Mining Terminological Knowledge in Large Biomedical Corpora

H. Liu, C. Friedman

Pacific Symposium on Biocomputing 8:415-426(2003)

MINING TERMINOLOGICAL KNOWLEDGE IN LARGE BIOMEDICAL CORPORA

HONGFANG LIU, CAROL FRIEDMAN
Department of Medical Informatics, Columbia University
NEW YORK, NY 10032

Terminological knowledge of the biomedical domain is important for natural language processing (NLP) and information retrieval (IR) applications, and a number of terminological knowledge sources, such as LocusLink, GeneBank, and the UMLS, already exist. However, because of the tremendous amount of research activity in the field, new terms and symbols are continually being created, many of which are published in the literature, but are not available in any of the other resources. Therefore, effective mining of the literature for new terminology is critical for furthering NLP and IR applications. Abbreviations are widely used in the biomedical domain, and the understanding of abbreviations requires a terminological knowledge base that consists of abbreviations with their associated senses. In previous work, several methods have been developed for automatic construction of abbreviation knowledge bases from parenthetical expressions. However, these methods pair abbreviations and their expansions based on manually crafted patterns or rules. In this paper, we propose an automatic method, which is not based on patterns or rules but is based on the use of collocations, to extract a set of related terms from parenthetical expressions including abbreviations associated with their expansions and other types of related terms such as synonyms, or hyponyms etc. Our method is based on the observation that terms associated with parenthetical expressions i) are usually related, and ii) are often collocations because they tend to co-occur more often than expected by chance. Our method was applied to the collection of MEDLINE abstracts. The method and the results were evaluated using two collections: Berman's handcrafted abbreviation list and the LocusLink collection.

1 Introduction

In recent years, there has been a growing interest in automatic methods that construct and manage terminology resources for natural language processing (NLP) applications using large online collections of documents [1-4]. Although it is feasible to construct and manage terminologies manually in very limited domains, automatic or at least semi-automatic methods are required for applications that apply to domains with rich genres of context [5;6]. The domain we consider for this study is biomedical literature. We seek to explore the use of parenthetical expressions in text for automatic acquisition of terminological knowledge for NLP applications that are applied to biomedical text. Specifically, we consider two tasks: i) to find expansions for abbreviations, and ii) to find other types of semantically related terms.

It has been known that the wide use of abbreviations in the biomedical domain affects NLP applications, such as information retrieval systems or information extraction systems [6-8]. In order for NLP applications to process and interpret abbreviations that are not defined in documents, the associated terminology should include abbreviations together with their corresponding expansions in that domain. In addition, terminologies are more useful for NLP purposes if they not only list

single-word or multi-word terms but also provide semantic knowledge, such as semantic categories and various other kinds of semantic relations (such as hypernymy/hyponymy, or synonymy) [9].

Observing that terms associated with parenthetical expressions are usually related terms, and expansions together with their abbreviations are often collocations (i.e., they tend to co-occur more often than expected by chance), we propose an automatic method for acquiring terminological knowledge from parenthetical expressions. In the following, we first introduce background material and related work for this study. We then describe our method in detail. The acquisition of terminological knowledge from the 2002 version of the MEDLINE abstracts, the evaluation of the method, and the results are presented next. We then discuss the results. Finally, we point out future directions of this work and conclude.

2 Background and Related Work

In this study, the uses of parentheses that we consider are parenthetical expressions "B (**A**)" where there is a space separating B from **A**[1]. **A** is the complete text string inside the parentheses (called the inner-text) and B is the text string to the left of the parentheses (called the outer-text) within a certain window size. In the biomedical literature domain, parenthetical expressions are popularly used to define abbreviations as in "estrogen receptor (**ER**)" or as in "GABA (**gamma-aminobutyric acid**)". They are also used to specify semantic relations such as synonymy as in "natural toxin (i.e., **aflatoxin**)" or hypernymy as in "an inactive H-Ras protein (**RasN17**)". Parenthetical expressions can also be citations as in "here by using a recently developed ultrasensitive HPLC technique (**Sakhi et al. J. Chromatogr. A 828:451-460, 1998**)" or measures as in "CGRP failed to inhibit glucose-stimulated (**16.7 mM**)", etc.

