Paradigmatic Modifiability Statistics for the Extraction of Complex Multi-Word Terms

Joachim Wermter

Jena University Language & Information Engineering JULIE Lab Friedrich-Schiller-Universität Jena Fürstengraben 30 D-07743 Jena, Germany joachim.wermter@uni-jena.de

Abstract

We here propose a new method which sets apart domain-specific terminology from common non-specific noun phrases. It is based on the observation that terminological multi-word groups reveal a considerably lesser degree of distributional variation than non-specific noun phrases. We define a measure for the observable amount of paradigmatic modifiability of terms and, subsequently, test it on bigram, trigram and quadgram noun phrases extracted from a 104-million-word biomedical text corpus. Using a community-wide curated biomedical terminology system as an evaluation gold standard, we show that our algorithm significantly outperforms a variety of standard term identification measures. We also provide empirical evidence that our methodolgy is essentially domain- and corpus-size-independent.

Introduction 1

As we witness the ever-increasing proliferation of volumes of medical and biological documents, the available dictionaries and terminological systems cannot keep up with this pace of growth and, hence, become more and more incomplete. What's worse, the constant stream of new terms is increasingly getting unmanageable because human curators are in the loop. The costly, often error-prone and timeconsuming nature of manually identifying new terminology from the most recent literature calls for

Udo Hahn

Jena University Language & Information Engineering JULIE Lab Friedrich-Schiller-Universität Jena Fürstengraben 30 D-07743 Jena, Germany udo.hahn@uni-jena.de

advanced procedures which can automatically assist database curators in the task of assembling, updating and maintaining domain-specific controlled vocabularies. Whereas the recognition of single-word terms usually does not pose any particular challenges, the vast majority of biomedical or any other domain-specific terms typically consists of multiword units.¹ Unfortunately these are much more difficult to recognize and extract than their singleton counterparts. Moreover, although the need to assemble and extend technical and scientific terminologies is currently most pressing in the biomedical domain, virtually any (sub-)field of human research/expertise in which we deal with terminologically structured knowledge calls for high-performance terminology identification and extraction methods. We want to target exactly this challenge.

Related Work 2

The automatic extraction of complex multi-word terms from domain-specific corpora is already an active field of research (cf., e.g., for the biomedical domain Rindflesch et al. (1999), Collier et al. (2002), Bodenreider et al. (2002), or Nenadić et al. (2003)). Typically, in all of these approaches term candidates are collected from texts by various forms of linguistic filtering (part-of-speech tagging, phrase chunking, etc.), through which candidates obeying various linguistic patterns are identified (e.g., noun-noun, adjective-noun-noun combinations). These candidates are then submitted to frequency- or statistically-based evidence measures

¹Nakagawa and Mori (2002) claim that more than 85% of domain-specific terms are multi-word units.

(such as the C-value (Frantzi et al., 2000)), which compute scores indicating to what degree a candidate qualifies as a term. Term mining, as a whole, is a complex process involving several other components (orthographic and morphological normalization, acronym detection, conflation of term variants, term context, term clustering; cf. Nenadić et al. (2003)). Still, the measure which assigns a *termhood value* to a term candidate is the essential building block of any term identification system.

For multi-word automatic term recognition (ATR), the C-value approach (Frantzi et al., 2000; Nenadić et al., 2004), which aims at improving the extraction of nested terms, has been one of the most widely used techniques in recent years. Other potential association measures are mutual information (Damerau, 1993) and the whole battery of statistical and information-theoretic measures (t-test, log-likelihood, entropy) which are typically employed for the extraction of general-language collocations (Manning and Schütze, 1999; Evert and Krenn, 2001). While these measures have their statistical merits in terminology identification, it is interesting to note that they only make little use of linguistic properties inherent to complex terms.²

More linguistically oriented work on ATR by Daille (1996) or on term variation by Jacquemin (1999) builds on the deep syntactic analysis of term candidates. This includes morphological and headmodifier dependency analysis and thus presupposes accurate, high-quality parsing which, for sublanguages at least, can only be achieved by a highly domain-dependent type of grammar. As sublanguages from different domains usually reveal a high degree of syntactic variability among each other (e.g., in terms of POS distribution, syntactic patterns), this property makes it difficult to port grammatical specifications to different domains.

