Finding New Terminology in Very Large Corpora

Joachim Wermter
Jena University Language and Information
Engineering (JULIE) Lab
Fürstengraben 30
D-07743 Jena, Germany
joachim.wermter@uni-jena.de

Udo Hahn
Jena University Language and Information
Engineering (JULIE) Lab
Fürstengraben 30
D-07743 Jena, Germany
udo.hahn@uni-jena.de

ABSTRACT

Most technical and scientific terms are comprised of complex, multi-word noun phrases but certainly not all noun phrases are technical or scientific terms. The distinction of specific terminology from common non-specific noun phrases can be based on the observation that terms reveal a much lesser degree of distributional variation than non-specific noun phrases. We formalize the limited paradigmatic modifiability of terms and, subsequently, test the corresponding algorithm on bigram, trigram and quadgram noun phrases extracted from a 104-million-word biomedical text corpus. Using an already existing and community-wide curated biomedical terminology as an evaluation gold standard, we show that our algorithm significantly outperforms standard term identification measures and, therefore, qualifies as a high-performant building block for any terminology identification system. We also provide empirical evidence that the superiority of our approach, beyond a 10-million-word threshold, is essentially domain- and corpus-size-independent.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; I.2.7 [**Artifical Intelligence**]: Natural Language Processing—*Text analysis*; J.3.7 [**Life and Medical Sciences**]: [Medical information systems]

General Terms

Algorithms, Measurement, Performance, Experimentation

Keywords

Natural Language Processing, Terminology, Knowledge Extraction

K-CAP'05, October 2–5, 2005, Banff, Alberta, Canada.
Copyright 2005 ACM 1-59593-163-5/05/0010 ...$5.00

1. INTRODUCTION

With proliferating volumes of medical and biological text available, the need to extract and manage domain-specific terminologies has become increasingly relevant in the recent years. Most available terminological dictionaries, however, are still far from being complete, and what's worse, a constant stream of new terms enters via the ever-growing biomedical literature. Naturally, the costly and time-consuming nature of manually identifying new terminology from text calls for 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 multi-word units[1], which are, thus, much more difficult to recognize and extract. Moreover, although the need to assemble and extend technical and scientific terminolgies is currently most pressing in the biomedical domain, virtually any (sub-)field of human research/ expertise in which structured knowledge needs to be extracted from text collections calls for high-performance terminology identification methods.

2. RELATED WORK AND PURPOSE

There have been many studies examining various methods to automatically extract scientific or technical terms from domain-specific corpora, such as from biomedical ones (see, e.g., [9], [18], [4], [2], [17] and [11]). Typically, approaches to multi-word term extraction collect term candidates from texts by making use of various degrees of linguistic filtering (e.g., part-of-speech tagging, phrase chunking etc.), through which candidates of various linguistic patterns are identified (e.g. *noun-noun*, *adjective-noun-noun* combinations etc.). These possible choices are then submitted to frequency- or statistical-based evidence measures (e.g., C-value [7]) which compute weights indicating to what degree a candidate qualifies as a terminological unit. While *term mining*, as a whole, is a complex process involving sev-

[1] According to [15], more than 85% of domain-specific terms are multi-word units.

eral other components (e.g., orthographic and morphological normalization, acronym detection, conflation of term variants, term context, term clustering, etc., see [17]), the measure which assigns a *termhood value* to a term candidate is an essential building block of any term identification system.

In multi-word automatic term recognition (ATR) the C-value approach [7, 16], 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 [5], and the battery of statistical and information-theoretic measures (t-test, log-likelihood, entropy) which is typically employed for the extraction of general-language collocations (see [13, 6]). While these measures have their statistical merits in terminology identification, it is interesting to note that they make little use of linguistic properties associated with terminological units.[2] However, such properties have proven to be helpful in the identification of general-language collocations [22]. Therefore, one may wonder whether there are linguistic features which may also be beneficial to ATR. One such feature we have identified is the *limited paradigmatic modifiability* of terms, which will be described in detail in Subsection 3.3.