There have been studies that report on the automatic construction of abbreviation terminologies using parenthetical expressions. Hisamitsu and Niwa [10] identified technical terms using parenthetical expressions that were statistically significant and then applied a set of simple rules to identify whether the text string inside parentheses was an abbreviation for a phrase that was at the left side of parentheses. Oh et al.[7] proposed a statistics-based model for constructing technical terminology by selecting similar phrase pairs "A (B)" including abbreviation pairs or translation pairs from parentheses, where a phrase pair was considered as an abbreviation pair if a half of the uppercase letters in A appeared in B sequentially. The method of Yoshida et al. [11] first identified terms representing biological substances using the PROPER system, and then extracted abbreviations using parenthetical expressions from terms identified by the PROPER system. Pustejovsky et al. [12] developed a system called ACROMED that applied a restricted pattern for

[1] Uses of parentheses without a space between B and **A** usually imply that **A** and B exist as a whole and they have no semantic relations with each other, for example, uses of parentheses in chemical names, e.g., Ca(OH)2.

identifying expansions for abbreviations from parentheses. Yu et al. [13] used a set of patterns to extract expansions for abbreviations from parenthetical expressions in full articles. Almost all above studies reported a precision of around 97% when matching abbreviations to their expansions.

However, all above studies used manually crafted patterns or rules for identification of expansions for abbreviations. Manually crafted patterns or rules are limited and often incomplete. For example, all above studies consider that alphabetic letters in an abbreviation occur in the corresponding expansions. Those methods may miss an abbreviation pair (***1H MRS***, *proton magnetic resonance spectroscopy*), or incorrectly identify the expansion as *magnetic resonance spectroscopy* from the sentence "*we used proton magnetic resonance spectroscopy (1H MRS) here*". In this study, we propose an automatic method to associate abbreviations with expansions for the purpose of automatic acquisition of terminological knowledge. Our method does not require patterns or rules, and is based on collocations. In addition to pairs of abbreviations and expansions, our method extracts other types of related terms from parenthetical expressions. The following provides background information about collocations as well as background information for resources used in this study.

Collocations considered in this paper are cohesive lexical clusters according to Smadja [14], where a collocation is a set of words such that the presence of one or several words of the set often implies or suggests the rest of the collocation. Parenthetical expressions, especially those used for defining abbreviations, are collocations. For example, given a parenthetical expression "*congestive heart failure* (*CHF*) ", the presence of words *CHF*, *heart*, and *failure* implies the presence of *congestive*. There are several methods to select collocations from text including simple frequency-based methods (such as eliminating all collocations with a frequency of less than a threshold) as well as complicated methods such as hypothesis-testing methods or mutual information methods [2;14]. Our method of selecting collocations is a complex frequency-based method.

We used the MEDLINE free-text collection for the experiment. MEDLINE [15] is the NLM bibliographic database that contains over 11 million references to journal articles in life sciences with a concentration in biomedicine. Each entry contains a unique MEDLINE identifier and citation information for the corresponding journal article, and often an abstract.

Our method, which is described in the Methods section, is based on word normalization that utilizes the SPECIALIST Lexicon, which is a UMLS Knowledge Source developed by the National Library of Medicine (NLM) as part of the Unified Medical Language System project [16]. The SPECIALIST Lexicon is a general English lexicon that includes a comprehensive set of biomedical terms. The lexical entry for each word or term provides syntactic information such as part of speech information (e.g., *acid* is a noun), and morphological information that maps textual variants to base forms (e.g., *discharging* and *discharged* to *discharge*). Additionally, the SPECIALIST Lexicon includes an abbreviation table which includes 11,051 pairs.

Two collections of abbreviations associated with their expansions were used as a gold standard. The first collection was a list of 12,098 pathology-related abbreviations that were manually collected by Berman [17]. The second collection consists of pairs of gene symbols together with their definitions. The list was extracted from LocusLink, which was developed by the NCBI of the NLM [18]. We excluded pairs where definitions contain the corresponding symbol as a sub-string with boundaries (e.g., *KIAA0042* vs *KIAA0042 gene product*). Additionally, there were 11,516 pairs that belong to RIKEN cDNA genes (i.e., *5730583K22Rik* vs *RIKEN cDNA 5730583K22*) that were excluded in this study in order to avoid biases in measures because they were formed using a simple pattern but contained almost 20% of the total entries. The resulting collection consisted of 42,875 unique pairs. After normalization the total collection obtained from the two sets amounted to 49,536 unique pairs.