Therefore, one may wonder whether there are cross-domain linguistic properties which might be beneficial to ATR and still could be accounted for by only *shallow* syntactic analysis. In this paper, we propose the *limited paradigmatic modifiability* of terms as a criterion which meets these requirements and will elaborate on it in detail in Subsection 3.3.

3 Methods and Experiments

3.1 Text Corpus

We collected a biomedical training corpus of approximately 513,000 MEDLINE abstracts using the following query composed of MESH terms from the biomedical domain: transcription factors, blood cells and human.³ We then annotated the resulting 104-million-word corpus with the GENIA partof-speech tagger⁴ and identified noun phrases (NPs) with the YAMCHA chunker (Kudo and Matsumoto, 2001). We restrict our study to NP recognition (i.e., determining the extension of a noun phrase but refraining from assigning any internal constituent structure to that phrase), because the vast majority of technical or scientific terms surface as noun phrases (Justeson and Katz, 1995). We filtered out a number of stop words (determiners, pronouns, measure symbols, etc.) and also ignored noun phrases with coordination markers ("and", "or", etc.).⁵

n-gram	cut-off	NP term candidates		
length	ength		types	
bigrams	no cut-off	5,920,018	1,055,820	
	$c \ge 10$	4,185,427	67,308	
trigrams	no cut-off	3,110,786	1,655,440	
	$c \ge 8$	1,053,651	31,017	
quadgrams	no cut-off	1,686,745	1,356,547	
	$c \ge 6$	222,255	10,838	

 Table 1: Frequency distribution for noun phrase term candidate tokens and types for the MEDLINE text corpus

In order to obtain the term candidate sets (see Table 1), we counted the frequency of occurrence of noun phrases in our training corpus and categorized them according to their length. For this study, we restricted ourselves to noun phrases of length 2 (word bigrams), length 3 (word trigrams) and length 4 (word quadgrams). Morphological normalization of term candidates has shown to be beneficial for ATR (Nenadić et al., 2004). We thus normalized the nom-

²A notable exception is the C-value method which incorporates a term's likelihood of being nested in other multi-word units.

³MEDLINE (http://www.ncbi.nlm.nih.gov) is the largest biomedical bibliographic database. For information retrieval purposes, all of its abstracts are indexed with a controlled indexing vocabulary, the *Medical Subject Headings* (MESH, 2004).

⁴http://www-tsujii.is.s.u-tokyo.ac.jp/ GENIA/postagger/

⁵Of course, terms can also be contained within coordinative structures (e.g., "*B and T cell*"). However, analyzing their inherent ambiguity is a complex syntactic operation, with a comparatively marginal benefit for ATR (Nenadić et al., 2004).

inal head of each noun phrase (typically the rightmost noun in English) via the full-form UMLS SPE-CIALIST LEXICON (UMLS, 2004), a large repository of both general-language and domain-specific (medical) vocabulary. To eliminate noisy low-frequency data (cf. also Evert and Krenn (2001)), we defined different frequency cut-off thresholds, *c*, for the bigram, trigram and quadgram candidate sets and only considered candidates above these thresholds.

3.2 Evaluating Term Extraction Quality

Typically, terminology extraction studies evaluate the goodness of their algorithms by having their ranked output examined by domain experts who identify the true positives among the ranked candidates. There are several problems with such an approach. First, very often only one such expert is consulted and, hence, inter-annotator agreement cannot be determined (as, e.g., in the studies of Frantzi et al. (2000) or Collier et al. (2002)). Furthermore, what constitutes a relevant term for a particular domain may be rather difficult to decide even for domain experts - when judges are just exposed to a list of candidates without any further context information. Thus, rather than relying on ad hoc human judgment in identifying true positives in a candidate set, as an alternative we may take already existing terminolgical resources into account. They have evolved over many years and usually reflect community-wide consensus achieved by expert committees. With these considerations in mind, the biomedical domain is an ideal test bed for evaluating the goodness of ATR algorithms because it hosts one of the most extensive and most carefully curated terminological resources, viz. the UMLS METATHE-SAURUS (UMLS, 2004). We will then take the mere existence of a term in the UMLS as the decision criterion whether or not a candidate term is also recognized as a biomedical term.

