MEDINFO 2007
HISA Ltd Australia
© 2007 The authors. All rights reserved.

Supplementing the UMLS with Acronym/Abbreviation Knowledge for Natural Language Processing Applications

Manabu Torii, Hongfang Liu

Department of Biostatistics, Bioinformatics and Biomathematics
Georgetown University Medical Center, USA

Abstract

Background: *With the widespread use of acronyms/abbreviations (denoted as short forms, abbreviated as SFs) in biomedical domain, a terminology knowledge source that associates SFs with their descriptive expressions (denoted as long forms, abbreviated as LFs) needs to be supplemented further for the improvement of natural language processing (NLP) systems in the biomedical domain.* ***Objectives:*** *We investigate a framework to supplement the UMLS, a terminology knowledge source, with information of SFs and LFs extracted from text.* ***Method:*** *Given a pair of (SF, LF) automatically collected from text using certain patterns such as parenthetical expressions, there will be three (exclusive) scenarios: (1) LF and SF are listed in the UMLS as synonyms, (2) only LF is listed in the UMLS, and (3) LF is not included in the UMLS. For the second scenario, we can safely supplement the UMLS with SF and mark SF with the annotations associated with LF. For the third scenario, even though we are unable to associate SFs with existing concepts in the UMLS, we can annotate LFs and SFs with UMLS semantic groups under the assumption that the UMLS semantic classification system is comprehensive. Utilizing a collection of (SF, LF) pairs obtained previously using five months' worth of Medline abstracts and also a semantic group assignment system constructed based on headwords and suffixes, we evaluate the framework.* ***Results:*** *We found that approximately 18k of SFs could be newly marked with existing concepts in the UMLS, and furthermore 40k of (SF, LF) pairs could be assigned with UMLS semantic categories with an average precision as 79%. Supplementing the UMLS with such acquired acronym/abbreviation knowledge would be critical for NLP applications that depend on the UMLS for their terminology needs.*

Keywords:

medical informatics; natural language processing; vocabulary, controlled

Introduction

From healthcare news to journal articles in medicine, acronyms/abbreviations have been widely used to provide concise presentations for domain-specific concepts. However, they pose a great challenge to automated text processing systems as well as to human readers. Unlike their corresponding full forms, acronyms/abbreviations are usually not descriptive representations of their referred concepts. For example, when skimming through paper titles at PubMed, one may have difficulty in understanding the meaning of acronyms such as VHA[1] and HIT[2] in the title *"An HIV collaborative in the VHA: do advanced HIT and one-day sessions change the collaborative experience?"* (PMID: 16776387). Since most acronyms/abbreviations represent important domain-specific concepts, and they are highly ambiguous, associating them with the correct concept is a critical task for building natural language processsing (NLP)-based systems for clinical decision support (CDS) or Evidence-Based Medicine (EBM) (see, e.g., [1]). A comprehensive terminology knowledge source that lists all possible full forms associated with an acronym/abbreviation is needed for the task. In the following, for simplicity, we call acronyms/abbreviations as short forms (SFs) and the corresponding full forms as long forms (LFs).

In the UMLS, SFs and their corresponding LFs are considered as synonyms, but the coverage of the UMLS regarding to acronym/abbreviation is not comprehensive. In this paper, we propose a framework for supplementing the UMLS with the acronym/abbreviation knowledge [2]. The framework consists of two steps: (i) automatic extraction of SFs and their corresponding LFs from text, and (ii) assignment of semantic categories with SFs. Both tasks in these steps have been tackled in this domain (e.g., [3, 4]), and the automatic supplementation of terms to the UMLS has also been considered (e.g., [5]). In this work, we aim to integrate theses two tasks and propose a framework specialized in supplementing the UMLS with SF knowledge and evaluate the framework utilizing results from two previous studies: one is to compare different SF definition detection systems, and another is to build a UMLS semantic category assignment system based on headwords and suffixes [6].