3 Methods

The method contains three steps. The first step, COLLECT, collects parenthetical expressions from a large collection of text and filters out certain expressions (e.g., citations and measures) since they are not helpful for acquiring terminological knowledge. The second step, DETECT, uses the results of the first step to derive a set of pair-wise terms. In the third step, SEPARATE, we assess the set of pair-wise terms and separate them into two sets: a set of (abbreviation, expansion) pairs and a set of other types of related terms such as synonyms and hyponyms, etc.

3.1 COLLECT

We collect all uses of parenthetical expressions "B (**A**)" from sentences in a large collection of text (we use a heuristic to determine sentence boundaries and the window size for extraction is generally twice the length of A). Additionally, we collect all uses of parenthetical expressions within nested parentheses and all uses of square brackets because authors use them to avoid nested parentheses. After each successful extraction, the text string inside the parentheses is deleted and the resulting string is used for subsequent extraction. For example, we extract four uses of parenthetical expressions (i.e., *the substitution of asparagine for threonine at position 1405* ***(T1405N),*** *… by a functional polymorphism* (***the substitution of asparagine for threonine at position 1405***), *…the current produced by the Na(+)-Ca(2+) exchanger* (***I(NCX)***), and *the release of* endothelin (**ET**)) from the following two sentences a) and b).

a. … by a functional polymorphism (the substitution of asparagine for threonine at position 1405 **[**T1405N**]**) in ….
b. …the current produced by the Na(+)-Ca(2+) exchanger (I(NCX)) working in … the possible autocrine role played by the release of endothelin (ET) in …

Parenthetical expressions where the inner-texts occur only once in the corpus are filtered out since most likely they occur together by chance. In addition, we filter

out certain expressions such as citations or measures since they are useless for the automatic acquisition of terminological knowledge. The filtering process is achieved using several heuristics that are consistent with patterns for citations and measures. For example, *here by using a recently developed ultrasensitive HPLC technique* (***Sakhi et al. J. Chromatogr. A 828:451-460, 1998***) and *CGRP failed to inhibit glucose-stimulated* (***16.7 mM***) are filtered out since the former one is a citation and the latter is a measure.

3.2 DETECT

After we have a collection of parenthetical expressions, we need to detect collocations from all outer-text strings of expressions that share the same inner-text. This step contains several components: a normalization module, a collocation generator, and a collocation selector. Table 1 shows an example of the overall process, where the input to the process consists of all unique outer-text strings (e.g., *treatment of community acquired pneumonia*, …, *hospitalized community acquired pneumonia*) that correspond to the same inner-text (i.e., ***CAP***) and their frequency. The output of the overall process is a set of pair-wise terms (e.g. CAP, community acquired pneumonia)

The purpose of normalization is to unify textual variants since they usually represent the same sense in terminologies. The normalization module changes an outer-text string into lower case, removes all non-letter characters and a small set of stop words. It then normalizes each word in the text string by transforming it to the base form in the SPECIALIST Lexicon, if applicable, and represents the normalized words as an array. For example, the output for an outer-text string for CAP, *Patients with community acquired pneumonia*, is an array (*patient, community, acquired, pneumonia*). Note that for simplicity, we use the last base form listed in the SPECIALIST Lexicon if there are multiple entries for the corresponding word (e.g., *acquired* has two base forms: *acquire* when it occurs as a past particle, and *acquired* when it occurs as an adjective); otherwise, a syntax parser would be needed in order to choose the most likely one.

The collocation generator generates candidate collocations associated with frequency information. For an array that contains l words, we generate l potential collocations $\{pc_j^l$, j ranges from 1 to $l\}$, where pc_j^l is formed by concatenating the last l-j+1words in the array. For example, the above outer-text string for *CAP* after normalization generates four potential collocations (i.e., "*pneumonia*", "*acquired pneumonia*", "*community acquired pneumonia*", "*patient community acquired pneumonia*"). The number of occurrences for each potential collocation is then counted. For the sake of saving computational resources, we eliminate all potential collocations that occur only once since these words most likely occur together by chance (note that for parenthetical expressions that are not used for capturing abbreviations, this statement may not be true). The last module selects a set of collocations based on frequency. The collocation selector selects a set of collocations based on frequency information. There are two main processing phases

