. We aim at extracting noun phrases (NPs) of different levels of complexity, such as adjective and noun combinations as well as NPs containing a genitive or prepositional modifier.

Nominal terms may contain embedded prepositional phrases (PPs), such as in example (1).

(1) *Bohrer mit Diamantspitze*[1]
(drill with diamond bit)

However, we do not want to extract PPs that are not syntactically attached to a term, e.g. because they are verb-dependent, such as in example (2).

(2) *die *Oberfläche mit Leinölfirnis bedecken*
*(cover the *surface with linseed oil varnish)*

Thus, one of the noise reduction steps is to ensure that the extracted nominal candidates are syntactically valid, and do not cover too long spans.

There are also cases where the term extraction may return too short candidates. Sometimes, the extracted terms only occur as part of bigger terms, and are not valid on their own, such as in example (3).

(3) *elektromagnetisch *angetriebene Spritzpistole*
*(electromagnetically *operated spray gun)*

There are both statistical and hybrid approaches to term extraction (Cabre and Vivaldi Palatresi, 2013). Association measures (cf. e.g. Evert (2005)) are designed to extract collocations ("unithood", cf. Kageura and Umino (1996)) and have been used for terminology extraction, e.g. by Couturier et al. (2006). However, Roche et al.

[1]Extracted term candidates are underlined. The * here denotes wrongly extracted MWT candidates.

(2004) investigated the use of association measures for this task and came to the conclusion that these standard measures are outperformed by more sophisticated approaches. They do not focus on domain-specificity ("termhood") and thus do not perform better at terminology extraction than mere frequency based approaches (Pazienza et al. (2005), confirmed by our own experiments, where the maximal F_1-score obtained with association measures is 0.45).

Termhood is addressed through statistical measures that only use the candidate's frequency in a domain-specific corpus (e.g. Frantzi et al. (2000)), as well as by measures based on a comparison of a candidate's frequency in a domain and in general-language corpora (cf. Ahmad et al. (1992) and section 2.3). However, among others due to data sparseness in small size specialized corpora, both approaches perform much better on single-word terms (SWTs) than on MWTs.

Most hybrid systems (linguistic pattern-based search for candidates plus statistical ranking) often do not address variable length and syntactic validity satisfactorily (with a few noteworthy exceptions, cf. for example Chen et al. (2008)): part-of-speech (POS) sequence patterns are typically flat and cannot identify phrase boundaries and grammatical functions (cf. example (2)). As the POS patterns do not model phrase structure, they may cut off essential parts from a multi-word, returning unattested candidates (cf. example (3)).

We address the above issues by means of a three-step approach which modifies and extends the classical hybrid scenario: (i) nominal candidates are selected via part-of-speech patterns; (ii) they are filtered wrt syntactic validity and embedding and finally (iii) ranked according to statistical measures that involve a comparison between specialized and general-language corpus. The system with which we experiment extracts lemma combinations, morphosyntactic properties and text-specific metadata. Our method is evaluated on a 2.7 million word corpus of German do-it-yourself (DIY) instruction texts against a gold standard.

The main contributions of this paper are a study on the suitability of standard statistical measures for the extraction of nominal single- and multi-word terms, as a basis for further adaptations and methods to improve the noise-silence ratio, based on experiments with linguistic filters (we use parsing information to ensure the MWE is a valid nominal phrase (NP)), and experiments on the combination of statistical measures. We believe that the methods we propose are generalizable to other domains of specialization and to other languages. More experiments will however be needed to confirm this.

2 Improving term extraction quality

In the present work, we only deal with nominal candidates. To maximize recall on (comparatively) small specialized corpora, we use POS patterns that account for basic terms[2] (N, Adj N, N P N, N D $N_{genitive}$) and for their potential variants (step (i) in the summary above). The set of patterns is described by the regular expressions given below.