The purpose of our study is to present a novel term recognition measure which directly incorporates this linguistic criterion, and in evaluating it against some of the standard procedures, we show that it substantially outperforms them on the task of term extraction from the biomedical literature.

3. METHODS AND EXPERIMENTS

3.1 Construction and Statistics of the Training Set

We collected a biomedical training corpus of approximately 513,000 MEDLINE abstracts using the following MESH-terms query: *transcription factors*, *blood cells* and *human*.[3] We then annotated this 104-million-word corpus with the GENIA part-of-speech tagger[4] and identified noun phrases (NPs) with the YAMCHA-Chunker [12]. In this study, we restricted ourselves 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 terminology (and terms in general) is contained within noun phrases [10]. We filtered out a number of stop words (i.e., determiners, pronouns, measure symbols etc.) and also ignored noun phrases with coordination markers (e.g., *and*, *or* etc.).[5]

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

Table 1: Frequency distribution for noun phrase term candidate tokens and types for our 104-million-word MEDLINE text corpus

In order to obtain our 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 [16]. We thus normalized the nominal head of each noun phrase (typically the rightmost noun in English) via the full-form UMLS SPECIALIST LEXICON [3], a large repository of both general-language and domain-specific (medical) vocabulary. To eliminate noisy low-frequency data, we set different frequency cut-off thresholds c for the bigram, trigram and quadgram candidate sets and only considered candidates above these thresholds (see Table 1).

3.2 Evaluating Terminology Extraction Algorithms

(Domain-specific) terms are usually referred to as the linguistic surface manifestation of (domain-specific) concepts. Typically, terminology extraction studies evaluate the goodness of their algorithms by having their ranked output examined by so-called *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 so inter-annotator agreement is not accounted for (e.g. in the studies of [7], [4]). Furthermore, what constitutes a relevant term for a particular domain may be rather difficult to decide – even for domain experts – if all they have in front of them is a list of candidates without any further context. Thus, rather than relying on direct human judgement in identifying true pos-

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

[3] MEDLINE is a large biomedical bibliographic database (see `http://www.ncbi.nlm.nih.gov`). For information retrieval purposes, all its abstracts are indexed with a controlled indexing vocabulary, viz. MESH. Our query is aimed at the molecular biology domain, with the publication period from 1978 to 2004.

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

[5] Of 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 [16].

itives among a candidate set, a better solution may be to take already existing terminolgical resources, which have developed over years and have gone through various modifications and editions by expert committees. In this sense, the biomedical domain is an ideal test bed for evaluating the goodness of ATR algorithms because it hosts one of the most extensive and curated terminological resources, viz. the UMLS METATHESAURUS [21], and thus offers a well-established source of curated and agreed judgements about what constitutes a biomedical term.

Accordingly, for our purposes of evaluating the quality of different measures in recognizing multi-word terminology from the biomedical literature, we take every word bigram, trigram, and quadgram in our candidate sets to be a term (i.e., a true positive) if it was found in the 2004 UMLS METATHESAURUS.[6] For example, the word trigram "*long terminal repeat*" is listed as a term in one of the UMLS vocabularies, *viz.* MESH [20], 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, the number was 3,590 (11.6%), and for 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, frequency of occurrence of the term candidates either plays a major role (e.g., C-value) or at least has a significant impact concerning the degree of *termhood* assigned (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. For example, a trigram multi-word expression such as "*long terminal repeat*" contains three word/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, i.e., the tendency not to let other words appear in particular slots. To arrive at the various combinatory possibilities that fill these slots, the standard combinatory formula without repetitions can be used. For an n-gram (of size n) to select k slots (i.e., in an unordered selection) we define:

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

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

n-gram	freq	$P\text{-}Mod$ (k=1,2)
long terminal repeat	434	0.03

k slots	possible selections sel	freq	mod_{sel}
$k = 1$	k_1 terminal repeat	460	0.94
	long k_2 repeat	448	0.97
	long terminal k_3	436	0.995
			mod_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
			mod_2 =0.03