Step	Examples	FREQ
Outer-texts for CAP	treatment of community acquired pneumonia Patients with community acquired pneumonia patients with community acquired pneumonia group of patients with community acquired pneumonia blood culture of community- acquired pneumonia hospitalized community acquired pneumonia including pneumonia	14 10 3 2 2 1 1
Normalization	(treatment, community, acquired, pneumonia) (14) (patient, community, acquired, pneumonia) (13) (group, patient, community, acquired, pneumonia) (2) (blood, culture, community, acquired, pneumonia) (2) (hospitalize, community, acquired, pneumonia) (1) (include, pneumonia)	14 13 2 2 1 1
Potential Collocations	pneumonia acquired pneumonia include pneumonia community acquired pneumonia treatment community acquired pneumonia patient community acquired pneumonia culture community acquired pneumonia hospitalize community acquired pneumonia blood culture community acquired pneumonia group patient community acquired pneumonia	33 32 1 32 14 15 2 1* 2 2
After Eliminating	pneumonia acquired pneumonia community acquired pneumonia	33 32 32
After Subsuming	community acquired pneumonia	32
Final Collocation	(CAP, community acquired pneumonia)	32

Table 1: An example illustrating the process of detecting collocations from all outer-texts that share the same inner-text CAP.

for the selection: eliminating, and subsuming, as shown in Figure 1. Both phases contain a loop associated with the number of words (LEN, ranges from 2 to the maximum number of words in a potential collocation, i.e., MAX) in a potential collocation. Let pc be a collocation with LEN words, pc' be a collocation formed from the last LEN-1 words in pc, and PC(pc) be the set of potential collocations formed by adding a prefix word to pc. The elimination of potential collocations is achieved using two formulas a and b, which are described as following:

Formula a) states that if the number of elements in PC(pc) is larger than a certain threshold t_0, we consider these prefix words to occur with pc by chance, and therefore all elements in PC(pc) are eliminated from the final set. For example in Table 1, when LEN = 3, pc = "*community acquired pneumonia*", t_0 =3, all potential collocations containing pc and having more than three words are eliminated from the final set.

Selecting:
FOR LEN = 2 to MAX {
FOR each *pc* that is a potential collocation with LEN words {
Let pc' be the potential collocation formed by the last LEN-1 words of pc
IF(a). $|PC(pc)| > t_0$ {
Delete all potential collocations from **PC** where the last LEN words are the same as *pc*
}
IF (b). $\frac{freq(pc)}{freq(pc')} < t_1$ {
Delete *pc* and all potential collocations from **PC** where the last LEN words are the same as *pc*
}
}
}
Subsuming:
FOR LEN = 2 to MAX {
FOR each *pc* that is a potential collocation with LEN words {
Let *sc* be the summation of the frequency of all collocations formed by adding one more word to the left of *pc*
IF (c). $\frac{sc}{freq(pc)} > t_2$ {
Delete *pc* from **PC**
}
}
}
RETURN **PC**

Figure 1: The process of selecting collocations.

Formula b) states that if the ratio of the frequency of pc compared to the frequency of pc' is less than a certain threshold t_1, we consider it to be relatively less frequent, and eliminate it together with potential collocations that have it as the postfix string. For example in Table 1, when LEN = 2, pc = "*include pneumonia*", and t_1 = 0.1, the string "*include pneumonia*" is eliminated since the ratio of the frequency of "*include pneumonia*" to the frequency of "*pneumonia*", i.e., 1/35, is less than 0.1.

Let sc be the summation of the frequency of all collocations in PC(pc). The subsuming process is achieved using the formula c in Figure 1, i.e., if the ratio of sc to the frequency of pc is larger than certain threshold t_2, we consider pc to be subsumed by elements in PC(pc), and delete it from the final set. For example in Table 1, when t_2=0.9, "*pneumonia*" is subsumed by "*acquired pneumonia*" and "*acquired pneumonia*" is then subsequently subsumed by "*community acquired pneumonia*". Note that thresholds can be set based on experience or learned from a gold standard. Pairs generated by pairing the inner-text string with each collocation in the final set (e.g., (***CAP***, *community acquired pneumonia*)) become the output of the DETECT phase.