Accordingly, for the purpose of evaluating the quality of different measures in recognizing multiword terms from the biomedical literature, we assign every word bigram, trigram, and quadgram in our candidate sets (see Table 1) the status of being a term (i.e., a true positive), if it is found in the 2004 edition of the UMLS METATHESAURUS.⁶ For example, the word trigram "long terminal repeat" is listed as a term in one of the UMLS vocabularies, *viz.* MESH (2004), whereas "*t cell response*" is not. Thus, among the 67,308 word bigram candidate types, 14,650 (21.8%) were identified as true terms; among the 31,017 word trigram candidate types, their number amounts to 3,590 (11.6%), while among the 10,838 word quadgram types, 873 (8.1%) were identified as true terms.⁷

3.3 Paradigmatic Modifiability of Terms

For most standard association measures utilized for terminology extraction, the frequency of occurrence of the term candidates either plays a major role (e.g., C-value), or has at least a significant impact on the assignment of the degree of termhood (e.g., t-test). However, frequency of occurrence in a training corpus may be misleading regarding the decision whether or not a multi-word expression is a term. For example, taking the two trigram multi-word expressions from the previous subsection, the non-term "*t cell response*" appears 2410 times in our 104-million-word MEDLINE corpus, whereas the term "*long terminal repeat*" (long repeating sequences of DNA) only appears 434 times (see also Tables 2 and 3 below).

The linguistic property around which we built our measure of termhood is the *limited paradigmatic modifiability* of multi-word terminological units. A multi-word expression such as "*long terminal repeat*" contains three token slots in which slot 1 is filled by "*long*", slot 2 by "*terminal*" and slot 3 by "*repeat*". The *limited paradigmatic modifiability* of such a trigram is now defined by the probability with which one or more such slots *cannot* be filled by other tokens. We estimate the likelihood of precluding the appearance of alternative tokens in particular slot positions by employing the standard combinatory formula without repetitions. For an n-gram (of size n) to select k slots (i.e., in an unordered selection) we thus define:

$$C(n,k) = \frac{n!}{k!(n-k)!}$$
 (1)

⁶We exclude UMLS vocabularies not relevant for molecular biology, such as nursing and health care billing codes.

⁷As can be seen, not only does the number of candidate types drop with increasing n-gram length but also the proportion of true terms. In fact, their proportion drops more sharply than can actually be seen from the above data because the various cut-off thresholds have a leveling effect.

For example, for n = 3 (word trigram) and k = 1and k = 2 slots, there are three possible selections for each k for "*long terminal repeat*" and for "t cell response" (see Tables 2 and 3). k is actually a placeholder for any possible token (and its frequency) which fills this position in the training corpus.

n-gram		freq	<i>P-Mod</i> (k=1,2)		
long terminal repeat		434		0.03	
k slots	possible select	ions <i>sel</i>	freq	mod_{sel}	
k = 1	k_1 terminal repeat		460	0.940	
	long k_2 repeat	448	0.970		
	long terminal k	436	0.995		
			mo	$d_1 = 0.91$	
k = 2	$k_1 k_2$ repeat		1831	0.23	
	k_1 terminal k_3		1062	0.41	
	long $k_2 k_3$		1371	0.32	
			mo	$d_2 = 0.03$	

Table 2: *P*-*M*od and *k*-modifiabilities for k = 1 and k = 2 for the trigram term "*long terminal repeat*"

n-gram		freq	<i>P-Mod</i> (k=1,2)		
t cell response		2410	0.00005		
k slots	possible selections sel		freq	mod_{sel}	
k = 1	k_1 cell response		3248	0.74	
	t k_2 response		2665	0.90	
	t cell k_3		27424	0.09	
			mo	$d_1 = 0.06$	
k = 2	$k_1 k_2$ response	;	40143	0.06	
	$k_1 \text{ cell } k_3$		120056	0.02	
	t $k_2 k_3$		34925	0.07	
			$mod_2 =$	=0.00008	

Table 3: *P*-*Mod* and *k*-modifiabilities for k = 1 and k = 2 for the trigram non-term "*t cell response*"

Now, for a particular k $(1 \le k \le n; n = \text{length of } n\text{-gram})$, the frequency of each possible selection, *sel*, is determined. The paradigmatic modifiability for a particular selection *sel* is then defined by the n-gram's frequency scaled against the frequency of *sel*. As can be seen in Tables 2 and 3, a *lower* frequency induces a *more limited* paradigmatic modifiability for a particular *sel* (which is, of course, expressed as a higher probability value; see the column labeled mod_{sel} in both tables). Thus, with *s* being the number of distinct possible selections for a particular *k*, the *k*-modifiability, mod_k , of an n-gram can be defined as follows (*f* stands for frequency):

$$mod_k(n\text{-}gram) := \prod_{i=1}^s \frac{f(n\text{-}gram)}{f(sel_i, n\text{-}gram)}$$
 (2)