Table 2: $P\text{-}Mod$ **and** k**-modifiabilities for** $k = 1$ **and** $k = 2$ **for the trigram term** ***long terminal repeat***

n-gram	freq	$P\text{-}Mod$ (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
			mod_1 =0.06
$k = 2$	k_1 k_2 response	40143	0.06
	k_1 cell k_3	120056	0.02
	t k_2 k_3	34925	0.07
			mod_2 =0.00008

Table 3: $P\text{-}Mod$ **and** k**-modifiabilities for** $k = 1$ **and** $k = 2$ **for the trigram non-term** ***t cell response***

Now, for a particular k ($1 \leq k \leq n$; n = length of n-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

[6] We excluded those UMLS source vocabularies that were definitely not deemed relevant for molecular biology, such as nursing and health care billing codes.

[7] As 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.

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 the 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 derived as follows:

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

Then, the *paradigmatic modifiability*, $P\text{-}Mod$, of an n-gram is the product of all its k-modifiabilities:[8]

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

Comparing the trigram $P\text{-}Mod$ 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 output list rank (see Subsection 4.1 for details) assigned to both trigrams by t-test and by our $P\text{-}Mod$ measure. While "*t cell response*" has rank 24 on the t-test output list (which has to be attributed to its high frequency), $P\text{-}Mod$ puts it on the 1249th rank. Conversely, "*long terminal repeat*" is ranked on 242 by t-test, whereas it is ranked on 24 by $P\text{-}Mod$. In fact, even lower-frequency multi-word units gain a prominent ranking if they exhibit limited paradigmatic modifiability. For example, the trigram term "*porphyria cutanea tarda*" is ranked on 28 by $P\text{-}Mod$ although its frequency is only 48 (which results in rank 3291 on the t-test output list). Despite its lower frequency, this term may be judged relevant for the molecular biology domain.[9] It should be noted that the termhood values (and the corresponding list ranks) computed by $P\text{-}Mod$ also include $k = 3$ and hence take into account some frequency factor. As can be seen from the previous ranking examples, however, this factor does not override the paradigmatic modifiability factors of the lower ks.

On the other hand, $P\text{-}Mod$, of course, 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 (on 31) by t-test. $P\text{-}Mod$, however, ranks this term on 550 because the token *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 rather the exception than the rule and that terms are in fact linguistically more fixed than non-terms, which is exactly what our measure of limited paradigmatic modifiability aims at quantifying.

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 an (usually) arbitrarily set number m of the highest ranked candidates returned by a particular measure, which is usually done by a domain expert. Because this is also labor-intensive (besides being unreliable), m is usually small, ranging from 50 to several hundreds.[10] In contrast, by taking a large and already established terminology as an evaluation gold standard, we are able to dynamically examine various m-highest ranked samples, which allows for the plotting of standard precision and recall graphs for the whole candidate set. Through this, we provide a much more reliable evaluation metric for ATR measures than what is typically employed in the literature.

We evaluate our $P\text{-}Mod$ algorithm against the t-test measure,[11] which, of all standard mesures, yields the best results in general-language collocation extraction studies [6], and against the widely used C-value, which aims at enhancing the common frequency of occurrence measure by making it sensitive to nested terms [7]. Our baseline is defined by the proportion of true positives (i.e., the proportion of terms) in our bi-, tri- and quad-gram candidate sets, which is equivalent to the likelihood of finding one by blindly picking from one of the different sets (see Subsection 3.2 above).

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 found, according to the UMLS METATHESAURUS) was

[8] Setting the upper limit of k to n (which would be $n = 3$ 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.

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

[10] Studies on collocation extraction (e.g. [6]) also point out the inadequacy of such evaluation methods claiming they usually lead to very superficial judgements about the measures to be examined.

[11] See [13] for a description how this measure can be used for the extraction of multi-word expressions.