3.3 SEPARATE

After collocations are detected and a set of pair-wise terms are generated, we then separate them into two sets so that relations that are not abbreviations are treated separately: i) abbreviations associated with expansions, and ii) other semantically related terms, according to lengths of two items in a pair and the existence of capitalized letters since usually, an abbreviation is short and contains capitalized letters, and an expansion is much longer than the corresponding abbreviation. Note that the semantic relations for some pairs can be determined by checking the use of simple patterns in the original text. For example, if we assume NP represents a noun phrase, the patterns "a NP (NP)" or "NP (a NP)" such as, "*indomethacin* ***(a cyclooxygenase inhibitor)***", usually implies a hypernymy relation while the pattern "NP (i.e., NP)" such as "*the congenital neutropenia* ***(i.e., Kostmann's syndrome)*** " implies a synonymy relation.

4 Experiments and Results

Our method was applied to the 2002 version of MEDLINE. The threshold values for t_0, t_1, and t_2 (3, 0.25, 0.9 respectively) were estimated using the abbreviation list in the SPECIALIST Lexicon. In the following, we described the evaluation of our method and assessment of the results.

4.1 Evaluation

We first evaluated the soundness of ignoring potential collocations with frequency equal to one during the selection process by counting the number of pairs in the gold standard set (i.e., the combination of Locus-link collection and Berman's abbreviation list) that occurred only once in MEDLINE and the number of pairs that occurred more than once after normalization using the SPECIALIST Lexicon. We then evaluated the selection process by computing the percentage of pairs with the expansions that were correctly identified to pairs that occurred more than once in MEDLINE.

We performed a manual analysis of the performance of the SEPARATE step using 50 randomly chosen pairs from each set (i.e., a set of abbreviations associated with expansions, and a set of other types of pairs). We then computed the precision based on the manual analysis.

The assessment of the results concentrated on the acquired abbreviation knowledge base. We did not assess other types of pairs since it required a detailed investigation of the original context as well as expert knowledge.

The acquired abbreviation knowledge base was assessed through several measures: frequency distribution, ambiguity (i.e., the number of unique expansions for the same abbreviation), the coverage of the abbreviation knowledge base, the relation of ambiguity to the length of abbreviations, the percentage of the number of abbreviations to the number of abbreviations that were ambiguous, and the percentage of abbreviations that did/did not contain digits, which were ambiguous

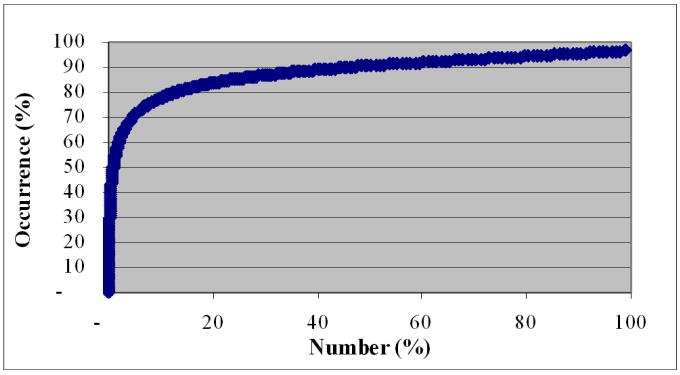

Figure 2: The most frequent pairs in relation with their occurrences. The X-Axis denotes the percentage of the number of the most frequent pairs, and the Y-Axis denotes the percentage of the number of their occurrences.

Len	With Digits		Without Digits	
	# Pairs (% Amb)	AVG	# Pairs (% Amb)	AVG
1	2,039 (37.0)	3.96	240 (56.3)	48.11
2	5,676 (28.3)	1.95	2,571 (56.3)	19.21
3	6,094 (26.1)	1.74	12,016 (66.9)	7.68
4	3,433 (19.7)	1.50	27,540 (36.0)	2.01
>=5	7,802 (8)	1.22	32,657 (17.1)	1.44

Table 2: The ambiguity study results with respect to the existence of digits in abbreviations and the number of letters in the abbreviations.

(all measures were computed by transforming pairs to lower-case except frequency distribution).

4.2 Results

We collected 16,068,562 uses of parenthetical expressions. After filtering out certain uses, such as citations and measures as well as uses where the inner-text occurred only once in the current version of MEDLINE, there were 6,626,790 uses remaining for generating pair-wise terms.

Among 49,536 unique normalized pairs in the gold standard, 5,232 occurred in MEDLINE using parenthetical expressions, and 4,809 (91.9%) occurred more than once in MEDLINE. Expansions associated with 96.3% of the pairs were detected correctly, which suggests that the recall of the method was around 88.5% (i.e., the product of 91.9% and 96.3%) if abbreviations were defined using parenthetical expressions and they were presented in the gold standard.

