Effective Grading of Termhood in Biomedical Literature

Joachim Wermter Udo Hahn
Jena University Language and Information Engineering (JULIE) Lab
http://www.coling.uni-jena.de

Abstract

The ever-increasing amount of textual information in biomedicine calls for effective procedures for automatic terminology extraction which assist biomedical researchers and professionals in gathering and organizing terminological knowledge encoded in text documents. In this study, we propose a new, linguistically grounded measure for automatically identifying multi-word terms from the biomedical literature. Our approach is based on the limited paradigmatic modifiability of terms and is tested on bigram, trigram and quadgram noun phrases extracted from a 104-million-word text corpus comprised of MEDLINE *abstracts. Using the* UMLS METATHESAURUS *as a gold standard, we show that our algorithm substantially outperforms the standard term identification measures and, therefore, qualifies as a high-performing building block for any biomedical terminology mining system.*

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. Thus, there have been many studies examining various methods to automatically extract terms from domain-specific corpora, such as from medical and biological ones (see, e.g., [1], [2] and [3]). Whereas the recognition of single-word terms usually does not pose any particular challenges, the vast majority of biomedical terms typically consists of multi-word units[1] and are, thus, much more difficult to recognize and extract. Typically, approaches to multi-word term extraction collect term candidates from domain-specific literature 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 candidates are then submitted to frequency- or statistical-based evidence measures (e.g., C-value [5]) which compute weights indicating to what degree a candidate qualifies as a terminological unit. While biomedical *term mining*, as a whole, is a complex process involving several other components (e.g., orthographic and inflectional normalization, term variant conflation, acronym detection, term context, term clustering, etc. [6, 5]), the measure which assigns such a termhood value is an essential building block of any term identification system.

In multi-word automatic term recognition (ATR) the C-value approach [5], which aims at improving the extraction of nested terms, has been one of the most widely used techniques in recent years. Other association measures are mutual information (MI) [7], 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 [8, 9]). 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. However, such properties have proven to be helpful in the identification of general-language collocations [10]. 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 the subsequent section.

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.

Methods and Experiments

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*

[1] According to [4], more than 85% of domain-specific terms are multi-word units.

cells and *human*.[2] We then annotated this 104-million-word corpus with the GENIA part-of-speech tagger[3] and identified noun phrases (NPs) with the YAMCHA-Chunker [11]. 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 biomedical terminology (and terms in general) is contained within noun phrases. 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.).

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 term candidate tokens (= any given instance of an NP) and types (= each unique NP) 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). We also morphologically normalized the nominal head of each noun phrase (typically the rightmost noun in English) via the full-form UMLS SPECIALIST LEXICON [12]. 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.

Biomedical Terminology

Terms are usually referred to as the linguistic surface manifestation of concepts. 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.[4] For example, the word trigram "*long terminal repeat*" is listed as a term in one of the UMLS vocabularies, *viz.* MESH [13], whereas "*t cell response*" is not. Thus, among the 67,308 word bigram candidate types, 14,650 (21.8%) true terms were identified; among the 31,017 word trigram types, the number was 3,590 (11.6%), and for the 10,838 word quadgram types, 873 (8.1%) true terms were identified.[5]

Paradigmatic Modifiability of Terms

For most standard association measures used for terminology extraction, (normalized) 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, occurrence frequency 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 UMLS-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:

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

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. Now, for a particular k ($1 \leq k \leq n$; n = length of n-gram), the frequency of each possible selection, sel,

[2] Our query is aimed at the molecular biology domain, with the publication period from 1978 to 2004.

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

[4] Exluding those UMLS vocabularies not relevant for molecular biology, e.g. nursing and health care billing codes.

[5] 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.

n-gram	freq	P-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-Mod and k-modifi abilities for $k=1$ and $k=2$ for the trigram term *long terminal repeat*

n-gram	freq	P-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-Mod and k-modifi abilities for $k=1$ and $k=2$ for the trigram non-term *t cell response*

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 score; 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-Mod, of an n-gram is the product of all its k-modifiabilities:[6]

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

[6]Setting the upper limit of k to n (which would be $n=3$ for trigrams) has the pleasant side effect of including frequency in our P-Mod measure: 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.