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 ranked list considered	**Precision** scores of measures		
		P-Mod	t-test	C-value
Bigrams	1%	0.82	0.62	0.62
	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
Trigrams	1%	0.62	0.55	0.54
	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
Quadgrams	1%	0.43	0.50	0.50
	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 of termhood. As can be clearly seen, however, our *P-Mod* algorithm 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) the precision value for *P-Mod* is 20 points higher than for t-test and for 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), 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 (bi-, tri-, and quadgram) ranked lists considered, the precision curves start to converge toward the baseline, but *P-Mod* maintains a steady advantage.

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, according to the UMLS METATHESAURUS) is identified by a particular measure at a certain point of the ranked list. In this sense, recall is an even better indicator of a particular measure's performance.

Again, our linguistically motivated terminology extraction algorithm outperforms all others, 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 t-test and C-value need to winnow 35% and 37%, respectively. For trigrams and quad-

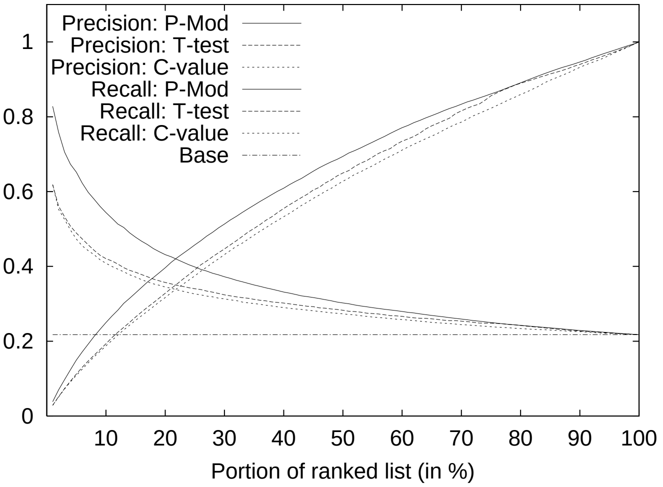

Figure 1: Precision/Recall for Bigram Biomedical Term Extraction

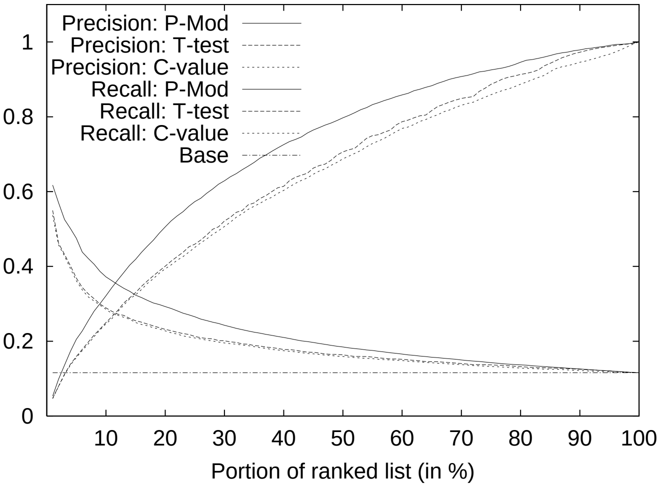

Figure 2: Precision/Recall for Trigram Biomedical Term Extraction

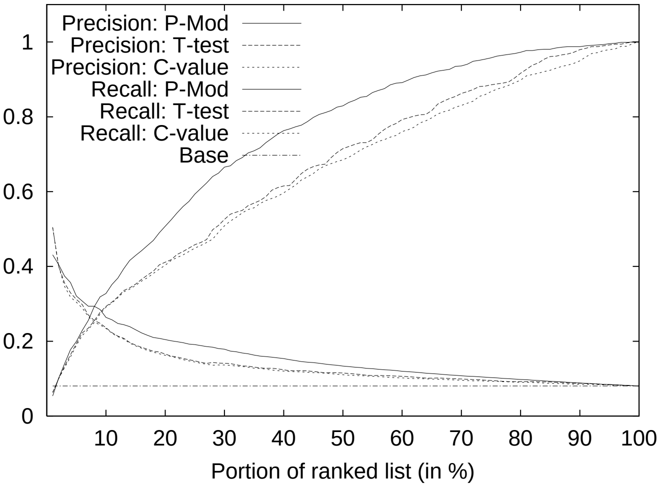

Figure 3: Precision/Recall for Quadgram Biomedical Term Extraction

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

Table 5: **Portions of the Ranked List to consider to obtain Selected Recall Scores for Biomedical Term Extraction**