We collected 381,126 unique pairs, where 308,339 were incorporated into the abbreviation knowledge base, and 72,787 were considered as other types of pair-

wise terms. Two of the 50 pairs from the abbreviation knowledge base that we manually checked were not abbreviation pairs (i.e., and two of the 50 pairs from the set of other types of pair-wise terms were actually abbreviations (i.e., (**NESP, darbepoetin alfa,** *novel erythropoiesis stimulate protein*), and (**GnRH-A,** *GnRH agonist*).

Figure 2 shows the frequency distribution where the X axis denotes the percentage of the number of the most frequent pairs, and the Y axis denotes the percentage of the number of their occurrences. The coverage of the acquired abbreviation knowledge base differed in the two collections: 38.3% for Berman's abbreviation list, and 3% for Locuslink collection. The ambiguity of an abbreviation was related to the existence of digits in the abbreviation as well as to the number of alphabetic letters in the string. Table 2 shows the result, where LEN is the number of letters, and AVG is the average number of expansions.

5 Discussion

We have presented an automatic method for the purpose of automatic acquisition of terminological knowledge for NLP applications using parenthetical expressions in large corpora. A novel aspect of the method is consideration of parenthetical expressions as collocations. Utilizing a collocation selector based on frequency, expansions for abbreviations are automatically recognized. The method has an advantage that it does not require manually crafted patterns or rules for the recognition of expansions. Many abbreviations (i.e., symbols) for biological substances do not follow rules or patterns, and including them in terminologies for NLP applications is important. We believe our method has a higher sensitivity for acquisition of pairs of abbreviations and expansions that occur frequently and should be useful as a complement to pattern-based methods. However, we were unable to. compare our method to other methods because of the absence of a common gold standard set

Our method is not suitable for recognizing expansions that occur only once in the corpora. However, since the purpose of our method is to automatically acquire terminological knowledge, the exclusion of expansions of abbreviations that occur only once from terminological knowledge has almost no impact on NLP applications. Newly defined pairs of abbreviations and expansions will most likely be captured using the most current version of the corpus based on the intuition that if they are accepted by the community, they will be repeatedly referred to and defined in literature; otherwise, they are likely to occur infrequently.

The frequency distribution of pairs of abbreviations and expansions follows the 20/80 rule, i.e., 20% of the pairs contribute to 80% of the total occurrences.

From the result of the coverage study, we found that abbreviations suggested or collected by humans may not be used in the literature. For example, around 10% of expansions from both collections existed in our abbreviation set obtained from MEDLINE but were associated with different abbreviations (1,361 out of 12,098 for Berman's abbreviation collection, and 4,376 out of 42,875 for LocusLink collection). The low coverage of LocusLink was because there are many pairs

associated with digits or letters specifying members, sub-families, or types. For example, there are 50 members in the gene family of *ATP-binding cassette* where the symbol for *ATP-binding cassette* is *ABC* (which is not listed in LocusLink), such that each has been assigned a symbol by attaching a sub-family symbol (from *A* to *G*) and a member number (e.g., *ABCA1*, *ABCB1*, and *ABCG8* etc). The resulting abbreviation knowledge base contains 429 occurrences of the pair (***ABC***, *ATP-binding cassette*), but does not contain symbols of most members in the family, signifying that researchers did not use the same conventions as the LocusLink curators.

The assessment of the ambiguity study shows that abbreviations are highly ambiguous, which is consistent with our previous study [19]. For example, the abbreviation *CAP* represents not only *community-acquired pneumonia*, but also dozens of others, including the following protein names: *catabolic activator protein*, *cystine aminopeptidase*, *cellulose acetate phthalate*, *cyclase-associated protein*, *cementum attachment protein*, *calcium-activated protease*, *capsid-associated protein* etc. We noticed that abbreviations containing digits are much less ambiguous than those without digits. The ambiguity of an abbreviation depends on the number of letters it contains: ones with fewer characters are more ambiguous. In order to allow NLP applications to process and interpret ambiguous abbreviations that are not defined in documents, a disambiguation method is needed. In several previous studies [19], we developed disambiguation methods with a precision of 97% for disambiguating abbreviations given a set of known expansions. The abbreviation knowledge base acquired here can be used as the set of expansions for our disambiguation methods.