1 "VHA" = "Veterans Health Administration"
2 "HIV" = "health information technology"

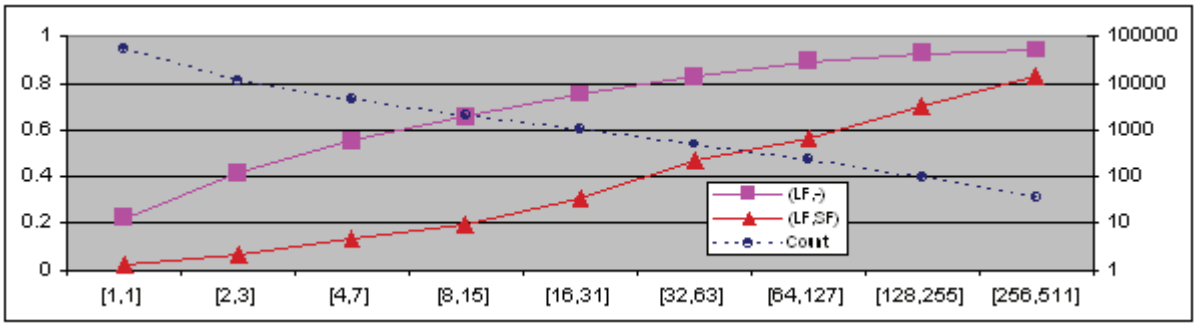

Figure 1 - X-axis represents bins of SF.LF pairs with different frequency ranges. A dotted line shows the number of pairs in each bin (read the right axis). Two solid lines show (i) the coverage of LF, i.e., (LF,-), and (ii) the coverage of LF.SF pairs, i.e., (LF,SF), in UMLS (read the left axis)

Resources and previous work

The UMLS

The UMLS, developed and maintained by National Library of Medicine (NLM), is an extensive terminology knowledge source in the medical domain. We used two components of the UMLS for this study, Metathesaurus (META) and SPECIALIST lexicon. In META, terms referring to the same biomedical concept (i.e., synonyms) are assigned with the same concept ID, e.g., "human immunodeficiency virus" and "HIV" are assigned with a concept ID C0019682. Moreover, in META, each concept (represented by a concept ID) is associated with UMLS semantic types, e.g., C0019682 is associated with a type Virus. As for SPECIALIST lexicon, we used one of its tables LRABR that records SF and LF phrase pairs. We also used another table LRAGR for normalizing noun tokens.

The SF definition detection comparison study

In this study [7], we used three independent systems ([8-10]) to extract pairs of SF and LF from *parenthetical expressions*, e.g., given a sentence "An electronic medical record (EMR) is a …", the systems will extract phrases "electronic medical record" and "EMR" as LF and SF. The three systems were applied to Medline abstracts published between January and May 2006 (210k records), and we collected 224k pairs (denoted as SF.LF) that are recognized by all of the three systems. After merging identical pairs[3], we obtained 74,428 unique SF.LF pairs. We manually inspected over 500 pairs randomly sampled from the obtained SF.LF, and found that all the pairs were valid SF and LF pairs.

Given this set of SF.LF pairs, we repeat the procedure described in [7] and summarized below to show the coverage of the UMLS over SF.LF pairs. We first grouped SF.LFs according to the frequency of their occurrences in their corpus to several bins. We then reported (i) the coverage of LF in the UMLS, and (ii) the coverage of both SF and LF in UMLS. The result is shown in Figure 1 where X-axis indicates bins according to frequency and Y-axis indicates the percentage. From Figure 1, we can see that (i) the coverage of SF.LF pairs lags behind that of LF (i.e., the gaps between the two solid lines), and (ii)even for frequent pairs, LFs may not be included in the UMLS

Semantic group assignment based on headwords and suffix

Headwords and suffixes of headwords are the two most effective features for automatic semantic group assignment (see, e.g., [4-6]). For example, a headword "virus" as in "Human Immunodeficiency Virus" strongly implies the semantic category of the concept referred. Similarly, a suffix "cyte" as in "Acanthocyte" implies that the referred entity is a cell. Hence, as in our previous work [6], we built a simple assignment system that assigns semantic groups to name phrases using headword/suffix information. In short, the system first extracts headwords/suffixes from a category-annotated list of name phrases (e.g., META), and then associate each headword/suffix with the most prevalent semantic category assigned to the phrases bearing the same headword/suffix.