grams, $P\text{-}Mod$ only needs to examine 19% and 20% of the list, whereas the other two measures need 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, $P\text{-}Mod$ only needs to winnow 27% of its output list while t-test and C-value need to analyze 38% and 40%, respectively. To get 0.7 recall, $P\text{-}Mod$ only needs to analyze 36%, and the second-placed t-test already 50% of the ranked list. For a 0.8 recall, this relation is 50% ($P\text{-}Mod$) to 63% (t-test), and at recall point 0.9, 68% ($P\text{-}Mod$) to 77% (t-test). For quadgram term identification, the results for $P\text{-}Mod$ 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.

# of measure points	# of significant differences comparing *P-Mod* with					
	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		quadgrams	

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

We also tested the significance of differences for our results, both between $P\text{-}Mod$ and t-test and between $P\text{-}Mod$ and C-value. Because in all cases the ranked lists were taken from the same set of candidates (*viz.* the set of bigram candidate types, the set of trigram candidate types, or the set of quadgram candidate types), and hence constitute dependent samples, we applied the McNemar test [19] 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 $P\text{-}Mod$ and C-value, all of them are significant, and between $P\text{-}Mod$ and t-test, all are significantly different up to measure point 70.[12] Looking at the tri- and quadgrams, although the number of significant differences is less than for bigrams, the vast majority of measure points still is significantly different and thus underlines the superiour performance of the $P\text{-}Mod$ measure.

4.2 Domain Independence and Corpus Size

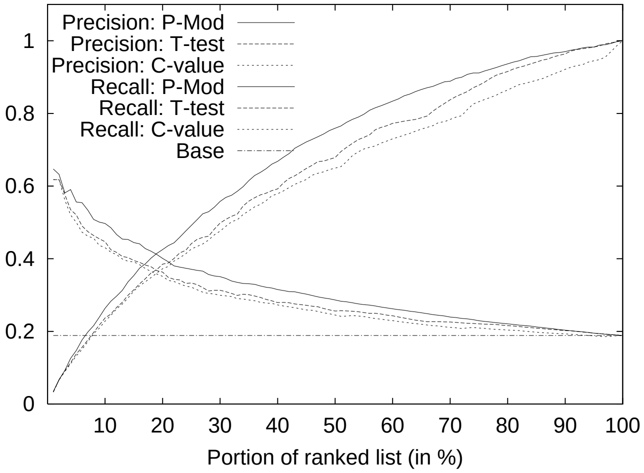

Figure 4: **Precision/Recall for Trigram Biomedical Term Extraction on 10-million-word Corpus (cutoff $c \geq 4$ with 6,760 term candidate types)**

It could be reasonably well argued that the results reported above are mainly due to corpus size. Indeed, the corpus employed in our study is rather large (104 million words) because the molecular biology domain offers a lot of free-text literature via the Medline bibliographic database. Other domains (e.g. clinical narratives, various engineering domains) or even more specialized subdomains (e.g. plant biology) do not offer such a wealth of free-text material and, therefore, terminology mining would have to make do with smaller-sized corpora. In order to test the effect that a drastically reduced corpus size would have, we ran the terminology extraction

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

methods for trigrams on a much smaller-sized subset of our original corpus, viz. on a subset of 10 million words. The results are visualized in Figure 4 above.