The *paradigmatic modifiability*, P-Mod, of an n-gram is the product of all its k-modifiabilities:⁸

$$P-Mod(n-gram) := \prod_{k=1}^{n} mod_k(n-gram)$$
(3)

Comparing the trigram P-Mod values for k =1, 2 in Tables 2 and 3, it can be seen that the term "long terminal repeat" gets a much higher weight than the non-term "t cell response", although their mere frequency values suggest the opposite. This is also reflected in the respective list rank (see Subsection 4.1 for details) assigned to both trigrams by the t-test and by our P-Mod measure. While "t cell response" has rank 24 on the t-test output list (which directly reflects its high frequency), P-Mod assigns rank 1249 to it. Conversely, "long terminal repeat" is ranked on position 242 by the t-test, whereas it occupies rank 24 for P-Mod. In fact, even lowerfrequency multi-word units gain a prominent ranking, if they exhibit limited paradigmatic modifiability. For example, the trigram term "porphyria cutanea tarda" is ranked on position 28 by P-Mod, although its frequency is only 48 (which results in rank 3291 on the t-test output list). Despite its lower frequency, this term is judged as being relevant for the molecular biology domain.9 It should be noted that the termhood values (and the corresponding list ranks) computed by P-Mod also include k = 3 and, hence, take into account a reasonable amount of frequency load. As can be seen from the previous ranking examples, still this factor does not override the paradigmatic modifiability factors of the lower ks.

On the other hand, *P-Mod* will also demote true terms in their ranking, if their paradigmatic modifiability is less limited. This is particularly the case if one or more of the tokens of a particular term often occur in the same slot of other equal-length n-grams. For example, the trigram term *"bone marrow cell"* occurs 1757 times in our corpus and is thus ranked quite high (position 31) by the t-test. *P-Mod*, however, ranks this term on position 550 because the to-

⁸Setting the upper limit of k to n (e.g., n = 3 for trigrams) actually has the pleasant side effect of including frequency in our modifiability measure. In this case, the only possible selection $k_1k_2k_3$ as the denominator of Formula (2) is equivalent to summing up the frequencies of all trigram term candidates.

⁹It denotes a group of related disorders, all of which arise from a deficient activity of the heme synthetic enzyme uroporphyrinogen decarboxylase (URO-D) in the liver.

ken "*cell*" also occurs in many other trigrams and thus leads to a less limited paradigmatic modifiability. Still, the underlying assumption of our approach is that such a case is more an exception than the rule and that terms are linguistically more 'frozen' than non-terms, which is exactly the intuition behind our measure of limited paradigmatic modifiability.

3.4 Methods of Evaluation

As already described in Subsection 3.2, standard procedures for evaluating the quality of termhood measures usually involve identifying the true positives among a (usually) arbitrarily set number of the m highest ranked candidates returned by a particular measure, a procedure usually carried out by a domain expert. Because this is labor-intensive (besides being unreliable), m is usually small, ranging from 50 to several hundreds.¹⁰ By contrast, we choose a large and already consensual terminology to identify the true terms in our candidate sets. Thus, we are able to dynamically examine various *m*-highest ranked samples, which, in turn, allows for the plotting of standard precision and recall graphs for the entire candidate set. We thus provide a more reliable evaluation setting for ATR measures than what is common practice in the literature.

We compare our P-Mod algorithm against the t-test measure,¹¹ which, of all standard measures, yields the best results in general-language collocation extraction studies (Evert and Krenn, 2001), and also against the widely used C-value, which aims at enhancing the common frequency of occurrence measure by making it sensitive to nested terms (Frantzi et al., 2000). Our baseline is defined by the proportion of true positives (i.e., the proportion of terms) in our bi-, tri- and quadgram candidate sets. This is equivalent to the likelihood of finding a true positive by blindly picking from one of the different sets (see Subsection 3.2).