– (Adv? Adj? Adj)? N
– (N D)? (Adv? Adj)? N P D? (Adv? Adj)? N
– (Adv? Adj)? N D (Adv? Adj)? $N_{genitive}$

These patterns are flat and thus do not adequately represent syntactic structure. In particular, they cannot distinguish between cases (a) where NP and PP are sister nodes vs. (b) where the PP is embedded in the NP (cf. examples (1) and (2) above). Thus, step (ii) is added to remove noise: we exclude items from our candidate set which are syntactically invalid (too long ones) by checking phrase boundaries and we use the C-value score (Frantzi et al., 2000) to remove too short items, i.e. those occurring only embedded in other candidates. In step (iii) we combine statistical measures to rank the selected candidates by domain specificity.

2.1 Ensuring syntactic validity

Candidates covering too long spans typically occur when part of the extracted MWT is actually attached to the verbal phrase, e.g. in example (4) and example (5).

(4) *die *Schablone mit Farbe besprühen*
*(spray the *template with paint)*

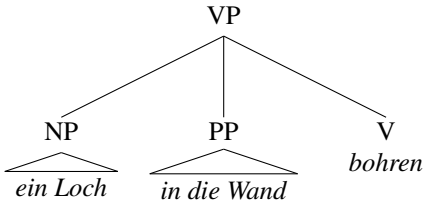
(5) *ein *Loch in die Wand bohren*
*(drill a *hole into the wall)*

[2]POS tags: N-noun, Adj-adjective, P-preposition, Adv-adverb, D-determiner.

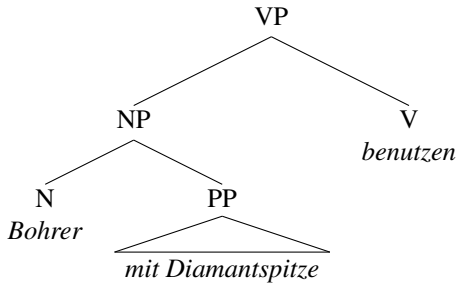
We filter these by using the dependency parser *mate* (Bohnet, 2010) to find start and end points of NPs.[3] This parser was chosen because, in the long run, we aim at relation extraction based on syntactic functions. Moreover, *mate* has been shown to produce the highest accuracy in a recent evaluation of the currently available dependency parsers (Choi et al., 2015). We are aware that *mate* has not been optimized to solve the PP attachment problem and that there is no evaluation on specialized text available yet (cf., however, Zollmann et al. (2016) on a partial evaluation).

The boundary violation filter works as follows: If an instance of a MWT candidate comprises two sister phrases, i.e. if the POS sequence identified goes beyond the end point of an NP, it is not counted as a valid occurrence of the respective lemma sequence. As an example for a violation, consider the following MWT candidate where '*ein Loch*' and '*in die Wand*' are sister phrases ('*ein Loch in die Wand bohren*' (drill a hole into the wall), (5)), whereas in the example (6) ('*Bohrer mit Diamandspitze benutzen*' (use a drill with diamond bit)) there is no violation.

(5)



(6)



The candidate sequence is not removed from the list of possible candidate terms, as other occurrences might not have been analyzed as violating syntactic boundaries. The filter is thus a "soft" one as it only affects the frequency of the lexeme combination candidate. We also experiment using a "hard" filter, where the lexeme combination candidate is removed altogether.

[3] In the current experiments only for subject and object phrases.

2.2 Filtering out invalid embedded phrases

An example to show the necessity for an accurate treatment of nested terms was found in our extraction result: '*zugängliche Stelle*' (accessible place) and '*schlecht zugängliche Stelle*' (poorly accessible place). It should be obvious that from occurrences of the latter term we do not want to extract the former.

Thus, with nested MWEs, not all fragments of a longer expression might be suitable candidate terms. C-value (Frantzi and Ananiadou, 1996) identifies embedded sequences as valid units under the following condiditons: (a) the embedded sequence also occurs on its own; (b) the embedded sequence occurs in lexically diverse longer sequences. The C-value for a candidate term a is defined as in formula 1.

$$\mathrm{C}(a) = \begin{cases} \log_2 |a| * f(a) & \text{if } a \text{ not nested} \\ \log_2 |a| * f(a) - \frac{\sum_{b \in T_a} f(b)}{P(T_a)} & \text{otherwise} \end{cases} \quad (1)$$

$|a|$ = term length of a (number of words)
$f(.)$ = frequency in the domain corpus
T_a = set of longer candidate terms that contain a
$P(T_a)$ = number of these longer candidate terms

Furthermore, C-value reflects the idea that longer sequences have a tendency to be more (domain-) specific than shorter ones.