As can be seen, our *P-Mod* extraction algorithm still clearly outperforms the other ones on the 10-million-word corpus, both in terms of precision and recall. Examining whether the differences are statistically significant, we again applied the two-tailed McNemar test on 100 selected measure points (see Table 7). Comparing *P-Mod* with t-test, most significant differences can be observed between measure points 20 and 80, with almost 80% to 90% of the points being significantly different. These significant differences are even more pronounced when comparing the results between *P-Mod* and C-value.

# of measure points	# of significant differences comparing *P-Mod* with	
	t-test	C-val
10	1	4
20	11	14
30	22	25
40	31	34
50	41	44
60	51	54
70	61	64
80	71	74
90	78	84
100	78	93
	trigrams	

Table 7: Significance testing of differences for trigrams on a 10-million-word corpus using the two-tailed McNemar test at 95% confidence interval

5. CONCLUSIONS

In our study, we proposed a new terminology identification algorithm and showed that it substantially outperforms some of the standard measures in distinguishing terms from non-terms in the biomedical literature. While mining technical and scientific literature for new terminological units and assembling those in controlled vocabularies is an overall complex task involving several components, one essential building block is a measure indicating the *degree of termhood* of a candidate. In this respect, our study has shown that an algorithm which incorporates a vital linguistic property of terms, *viz.* their *limited paradigmatic modifiability*, can be a much more powerful and valuable part of a terminology extraction system (like, e.g., proposed by [14]) than the standard measures that are typically employed. This is in line with our previous work on general-language collocation extraction [22] in which we showed that a linguistically motivated algorithm based on the limited syntagmatic modifiability of collocations outperforms several of the standard association measures. Furthermore, we also showed that *P-Mod* is superiour to the other term extraction algorithms regardless of corpus size. This is particularly important considering the fact that there are domains in which not such a wealth of free-text material is available and in which terminology mining thus may be restricted to smaller corpus sizes. Consequently, we may also conclude that, although our methodology has been tested on the biomedical domain, there are essentially no domain-specific restrictions to it.

In general, a high-performing term identification system is not only valuable for collecting new terms per se but is also essential in updating already existing terminology resources. As a concrete example, the term "*cell cycle*" is contained in the hierarchically-structured biomedical MeSH terminology and the term "*cell cycle arrest protein BUB2*" in the MeSH supplementary concept records which include many proteins with a GenBank[1][13] identifier. The word trigram *cell cycle arrest*, however, is not included in MeSH although it is ranked in the top 10% of *P-Mod*. Utilizing this prominent ranking, the missing semantic link can be established between these two terms (i.e., between *cell cycle* and *cell cycle arrest protein BUB2*), both by including the trigram *cell cycle arrest* in the MeSH hierarchy and by linking it via the comprehensive terminological umbrella system for biomedicine, *viz.* Umls, to the Gene Ontology (GO [8]), in which it is listed as a stand-alone term.

6. REFERENCES

[1] D. A. Benson, M. S. Boguski, D. J. Lipman, J. Ostell, F. B. Ouellette, B. A. Rapp, and D. L. Wheeler. Genbank. *Nucleic Acids Research*, 27(1):12–17, 1999.

[2] O. Bodenreider, T. C. Rindflesch, and A. Burgun. Unsupervised, corpus-based method for extending a biomedical terminology. In *Proceedings of the ACL Workshop on Natural Language Processing in the Biomedical Domain.*, pages 53–60. Pittsburgh, PA, USA. Association for Computational Linguistics, 2002.

[3] A. C. Browne, G. Divita, V. Nguyen, and V. C. Cheng. Modular text processing system based on the Specialist lexicon and lexical tools. In C. G. Chute, editor, *AMIA'98 – Proceedings of the 1998 AMIA Annual Fall Symposium. A Paradigm Shift in Health Care Information Systems: Clinical Infrastructures for the 21st Century*, page 982. Orlando, FL, November 7-11, 1998. Philadelphia, PA: Hanley & Belfus, 1998.