4 **Results and Discussion**

4.1 Precision/Recall for Terminology Extraction

For each of the different candidate sets, we incrementally examined portions of the ranked output lists returned by each of the three measures we considered. The precision values for the various portions were computed such that for each percent point of the list, the number of true positives found (i.e., the number of terms) was scaled against the overall number of candidate items returned. This yields the (descending) precision curves in Figures 1, 2 and 3 and some associated values in Table 4.

	Portion of	Precision scores of measures			
	ranked list				
	considered	P-Mod	t-test	C-value	
	1%	0.82	0.62	0.62	
Bigrams	10%	0.53	0.42	0.41	
	20%	0.42	0.35	0.34	
	30%	0.37	0.32	0.31	
	baseline	0.22	0.22	0.22	
	1%	0.62	0.55	0.54	
Trigrams	10%	0.37	0.29	0.28	
	20%	0.29	0.23	0.23	
	30%	0.24	0.20	0.19	
	baseline	0.12	0.12	0.12	
	1%	0.43	0.50	0.50	
Quadgrams	10%	0.26	0.24	0.23	
	20%	0.20	0.16	0.16	
	30%	0.18	0.14	0.14	
	baseline	0.08	0.08	0.08	

 Table 4: Precision scores for biomedical term extraction at selected portions of the ranked list

First, we observe that, for the various n-gram candidate sets examined, all measures outperform the baselines by far, and, thus, all are potentially useful measures for grading termhood. Still, the *P-Mod* criterion substantially outperforms all other measures at almost all points for all n-grams examined. Considering 1% of the bigram list (i.e., the first 673 candidates) precision for *P-Mod* is 20 points higher than for the t-test and the C-value. At 1% of the trigram list (i.e., the first 310 candidates), P-Mod's lead is 7 points. Considering 1% of the quadgrams (i.e., the first 108 candidates), the t-test actually leads by 7 points. At 10% of the quadgram list, however, the *P*-Mod precision score has overtaken the other ones. With increasing portions of all ranked lists considered, the precision curves start to converge toward the baseline, but *P*-*Mod* maintains a steady advantage.

¹⁰Studies on collocation extraction (e.g., by Evert and Krenn (2001)) also point out the inadequacy of such evaluation methods. In essence, they usually lead to very superficial judgments about the measures under scrutiny.

¹¹Manning and Schütze (1999) describe how this measure can be used for the extraction of multi-word expressions.



Figure 1: Precision/Recall for bigram biomedical term extraction



Figure 2: Precision/Recall for trigram biomedical term extraction



Figure 3: Precision/Recall for quadgram biomedical term extraction

The (ascending) recall curves in Figures 1, 2 and 3 and their corresponding values in Table 5 indicate which *proportion of all true positives* (i.e., the proportion of all terms in a candidate set) is identified by a particular measure at a certain point of the ranked list. For term extraction, recall is an even better indicator of a particular measure's performance because finding a bigger proportion of the true terms at an early stage is simply more economical.

	Recall	Portion of Ranked List			
	scores of				
	measures	P-Mod	t-test	C-value	
	0.5	29%	35%	37%	
	0.6	39%	45%	47%	
Bigrams	0.7	51%	56%	59%	
	0.8	65%	69%	72%	
	0.9	82%	83%	85%	
	0.5	19%	28%	30%	
Trigrams	0.6	27%	38%	40%	
	0.7	36%	50%	53%	
	0.8	50%	63%	66%	
	0.9	68%	77%	84%	
	0.5	20%	28%	30%	
	0.6	26%	38%	40%	
Quadgrams	0.7	34%	49%	53%	
	0.8	45%	62%	65%	
	0.9	61%	79%	82%	

 Table 5: Portions of the ranked list to consider for selected recall scores for biomedical term extraction

Again, our linguistically motivated terminology extraction algorithm outperforms its competitors, and with respect to tri- and quadgrams, its gain is even more pronounced than for precision. In order to get a 0.5 recall for bigram terms, *P-Mod* only needs to winnow 29% of the ranked list, whereas the t-test and C-value need to winnow 35% and 37%, respectively. For trigrams and quadgrams, *P-Mod* only needs to examine 19% and 20% of the list, whereas the other two measures have to scan almost 10 additional percentage points. In order to obtain a 0.6, 0.7, 0.8 and 0.9 recall, the differences between the measures narrow for bigram terms, but they widen substantially for tri- and quadgram terms. To obtain a 0.6 recall for trigram terms, P-Mod only needs to winnow 27% of its output list while the t-test and 100 C-value must consider 38% and 40%, respectively. For a level of 0.7 recall, *P-Mod* only needs to analyze 36%, while the t-test already searches 50% of the ranked list. For 0.8 recall, this relation is 50% (P-Mod) to 63% (t-test), and at recall point 0.9, 68% (P-Mod) to 77% (t-test). For quadgram term identification, the results for P-Mod are equally superior to those for the other measures, and at recall points 0.8 and 0.9 even more pronounced than for trigram terms.