Based on this, we consider German noun compounds as pseudo-MWEs and compute the term length $|a|$ from formula 1 by using the result of compound splitting as produced by the compound splitting tool COMPOST (Cap, 2014) (cf. formula 2).

$$\text{termlength}(a) = \sum_{w \in a} (1 + \log(\text{cslen}(w))) \quad (2)$$

w = a word
$\text{cslen}(w)$ = number of compound elements in w

Frantzi and Ananiadou (1996) propose C-value as a termhood measure using only the frequencies in the domain corpus; thus, most general-language noise cannot be filtered out. We therefore suggest to use C-value as a corrected frequency and to combine it with further statistical measures.

2.3 Ranking by domain-specificity

In section 4, we will compare the following statistical measures designed to rank candidate terms by domain-specificity. A detailed description of these measures is given in Schäfer (2015). As these measures place general-language candidates

at the bottom of the list, a selection of top candidates shows a reduced amount of noise. The measures are defined as follows using the domain frequencies f, the general-language frequencies F as well as the sizes of the corpora: s for the domain corpus and S for the general-language corpus.

- Weirdness ratio for domain specificity (DS) (Ahmad et al., 1999): Identifies domain-specific terms by the ratio of the relative frequencies in the domain and in general language as in formula 3.

$$\text{Weirdness} = \frac{f/s}{F/S} \tag{3}$$

- Corpora-comparing log-likelihood (LL) (Rayson and Garside, 2000): Identifies units with significant frequency differences between the two corpora by formula 4[4]. Note that this version of LL differs from the standard log-likelihood collocation measure.

$$LL = 2\left(f * \log\left(\frac{f}{E_f}\right) + F * \log\left(\frac{F}{E_F}\right)\right) \tag{4}$$

With $E_f = \frac{s*(f+F)}{s+S}$ and $E_F = \frac{S*(F+f)}{S+s}$.

- Contrastive Selection via Heads (CSvH) (Basili et al., 2001): Computes the domain-specificity of a multi-word candidate (ct) using a contrastive filter based on the general-language frequency of its head ($h(ct)$) by formula 5.

$$cw_{ct} = \log(f_{h(ct)}) * \log(\frac{S}{F_{h(ct)}}) * f_{ct} \tag{5}$$

- Term Frequency Inverse Term Frequency (TFITF) (Bonin et al., 2010): Combines the term frequency in the domain corpus with the inverse term frequency in the general-language corpus as in formula 6.

$$w_t = \log(f(t)) * \log\frac{S}{F(t)} \tag{6}$$

[4]As the LL formula is obviously symmetric in the two corpora, we multiplied the result for candidates by -1 if their relative frequency in the domain is smaller than the one in general language, in order to place candidates with a significantly high general-language frequency at the bottom of the list.

- Contrastive Selection of multi-word terms (CSmw) (Bonin et al., 2010): Applies a contrastive filter using the general-language frequency including an $\arctan$ scaling to reduce variation in low-frequency candidates as in formula 7.

$$\text{CSmw}(t) = \arctan(\log(f(t)) * \frac{f(t)}{F(t)/S}) \tag{7}$$

3 Evaluation setup

Tool. We used the TTC[5] (Terminology extraction, translation tools and comparable corpora (2010-2012)) term extraction research prototype (Gojun et al., 2012), a standard hybrid tool that combines linguistic preprocessing with statistical measures, which has recently proven to outperform SDL MultiTerm[6], a purely statistical commercial state-of-the-art tool (George, 2014).