[4] N. Collier, C. Nobata, and J. Tsujii. Automatic acquisition and classification of terminology using a tagged corpus in the molecular biology domain. *Terminology*, 7(2):239–257, 2002.

[13] GenBank is a database containing an annotated collection of all publicly available DNA sequences.

[5] F. J. Damerau. Generating and evaluating domain-oriented multi-word terms from text. *Information Processing & Management*, 29(4):433–447, 1993.

[6] S. Evert and B. Krenn. Methods for the qualitative evaluation of lexical association measures. In *ACL'01 – Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 188–195. Toulouse, France, 2001.

[7] K. Frantzi, S. Ananiadou, and H. Mima. Automatic recognition of multi-word-terms: the C/NC value method. *International Journal of Digital Libraries*, 3(2):115–130, 2000.

[8] Gene Ontology Consortium. Creating the Gene Ontology resource: Design and implementation. *Genome Research*, 11(8):1425–1433, 2001.

[9] W. R. Hersh, E. Campbell, D. Evans, and N. Brownlow. Empirical, automated vocabulary discovery using large text corpora and advanced natural language processing tools. In J. J. Cimino, editor, *AMIA'96 – Proceedings of the 1996 AMIA Annual Fall Symposium (formerly SCAMC). Beyond the Superhighway: Exploiting the Internet with Medical Informatics*, pages 159–163. Washington, D.C., October 26-30, 1996. Philadelphia, PA: Hanley & Belfus, 1996.

[10] J. S. Justeson and S. M. Katz. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1(1):9–27, 1995.

[11] M. Krauthammer and G. Nenadić. Term idenfification in the biomedical literature. *Journal of Biomedical Informatics*, 37(6):512–526, 2004.

[12] T. Kudo and Y. Matsumoto. 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, 2001.

[13] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. Cambridge, MA; London, U.K.: Bradford Book & MIT Press, 1999.

[14] H. Mima, S. Ananiadou, and G. Nenadić. The Atract workbench: Automatic term recognition and clustering of terms. In V. Matusek, editor, *Text, Speech and Dialog (TSD 2001)*, volume 2166 of *Lecture Notes in Artificial Intelligence*, pages 126–133. Berlin: Springer, 2001.

[15] H. Nakagawa and T. Mori. Nested collocation and compound noun for term recognition. In *COMPUTERM '98 – Proceedings of the First Workshop on Comutational Terminology*, pages 64–70, 1998.

[16] G. Nenadić, S. Ananiadou, and J. McNaught. Enhancing automatic term recognition through recognition of variation. In *COLING 2004 – Proceedings of the 20th International Conference on Computational Linguistics*, pages 604–610. Association for Computational Linguistics, 2004.

[17] G. Nenadić, I. Spasic, and S. Ananiadou. Terminology-driven mining of biomedical literature. *Journal of Biomedical Informatics*, 33:1–6, 2003.

[18] T. C. Rindflesch, L. Hunter, and A. R. Aronson. Mining molecular binding terminology from biomedical text. In N. M. Lorenzi, editor, *AMIA'99 – Proceedings of the 1999 Annual Symposium of the American Medical Informatics Association. Transforming Health Care through Informatics: Cornerstones for a New Information Management Paradigm*, pages 127–131. Washington, D.C., November 6-10, 1999. Philadelphia, PA: Hanley & Belfus, 1999.

[19] L. Sachs. *Applied Statistics: A Handbook of Techniques*. New York: Springer, 2nd edition, 1984.

[20] MeSH. *Medical Subject Headings*. Bethesda, MD: National Library of Medicine, 2004.

[21] Umls. *Unified Medical Language System*. Bethesda, MD: National Library of Medicine, 2004.

[22] J. Wermter and U. Hahn. Collocation extraction based on modifiability statistics. In *COLING Geneva 2004 – Proceedings of the 20th International Conference on Computational Linguistics*, volume 2, pages 980–986. Geneva, Switzerland, August 23-27, 2004. Association for Computational Linguistics, 2004.