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 output list rank (see the *Results* section below for details) assigned to both trigrams by t-test and by our P-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-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-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-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.[7]

Methods of Evaluation

Standard procedures for evaluating the qualitiy 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 rather labor-intensive (besides being unreliable and superficial), m is usually small, ranging from 50 to several hundreds. In contrast, we examine increasing m-highest ranked samples, which allows for the plotting of standard precision and recall graphs for the whole candidate set.

We evaluate our P-Mod measure against the widely used C-value and also against the t-test measure, which, of all standard mesures (such as mutual information or log-likelihood), yields the best results in collocation extraction studies [9]. Our baseline is defined by the proportion of true positives (i.e., the proportion of terms) in our bi-, tri- and quadgram candidate sets, which is equivalent to the likelihood of finding one by blindly picking from one of the different sets (see the subsection on *Biomedical Terminology* above).

Results and Discussion

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 considered. The precision values for the various portions were computed such that for each percent point of the list,

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

the number of true terms found 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 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
	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	0.22
	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
	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* 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 quadgrams, P-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 or-

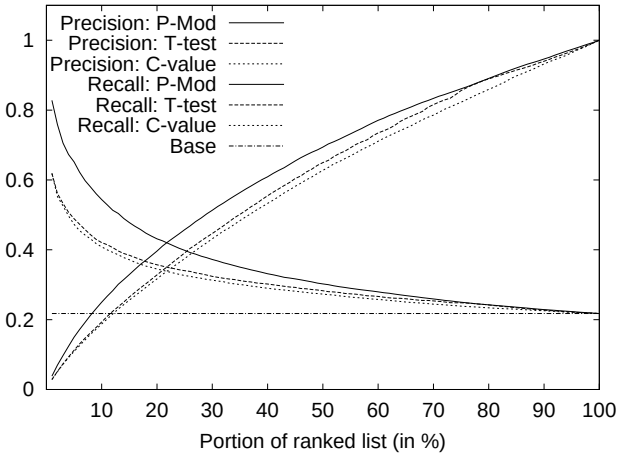

Figure 1: Precision/Recall for Bigram Biomedical Term Extraction

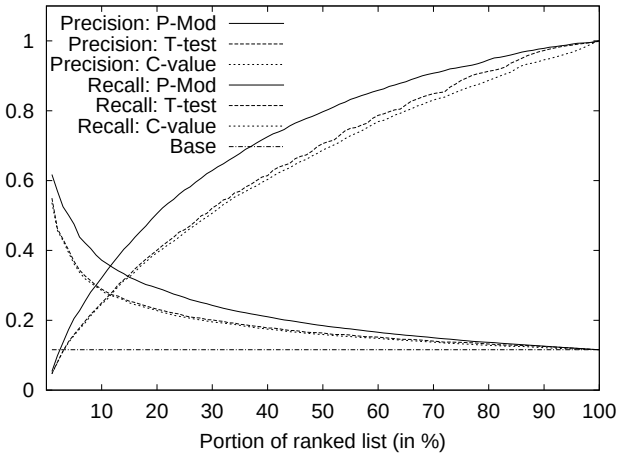

Figure 2: Precision/Recall for Trigram Biomedical Term Extraction

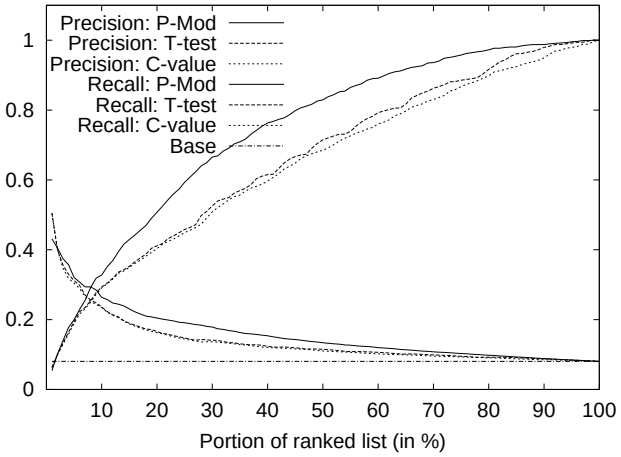

Figure 3: Precision/Recall for Quadgram Biomedical Term Extraction

	Recall scores of measures	Portion of Ranked List		
		P-Mod	t-test	C-value
Bigrams	0.5	29%	35%	37%
	0.7	51%	56%	59%
	0.8	65%	69%	72%
	0.9	82%	83%	85%
Trigrams	0.5	19%	28%	29%
	0.7	36%	50%	52%
	0.8	50%	63%	65%
	0.9	68%	77%	82%
Quadgrams	0.5	20%	28%	30%
	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

der to obtain a 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 get 0.7 recall for trigrams, P-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-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.