The pipeline involves the following components:

Figure 1: Term extraction pipeline

- Preprocessing:
  - Tokenization: sentence and word form delimitation and markup;
  - Word class tagging and preliminary lemmatization: annotation by means of the RFTagger (Schmid and Laws, 2008), including an annotation as "unknown" of word forms absent from the tagger lexicon;
  - Lemmatization: specific treatment of the word forms absent from the tagger lexicon, with a view to guessing their lemma and part of speech, by use of word form similarity, inflection-based rules and compound splitting.
- Pattern-based term candidate extraction: use of simple as well as extended POS-based patterns to identify term candidates;for the patterns used see section 2.

[5]TTC-project: `http://www.ttc-project.eu/`

[6]`http://www.sdl.com/de/cxc/language/terminology-management/multiterm/`

# tokens	text
62,131	do-it-yourself handbook
6,868	encyclopedia entries
5,150	list of FAQs with answers
15,104	tips and tricks for do-it-yourselfers
35,302	marketing texts
2,160,008	user generated project descriptions
444,381	user generated wiki content
2,728,944	total DIY corpus

Table 1: Number of tokens in the domain corpus

- Ranking:
  sorting of the candidate lists produced by the preceding step, according to different measures (cf. section 2.3).

Domain corpus. We use a corpus of expert and user-generated German texts from the DIY-domain, consisting a.o. of manuals, practical tips, marketing texts and project descriptions (cf. table 1). This corpus is highly heterogeneous since the domain texts were acquired by various methods resulting in fundamentally different types of texts. The texts also differ with regard to the level of expertise of the author and the intended reader. Some texts are written by a domain expert as instructions for users and some are user-generated context.

As the texts differ wrt authorship and text style, several statistical measures are implemented in order to identify different properties of terms providing multiple lists of term candidates. A domain expert can then select the most relevant lists for the construction of a terminological representation of the domain language. For the experiments presented in this work we treated the corpus as a single unit. A source identifier was included as meta data annotation. In future work on a larger version of the corpus, subsets by text type and author/intended reader may be analyzed separately.

General-language corpus. We use the SdeWaC corpus (Faaß and Eckart, 2013) as a general-language corpus. It consists of 880 million tokens. This corpus was chosen since its sentences are a broad collection of German web texts supposed to provide a statistically representative distribution of words in general language. [7]

[7] An alternative source would be Wikipedia, as it covers a broad variety of specialized topics.

POS pattern	number	example
N	4,238	*Kreissäge (circular saw)*
Adj N	604	*thermische Zersetzung (thermal decomposition)*
N P N	148	*Bohren von Dübellöchern (drilling of dowel holes)*
N D N_{gen}	107	*Viskosität der Farbe (viscosity of the paint)*

Table 2: Terms in the gold standard

Gold standard. A gold standard (GS) has been developed for the basic POS patterns (cf. section 2) which we take to capture the core terminology of the domain. Lemma sequences with a minimum frequency of four were extracted from the domain corpus matching these patterns. The gold standard contains those terms which were marked as terms by at least two out of three independent annotators carefully following defined guidelines (George, 2014). This process of creating a gold standard is based on the concept of monolingual reference lists (cf. Loginova et al. (2012)). Our gold standard contains 4,238 SWTs (nouns including compounds) and 826 MWTs (cf. table 2). This distribution is due to the fact that we derived the gold standard from the available text data and not e.g. from a test suite. Moreover, the frequency cut-off of four removed many MWT from being considered for the gold standard. As German compounds "count" as SWT, the MWT number is comparatively low. Our statistical methods are however also applied to compounds (cf. section 2.3). The inter-annotator agreement[8] ranges between moderate and substantial agreement, depending on the pattern. For multi-words, the kappa is 0.59 (moderate agreement) which is satisfactory considering the imbalanced distribution of categories.