Methods

Framework for supplementing the UMLS

- Our framework to supplement the UMLS consists of the two major subtasks (shown in Figure 2):
- Automatic extraction of SF.LF pairs from biomedical text
- Development of a semantic category assignment system

3 Identity of pairs is tested after the normalization of LF, which includes normalization of upper/lower cases, singular/plural forms, hyphen/blanks, etc.

After extracting SF.LF pairs in the first subtask, there would be three scenarios in terms of our objective:

1. Both SF and LF are recorded as synonyms in the UMLS
2. Only LF is in the UMLS but SF is not in the UMLS as a synonym of LF
3. LF is not in UMLS.

In the first scenario (1), we do not need to do anything because both SF and LF are already in the UMLS as synonyms. In the second scenario (2), we can safely mark SF with the concept ID of the paired LF. In the third scenario (3), we cannot associate LF with an existing concept ID, but we can assign the UMLS semantic category to SF and LF based on headwords and suffixes of LF.

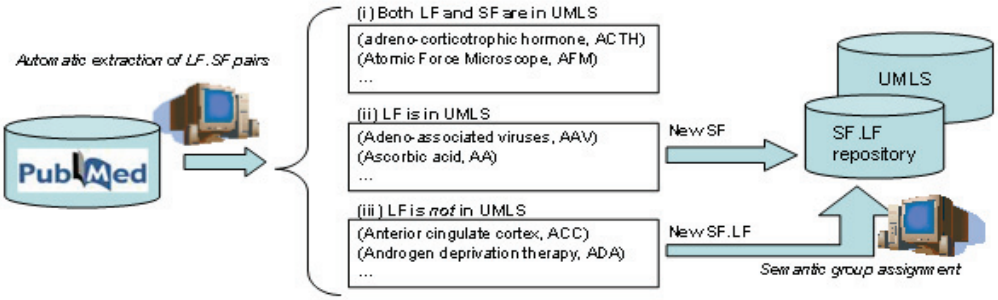

Figure 2 - Overview of the framework to suppplement UMLS

Evaluation method

For step 1, we used the collection of SF.LF pairs obtained in the comparison study to evaluate the framework (see the section Resources and Previous Work)[6]. For the semantic assignment step, instead of using the entire META for obtaining headwords and suffixes, we used only LFs from the collection that can be looked up in META through a procedure that normalizes textual variants (i.e., ignoring punctuations and normalizing using the Specialist lexicon). Similarly to [6], we used 15 semantic groups to provide reasonable granularity for the assignment system. After extracting headwords/suffixes from those LFs, we recorded them together with semantic groups assigned to the source name phrases, e.g., ”leukemia” appears as a headword in 39 phrases in META, of which 39 are annotated as Disorders, but one is annotated as Genes (note that a phrase can be annotated with more than one semantic type in the UMLS). When these headwords/suffixes are used for semantic group assignment, only the majority group is considered for assignment. For example, given a phrase “chronic mylogenous leukemia”, it is annotated with Disorders because the majority group of its headword “leukemia” is associated with Disorders. The system is then applied to pairs where LFs are not in the UMLS. Similarly to the comparison study in the previous section, we report the results with respect to different frequency bins. In order to investigate the precision of the semantic group assignment system, from each bin of LF.SF pairs (where bins from [32,63] to [256,511] were merged as [32,511]), we randomly picked as many as 100 LF.SF pairs where the assignment rules are applicable to LFs.

Results

Assignment of semantic groups to novel SFs

Of 74,428 unique SF.LF pairs automatically extracted from five month’ worth of Medline, approximately 5% of them were recognized as synonyms using the UMLS. However, in 30% of the cases, LF could be found in the UMLS. Therefore, by incorporating the SF.LF relations automatically recognized, SFs in 25% (30-25) of the pairs (i.e., 18k SFs) could be associated with existing concept IDs in the UMLS.

For the remaining 52,348 pairs where LFs cannot be found in the META, the semantic group assignment system was applicable to 81% of them. Among pairs where the system was applicable, the distribution of assigned semantic groups varied much: some with higher percentage (e.g., Chemicals & Drugs 18%, Procedures 16%, Disorders 13%), and some with lower percentage (e.g. Occupations 0.2%, Geographic Areas 0.1%). Given this highly skewed distribution, we looked into the performance of the five significant semantic groups, Procedure, Chemicals & Drugs, Disorders, Anatomy, and Organizations. The results about these five semantic groups are shown in Table 1. The first row of the table indicates that there are 42,579 pairs occurring once in the text. Around 80% of them can be automatically associated with UMLS semantic groups with a precision of 86%. There is a slight

increasing trend of percentage of pairs where the assignment system was applicable based on the third column of table 1, but not significant.