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 biomedical text 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 typically employed.

In general, a high-performing biomedical 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 MESH and the term "*cell cycle arrest protein BUB2*" in the MESH supplementary concept records which include many proteins with a GENBANK[15] 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 UMLS to the Gene Ontology (GO [16]), in which it is listed as a stand-alone term.

References

[1] Rindflesh TC, Hunter L, Aronson AR. Mining molecular binding terminology from biomedical text. In: AMIA '99; 1999. p. 127–131.

[2] Nenadić G, Spasic I, Ananiadou S. Terminology-Driven Mining of Biomedical Literature. Journal of Biomedical Informatics 2003;13:1–6.

[3] Krauthammer M, Nenadić G. Term Identification in the Biomedical literature. Journal of Biomedical Informatics 2004;37(6):512–526.

[4] Nakagawa H, Mori T. Nested collocation and compound noun for term recognition. In: COMPUTERM '98 – Proc of the First Workshop on Computational Terminology; 1998. p. 64–70.

[5] Frantzi K, Ananiadou S, Mima H. Automatic Recognition of Multi-Word-Terms: the C/NC value method. Int'l Journal of Digital Libraries 2000;3(2):115–130.

[6] Nenadić G, Ananiadou S, McNaught J. Enhancing automatic term recognition through recognition of variation. In: COLING '04 – Proc of the 20th Int'l Conf on Comp Ling; 2004. p. 604–610.

[7] FJ Damerau. Generating and evaluating domain-oriented multi-word terms from text. Information Processing & Management 1993;29(4):433–447.

[8] Manning CD, Schütze H. Foundations of Statistical Natural Language Processing. Cambridge, MA; London, U.K.: Bradford Book & MIT Press; 1999.

[9] Evert S, Krenn B. Methods for the Qualitative Evaluation of Lexical Association Measures. In: ACL'01 – Proc of the 39th Annual Meeting of the Assoc for Comp Ling; 2001. p. 188–195.

[10] Wermter J, Hahn U. Collocation extraction based on modifiability statistics. In: COLING '04 – Proc of the 20th Int'l Conf on Comp Ling; 2004. p. 980–986.

[11] Kudo T, Matsumoto Y. Chunking with support vector machines. In: NAACL'01 – Proc of the 2nd Meeting of the North American Assoc for Comp Ling. Pittsburgh, PA, USA, June 2-7, 2001; 2001. p. 192–199.

[12] UMLS. Unified Medical Language System. Bethesda, MD: National Library of Medicine; 2004.

[13] MESH. Medical Subject Headings. Bethesda, MD: National Library of Medicine; 2004.

[14] Mima H, Ananiadou S, G Nenadić. The ATRACT Workbench: Automatic Term Recognition and Clustering of Terms. In: Matusek V, editor. Text, Speech and Dialog (TSD 2001). vol. 2166. Berlin: Springer; 2001. p. 126–133.

[15] Benson DA, et al. GenBank. Nucleic Acids Research 1999;27(1):12–17.

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