4 Evaluation of noise reduction steps

4.1 Comparative evaluation of statistical measures

We compare the suitability of the measures mentioned in section 2.3, as a basis for further adjustments. The measures DS and CSmw seem to be the most suited overall (cf. figure 2), while TFITF

[8] We compute Fleiss' kappa (Fleiss, 1971). Interpretation according to (Landis and Koch, 1977).

term	freq	f-rank	DS-rank	TFITF-rank	CSmw-rank
Drehmomentvorwahl (torque pre-selection)	33	2,344	65	314	67
Bohrer (drill)	1,094	44	158,094	40	2,554
Mutter (screw nut/mother)	510	133	216,341	2,276	38,036

Table 3: Ranked candidate term examples

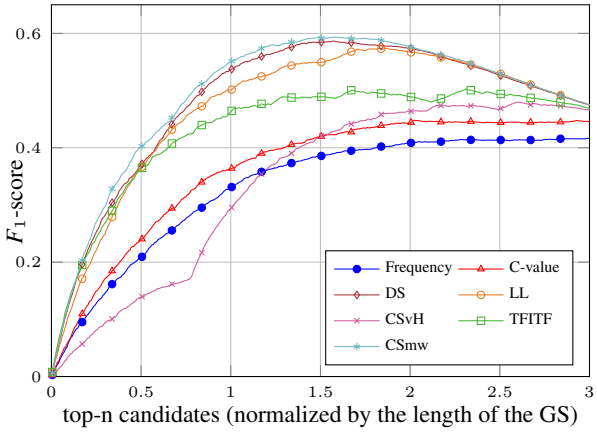


Figure 2: Statistical measures for term extraction

produces superior results only for very short candidate term lists. CSmw achieves a maximum F_1-score of 0.59 (48% precision and 77% recall) which is an improvement of 20 % over the simple frequency baseline. For very short lists TFITF achieves a precision of above 80%, for example 84% in the top 150 extracted candidates where other measures barely reach 70%. In the following we analyze these observations in detail and illustrate the reasons with examples from the extraction result as presented in table 3. This table shows in columns from left to right: extracted terms, their frequency in the domain corpus and their rank in the result lists according to the measures frequency, DS, TFITF and CSmw.

Sorting the candidates by their **DS**-value shows a high density of highly domain-specific terms at the top. For example, '*Drehmomentvorwahl*' (torque pre-selection) on rank 65 (out of 226,715 candidates). DS seems to strongly focus on very specialized terms of the domain texts which do not occur in general language, or only with very low frequency. However, as a consequence it misses important domain terms which also occur with a moderate frequency in general language. For example the term '*Bohrer*' (drill) which is essential for the domain (domain corpus frequency: 1,094) is only on DS-rank 158,094. This shows that the DS-value approach is not suitable to provide a list of important terms in the domain, but rather to identify its very specialized terms.

The **TFITF**-measure determines termhood by including the domain corpus frequency of a candidate term logarithmically (cf. formula in section 2.3) - unlike the DS-measure which uses it relatively. As a result several top candidate terms of the TFITF list are of a different kind than the best DS terms. For example, the above-mentioned term '*Bohrer*' (drill) is now at rank 40. TFITF even puts the candidate '*Mutter*' (screw nut/mother) which, due to its homography, is hard for a terminology extractor to identify as a term, at rank 2,276; this is acceptable, considering that there are 5,097 terms listed in the gold standard. The measure puts a stronger emphasis on the domain corpus frequency producing a considerable amount of noise in form of general-language candidates, which explains its rather mediocre overall performance. We thus suggest to use TFITF to extract a relatively small set of terms with a very high precision, for example for bootstrapping approaches or for an ontology learning which not only focuses on special technical terms of the domain but rather on its key topics.

The results acquired by the measure **CSmw** are in between TFITF and DS. It also gives more emphasis to the domain corpus frequency of a candidate term than DS, however not as much as TFITF. In the statistical analysis this approach reached the best overall F-scores. The CSmw-measure identifies the same top terms as the DS-measure, namely those which are highly domain specific and rare in general language, for example: '*Drehmomentvorwahl*' (torque pre-selection) on rank 67. Furthermore, it ranks comparatively high the essential objects of the domain which are also used in general language, for example: '*Bohrer*' (drill) is on rank 2,554. Consequently, the measure also produces some noise (as TFITF does) which is why it does not outperform the DS-measure. The ranking by CSmw turns out to be the most recommendable for a general terminology extraction which focuses on