We also tested the significance of differences for these results, both comparing *P*-Mod vs. t-test and *P-Mod vs.* C-value. Because in all cases the ranked lists were taken from the same set of candidates (viz. the set of bigram, trigram, and quadgram candidate types), and hence constitute dependent samples, we applied the McNemar test (Sachs, 1984) for statistical testing. We selected 100 measure points in the ranked lists, one after each increment of one percent, and then used the two-tailed test for a confidence interval of 95%. Table 6 lists the number of significant differences for these measure points at intervals of 10 for the bi-, tri-, and quadgram results. For the bigram differences between P-Mod and C-value, all of them are significant, and between P-Mod and t-test, all are significantly different up to measure point 70.12 Looking at the tri- and guadgrams, although the number of significant differences is less than for bigrams, the vast majority of measure points is still significantly different and thus underlines the superior performance of the *P*-*Mod* measure.

# of	# of significant differences comparing					
measure	P-Mod with					
points	t-test	C-val	t-test	C-val	t-test	C-val
10	10	10	9	9	3	3
20	20	20	19	19	13	13
30	30	30	29	29	24	24
40	40	40	39	39	33	33
50	50	50	49	49	43	43
60	60	60	59	59	53	53
70	70	70	69	69	63	63
80	75	80	79	79	73	73
90	84	90	89	89	82	83
100	93	100	90	98	82	91
	bigrams trigrams		quad	grams		

Table 6: Significance testing of differences for bi-, tri- and quadgrams using the two-tailed McNemar test at 95% confidence interval

4.2 Domain Independence and Corpus Size

One might suspect that the results reported above could be attributed to the corpus size. Indeed, the text collection we employed in this study is rather large (104 million words). Other text genres and domains (e.g., clinical narratives, various engineering domains) or even more specialized biological subdomains (e.g., plant biology) do not offer such a plethora of free-text material as the molecular biology domain. To test the effect a drastically shrunken corpus size might have, we assessed the terminology extraction methods for trigrams on a much smaller-sized subset of our original corpus, *viz.* on 10 million words. These results are depicted in Figure 4.



Figure 4: Precision/Recall for trigram biomedical term extraction on the 10-million-word corpus (cutoff $c \ge 4$, with 6,760 term candidate types)

The P-Mod extraction criterion still clearly outperforms the other ones on that 10-million-word corpus, both in terms of precision and recall. We also examined whether the differences were statistically significant and applied the two-tailed McNemar test on 100 selected measure points. Comparing P-Modwith t-test, most significant differences could be observed between measure points 20 and 80, with almost 80% to 90% of the points being significantly different. These significant differences were even more pronounced when comparing the results between P-Mod and C-value.

5 Conclusions

We here proposed a new terminology extraction method and showed that it significantly outperforms

¹²As can be seen in Figures 1, 2 and 3, the curves start to merge at the higher measure points and, thus, the number of significant differences decreases.

two of the standard approaches in distinguishing terms from non-terms in the biomedical literature. While mining scientific literature for new terminological units and assembling those in controlled vocabularies is a task involving several components, one essential building block is to measure the degree of termhood of a candidate. In this respect, our study has shown that a criterion which incorporates a vital linguistic property of terms, viz. their limited paradigmatic modifiability, is much more powerful than linguistically more uninformed measures. This is in line with our previous work on generallanguage collocation extraction (Wermter and Hahn, 2004), in which we showed that a linguistically motivated criterion based on the limited syntagmatic modifiability of collocations outperforms alternative standard association measures as well.