Although our method also extracts a set of related terms, we did not focus on this feature in the current paper. A further investigation will be performed in order to assign appropriate semantic relations and semantic categories to those pairs. Future work will also involve use of multiple biomedical databases and use of contextual clues other than parentheses to discover semantic collocations associated with terminology.

6 Conclusion

We proposed and evaluated an automatic method for automatic acquisition of terminological knowledge based on the observation that terms associated with parenthetical expressions are collocations. We identified expansions for abbreviations using frequency information associated with the collocations without the requirement of manually crafted patterns or rules. The method had a precision 96.3% and a recall of around 88.5% for abbreviations that were defined using parenthetical expressions and were presented in the gold standard. We acquired 381,126 unique pairs from the 2002 version of MEDLINE abstracts, where 308,339 were incorporated into the abbreviation knowledge base, and 72,787 were considered as other types of pair-wise terms. Abbreviations are highly ambiguous and the ambiguity is related to the number of alphabetic letters in the abbreviation and the existence of digits: ones with fewer characters are more ambiguous, and

ones with digits are less ambiguous. In the future, we plan on expanding this method in order to obtain a more comprehensive terminological knowledge using other corpora such as full articles and online databases.

Acknowledgment

This study was supported in part by grants LM06274 from the NLM and ILS-9817434 from the NSF.

Reference List

1. Lauriston A. Ph.D Dissertation *Automatic Term Recognition: performance of Linguistic and Statistical Techniques*. 1996 UMIST, Manchester.
2. Justeson J, Katz S. *Technical terminology: some linguistic properties and an algorithm for indentification in text*. Nat Lang Eng 1995; 1(1):9-27.
3. Frantzi K, Ananiadou S. *The C-value/NC-value domain indendent method for multi-word term extraction*. Journal of Natural Language Processing 1999; 6(3):145-180.
4. Georgantopoulos B, Piperidis S *Automatic Acquisition of Terminological Resources for Information Extraction Applications*. Proc. of New Information Technologies 1998
5. Friedman C, Hripcsak G, DuMouchel W, Johnson SB, Clayton PD. *Natural language processing in an operational clinical information system*. Nat Lang Eng 1995; 1(1): 83-108.
6. Friedman C *A Broad Coverage Natural Language Processing System*. Proc AMIA Symp 2000: 270-274.
7. Aronson A *Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program*. Proc. AMIA Symp 2001: 17-21.
8. Nadkarni P, Chen R, Brandt C. *UMLS concept Indexing for Production Databases: A Feasibility Study*. J Am Med Inf Assoc 2001; 8(1): 80-91.
9. Finkelstein-Landau M, Morin E *Extracting Semantic Relationships between Terms: Supervised vs Unsupervised Methods.* Proc. International Workshop on Ontological Engineering on the Global Information Infrastructure 1999:71-80
10. Hisamitsu T, Niwa Y *Extraction of useful terms from parenthetical expressions by using simple rules and statistical measures*. The First Workshop on Computational Terminology Computerm'98 1998: 36-42.
11. Yoshida M, Fukuda K, Takagi T. *PNAD-CSS: a workbench for construction a protein name abbreviation dictionary*. Bioinformatics 2000; 16(2):169-175.
12. Pustejovsky J, Castano J, Cochran B, Kotechi M, Morrell M, Rumshisky A *Extraction and Disambiguation of Acronym-Meaning Pairs in Medline*. Medinfo 2001
13. Yu H, Hripcsak G, Friedman C. *Mapping abbreviations to full forms in electronic articles*. JAMIA 2002: 9(3):262-272.
14. Smadja F. *Retrieving Collocations from Text: Xtract*. Computational Lingustics 1993;19:143-176.
15. MEDLINE. http://www.nlm.nih.gov . 2001.
16. NIH. Unified Medical Language System. NIH. 2000.
17. Berman J. A classification for Medical Abbreviations. 2002. http://www.pathinfo.com/jjb/berman1j.htm.
18. Maglott D, Katz K, Sicotte H, Pruitt K. *NCBI's LocusLink and RefSeq*. Nucleic Acids Research 2000; 28:126-128.
19. Liu H, Johnson SB, Friedman C. Automatic Resolution of Ambiguous Terms Based on Machine Learning and Conceptual Relations in the UMLS. J Am Med Inf Assoc 2002: 9(6): 621-636