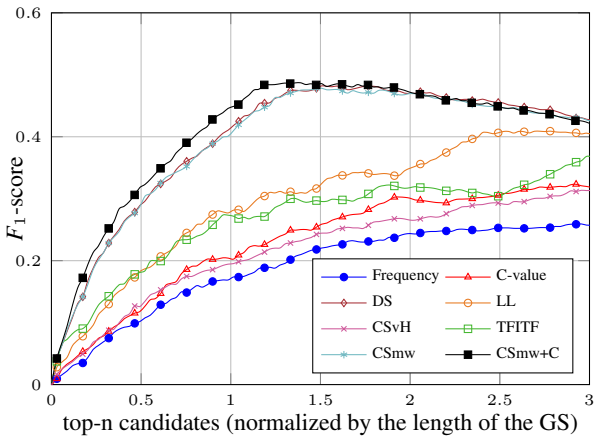


Figure 3: Statistical measures for MWT extraction

technical terms as well as on essential objects of the domain.

While **LL** outperforms the frequency baseline, for our domain it has proven to be one of the weaker measures and we could not identify any further useful characteristics.

The unsatisfactory result of the ranking by **CSvH** is based on its strong emphasis on MWTs which have a head with a very high domain corpus frequency. Thus, the result lists show many general-language candidates at their top which is why the measure underperforms the baseline.

The above results were obtained from an evaluation of a mixed set of SWTs and MWTs. A separate analysis of the MWT extraction (cf. figure 3) shows that the two best-performing measures (DS and CSmw) achieve an F_1-score of only about 0.49 (recall 64%, precision 40%). In comparison, the maximum F_1-score for the extraction of single-word terms was about 0.65. The low performance of the MWT extraction is due to the fact that this task also includes the determination of the right length of the MWT and therefore leads to more noise (in total approximately 80% noise in the MWT candidates selected by the basic POS patterns). Thus, a further filtering of the multi-word candidates is necessary.

4.2 Effect of C-value

Out of the 226,715 items that follow our extended patterns (frequency ≥ 1), C-value successfully removes 58,491 cases of noise that only occur embedded (25.8%). In our GS-based evaluation, C-value outperforms mere frequency, as shown in figure 2 in the extraction of the basic term patterns.

candidate term	freq	C-value
Band	301	296.50
Klebeband	376	707.33
doppelseitiges Klebeband	117	342.00

Table 4: Comparison of frequency to C-value

The positive effect of C-value is illustrated with a few examples in table 4. The frequency of occurrence in the domain corpus of the first candidate term '*Band*' (tape) is relatively similar to the one of the second candidate '*Klebeband*' (adhesive tape). They are both single-word nouns and thus would be considered almost equally as terms for the domain. After applying the C-value approach however their termhood values differ clearly with '*Klebeband*' having a value twice as high as the value of '*Band*'. This mainly follows from the term length computation based on the number of components of compounds ('*Klebeband*' is a compound with two components: '*kleben*' (to glue) and '*Band*'). Furthermore, the frequency of '*doppelseitiges Klebeband*' (double-sided adhesive tape) is only approximately a third of the frequency of the single noun '*Klebeband*'. The C-value method here also rewards the length of the multi-word and computes a value that is half of the C-value of the single noun despite the much lower absolute frequency of '*doppelseitiges Klebeband*'. Note that the termhood value of '*doppelseitiges Klebeband*' is also greater than the value for '*Band*' even though it has a lower frequency. This shows that the fine-grained measurement of the length characteristic of candidate terms including a special treatment for compounds is beneficial for terminology extraction.

However, it has to be noted that the ranking of candidate terms by C-value alone is not sufficient for term extraction, as extracted top lists with a recall of greater than 50% still contain a considerable amount of noise (at least 78%), mostly in the form of general-language candidates.

We found that CSmw, one of the best-performing measures in the comparative evaluation, improves when domain-specificity is computed on C-value instead of frequency (*CSmw+C*-plot in Figure 3, maximum F_1-score 0.51). This is due to C-value's sensitivity for nested terms which is combined with the domain-specificity filter from CSmw.