We also collected evidence that the superiority of the P-Mod method relative to other term extraction approaches holds independent of the underlying corpus size (given a reasonable offset). This is a crucial finding because other domains might lack large volumes of free-text material but still provide sufficient corpus sizes for valid term extraction. Finally, since we only require shallow syntactic analysis (in terms of NP chunking), our approach might be well suited to be easily portable to other domains. Hence, we may conclude that, although our methodology has been tested on the biomedical domain only, there are essentially no inherent domain-specific restrictions.

Acknowledgements. This work was partly supported by the European Network of Excellence "Semantic Mining in Biomedicine" (NoE 507505).

References

- Olivier Bodenreider, Thomas C. Rindflesch, and Anita Burgun. 2002. Unsupervised, corpus-based method for extending a biomedical terminology. In Stephen Johnson, editor, *Proceedings of the ACL/NAACL 2002 Workshop on 'Natural Language Processing in the Biomedical Domain'*, pages 53– 60. Philadelphia, PA, USA, July 11, 2002.
- Nigel Collier, Chikashi Nobata, and Jun'ichi Tsujii. 2002. Automatic acquisition and classification of terminology using a tagged corpus in the molecular biology domain. *Terminology*, 7(2):239–257.
- Béatrice Daille. 1996. Study and implementation of combined techniques for automatic extraction of terminology. In Judith L. Klavans and Philip Resnik, editors, *The Balancing Act: Combining Statistical and Symbolic Approaches to Language*, pages 49–66. Cambridge, MA: MIT Press.

- Fred J. Damerau. 1993. Generating and evaluating domainoriented multi-word terms from text. *Information Process*ing & Management, 29(4):433–447.
- Stefan Evert and Brigitte Krenn. 2001. Methods for the qualitative evaluation of lexical association measures. In ACL'01/EACL'01 – Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics and the 10th Conference of the European Chapter of the ACL, pages 188–195. Toulouse, France, July 9-11, 2001.
- Katerina T. Frantzi, Sophia Ananiadou, and Hideki Mima. 2000. Automatic recognition of multi-word terms: The Cvalue/NC-value method. *International Journal on Digital Libraries*, 3(2):115–130.
- Christian Jacquemin. 1999. Syntagmatic and paradigmatic representations of term variation. In *Proceedings of the 37rd Annual Meeting of the Association for Computational Linguistics*, pages 341–348. College Park, MD, USA, 20-26 June 1999.
- John S. Justeson and Slava M. Katz. 1995. Technical terminology: Some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1(1):9–27.
- Taku Kudo and Yuji Matsumoto. 2001. Chunking with support vector machines. In NAACL'01, Language Technologies 2001 – Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics, pages 192–199. Pittsburgh, PA, USA, June 2-7, 2001.
- Christopher D. Manning and Hinrich Schütze. 1999. Foundations of Statistical Natural Language Processing. Cambridge, MA; London, U.K.: Bradford Book & MIT Press.
- Hiroshi Nakagawa and Tatsunori Mori. 2002. A simple but powerful automatic term extraction method. In COMPU-TERM 2002 – Proceedings of the 2nd International Workshop on Computational Terminology, pages 29–35. Taipei, Taiwan, August 31, 2002.
- Goran Nenadić, Irena Spasić, and Sophia Ananiadou. 2003. Terminology-driven mining of biomedical literature. *Bioin-formatics*, 19(8):938–943.
- Goran Nenadić, Sophia Ananiadou, and John McNaught. 2004. Enhancing automatic term recognition through recognition of variation. In COLING 2004 – Proceedings of the 20th International Conference on Computational Linguistics, pages 604–610. Geneva, Switzerland, August 23-27, 2004.
- Thomas C. Rindflesch, Lawrence Hunter, and Alan R. Aronson. 1999. Mining molecular binding terminology from biomedical text. In AMIA'99 – Proceedings of the Annual Symposium of the American Medical Informatics Association, pages 127–131. Washington, D.C., November 6-10, 1999.
- Lothar Sachs. 1984. Applied Statistics: A Handbook of Techniques. New York: Springer, 2nd edition.
- MESH. 2004. *Medical Subject Headings*. Bethesda, MD: National Library of Medicine.
- UMLS. 2004. Unified Medical Language System. Bethesda, MD: National Library of Medicine.
- Joachim Wermter and Udo Hahn. 2004. Collocation extraction based on modifiability statistics. In COLING 2004 – Proceedings of the 20th International Conference on Computational Linguistics, pages 980–986. Geneva, Switzerland, August 23-27, 2004.