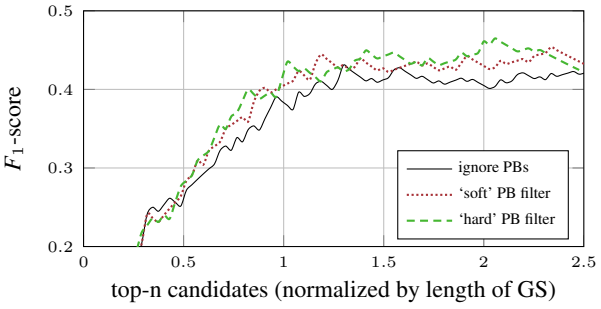


Figure 4: Syntactic validity filter for N P N extraction

4.3 Effect of phrase boundaries

As the syntactic filter only affects POS sequences with prepositions whose proportion in the GS is rather small, the effect of phrase boundaries (PBs) cannot be shown in Figure 3 and was thus tested in two other settings. First, we evaluated against the 107 N P N terms in the gold standard. Figure 4 shows the effects of applying the "soft" filter (frequency adjustments, section 2.1) and the "hard" filter (candidate removed altogether) on the F_1-score. Both filters clearly improve the standard extraction based on CSmw. However, the filter affects more than just the terms in the GS (17,4% of all NP+PP candidate occurrences affected) and we would like to observe the effects on all variants of prepositional patterns. Thus, in a precision-based evaluation, we ranked the MWT candidates by the number of times they violated the syntactic filter and manually checked, for the top 500 removal candidates, whether the removal was justified. The result, as shown in Table 5, indicates that the quality of the parser output is sufficient to predict syntactic validity: the overall precision for these top 500 candidates was 83%.

Top n	50	100	150	200	250
Precision	0.76	0.75	0.78	0.81	0.81
Top n	300	350	400	450	500
Precision	0.82	0.83	0.82	0.82	**0.83**

Table 5: Top-n manual plausibility check for "hard" filter

5 Conclusion and future work

We presented three steps to remove noise and to increase performance in nominal terminology extraction. We also suggested a combination of statistical measures that is particulary suitable for this task: Our best setting has proven to be a combination of C-value and CSmw, together with a syntactic validity check. A qualitative analysis of the extraction results showed that different termhood measures emphasize different characteristics of terms, as their top lists differ. Therefore, a combination of statistical measures can also be considered for further improvements, instead of only focusing on one single best performing measure. One could for example think of ways to combine the top lists of a set of best-performing measures, or try an approach that combines or ranks different scores of certain measures in one formula. Future work will also be based on English data where we will evaluate further steps to improve term extraction results, e.g. by combining the termhood measures also with association measures and by further improving the syntactic analysis through the use of an additional constituency parser. A further objective of this work will be to assess the generality of the approach on different domains.

References

Khurshid Ahmad, Andrea Davies, Heather Fulford, and Margaret Rogers. 1992. What is a term? The semi-automatic extraction of terms from text. *Translation Studies: An Interdiscipline: Selected papers from the Translation Studies Congress, Vienna, 1994*, pages 267–278.

Khurshid Ahmad, Lee Gillam, and Lena Tostevin. 1999. University of Surrey Participation in TREC8: Weirdness Indexing for Logical Document Extrapolation and Retrieval (WILDER). In *Text REtrieval Conference*.

Roberto Basili, Alessandro Moschitti, Maria Teresa Pazienza, and Fabio Massimo Zanzotto. 2001. A contrastive approach to term extraction. In *Terminologie et intelligence artificielle. Rencontres*, pages 119–128.

Bernd Bohnet. 2010. Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), Beijing, China*, pages 89–97. Association for Computational Linguistics.

Francesca Bonin, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2010. A contrastive approach to multi-word term extraction from domain corpora. In *Proceedings of the 7th International Conference on Language Resources and Evaluation, Malta*, pages 19–21.

Maria Teresa Cabre and Jorge Vivaldi Palatresi. 2013. Acquisition of terminological data from text: Approaches. In *Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguis-*

tics and Communication Science (HSK) 5/4, pages 1486–1497. DeGruyter Mouton.

Fabienne Cap. 2014. Morphological processing of compounds for statistical machine translation. PhD thesis, Institute for Natural Language Processing (IMS), University of Stuttgart, `http://elib.uni-stuttgart.de/opus/volltexte/2014/9768/`.

Chaomei Chen, Fidelia Ibekwe-SanJuan, Eric SanJuan, and Michael Vogeley. 2008. Identifying thematic variations in sdss research. In *6th International Conference on Language Resources and Evaluation Conference (LREC-08)*, pages 319–330. Presses Universitaires de Lyon.

Jinho D. Choi, Joel R. Tetreault, and Amanda Stent. 2015. It depends: Dependency parser comparison using a web-based evaluation tool. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 387–396.

Jean-François Couturier, Sylvain Neuvel, and Patrick Drouin. 2006. Applying Lexical Constraints on Morpho-Syntactic Patterns for the Identification of Conceptual-Relational Content in Specialized Texts. *Language Resources and Evaluation Conference (LREC-06), Genoa, Italy*.

Stefan Evert. 2005. The Statistics of Word Cooccurrences: Word Pairs and Collocations. PhD thesis, Universität Stuttgart, `http://elib.uni-stuttgart.de/opus/volltexte/2005/2371`.

Gertrud Faaß and Kerstin Eckart. 2013. SdeWaC – A Corpus of Parsable Sentences from the Web. In *Iryna Gurevych, Chris Biemann, and Torsten Zesch, editors, Language processing and knowledge in the Web: 25th International Conference, GSCL 2013, Darmstadt, Germany, volume 8105 of Lecture Notes in Computer Science*, pages 61–68. Springer.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Katerina T. Frantzi and Sophia Ananiadou. 1996. Extracting nested collocations. In *Proceedings of the 16th conference on Computational linguistics-Volume 1*, pages 41–46. Association for Computational Linguistics.

Katerina T. Frantzi, Sophia Ananiadou, and Hideki Mima. 2000. Automatic recognition of multi-word terms: the c-value/nc-value method. *International Journal on Digital Libraries*, 3(2):115–130.

Tanja George. 2014. Comparing a commercial term extraction tool with a research prototype: an evaluation study on DIY instruction texts. Bachelor thesis, ms. Institute for Natural Language Processing (IMS), University of Stuttgart.

Anita Gojun, Ulrich Heid, Bernd Weißbach, Carola Loth, and Insa Mingers. 2012. Adapting and evaluating a generic term extraction tool. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA).

Kyo Kageura and Bin Umino. 1996. Methods of automatic term recognition: A review. *Terminology*, 3(2):259–289.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Elizaveta Loginova, Anita Gojun, Helena Blancafort, Marie Guégan, Tatiana Gornostay, and Ulrich Heid. 2012. Reference lists for the evaluation of term extraction tools. In *Proceedings of the Terminology and Knowledge Engineering Conference (TKE'2012)*.

Maria Teresa Pazienza, Marco Pennacchiotti, and Fabio Massimo Zanzotto. 2005. Terminology extraction: an analysis of linguistic and statistical approaches. In *Knowledge Mining*, pages 255–279. Springer.

Paul Rayson and Roger Garside. 2000. Comparing Corpora Using Frequency Profiling. In *Proceedings of the Workshop on Comparing Corpora - Volume 9*, WCC '00, pages 1–6, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mathieu Roche, Jérôme Azé, Yves Kodratoff, and Michele Sebag. 2004. Learning interestingness measures in terminology extraction. a roc-based approach. In *"ROC Analysis in AI" Workshop (ECAI 2004), Valencia, Spain*, pages 81–88.

Johannes Schäfer. 2015. Statistical and parsing-based approaches to the extraction of multi-word terms from texts: implementation and comparative evaluation. Bachelor thesis, ms. Institute for Natural Language Processing (IMS), University of Stuttgart.

Helmut Schmid and Florian Laws. 2008. Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 777–784. Association for Computational Linguistics.

Marie Zollmann, Ina Rösiger, Ulrich Heid, and Michael Dorna. 2016. Nutzen von Parsing in der Termextraktion: eine qualitative und quantitative korpusbasierte Untersuchung. In *Poster session of the DGfS conference 2016 (to appear)*.