Discussion

As for 19% of the pairs where the assignment system was not applicable, we observed several reasons, e.g., typos (e.g., 12-lipoxygeanse), or adjectival phrases (e.g., single-stranded).

Though we found that the system achieved the precision of 80% or more for the most of the semantic groups, we concern the poor performance for the important and distinct semantic group Disorders. After looking at the error cases for the group Disorders, we found some of the shortest suffixes are used in the error cases, e.g., "sion" which is extracted from "hypertension", "occlusion", "convulsion", etc. In our previous study, short suffixes were not the major problem: those short suffixes are not used often because there are longer suffixes and headwords that have precedence over them. For example, given a phrase "rapid maxillary expansion" (RME), the phrase can be properly associated with a group Procedure owing to phrase entries with the headword "expansion" or even "maxillary expansion" in the UMLS. We can easily resolve the problem by increasing the amount of source name phrases when building the assignment system. For example, we can use the whole collection of Medline abstracts to have a large collection of LF.SF pairs or use all META phrases that appear in text at least once.

Freq. Range	Num. pairs	Type assigned	hand-checked	Procedures	Chemicals & Drugs	Disorders	Anatomy	Organizations	All
[1,1]	42,579	80%	100	95% (20)	88% (25)	68% (19)	90% (10)	100% (3)	86% (100)
[2,3]	6,712	81%	100	95% (13)	100% (14)	52% (21)	86% (14)	100% (2)	74% (100)
[4,7]	1,990	82%	100	72% (29)	85% (20)	47% (15)	75% (8)	100% (2)	70% (100)
[8,15]	698	82%	100	96% (23)	93% (14)	53% (17)	100% (4)	0% (1)	85% (100)
[16,31]	252	84%	100	79% (33)	88% (16)	75% (12)	100% (7)	100% (2)	83% (100)
[32,511]	117	83%	97	81% (16)	85% (26)	59% (13)	86% (7)	100% (3)	75% (97)

Table 1 - precision of semantic group assignment for five semantic groups

Conclusion

In this study, we proposed a framework to supplement a biomedical thesaurus with SFs as well as LFs. In general, SFs can be highly ambiguous semantically and it is difficult to determine their semantic categories. In our framework, SFs are annotated with semantic categories through the corresponding LFs. Another idea employed was that, in building a semantic category assignment system, we did not use the entire META, but we used only those phrases that are actually used in text so as to identify viable name phrases in biomedical text.

In sum, from the five month's worth of Medline records, approximately 18k SFs are newly associated with existing concept IDs in UMLS, and furthermore, 40k SFs and LFs are supplemented to the UMLS, where these phrases were assigned with semantic categories with the over all precision of 79%.

While we investigated the framework over the five month's worth of Medline records, the framework can be extended to a larger set of records. With a larger corpus, we could not only supplement the UMLS further, but we might incorporate more source name phrases in building a semantic category assignment system, which may improve the performance of semantic category assignment system further.

Acknowledgments

This project is supported by IIS-0430743 from the National Science Foundation. We thank Hiroko Ao, Jeffrey Chang, Ariel Schwartz, and each of their colleagues for making their extraction systems available/accessible.

References

[1] W. R. Hersh, E. H. Campbell, D. A. Evans, and N. D. Brownlow, "Empirical, automated vocabulary discovery using large text corpora and advanced natural language processing tools," *Proc AMIA Annu Fall Symp*, pp. 159-63, 1996.

[2] O. Bodenreider, "The Unified Medical Language System (UMLS): integrating biomedical terminology," *Nucleic Acids Res*, vol. 32, pp. D267-70, 2004.

[3] J. Pustejovsky, J. Castaño, B. Cochran, M. Kotecki, M. Morrell, and A. Rumshisky, "Extraction and Disambiguation of Acronym-Meaning Pairs in Medline," *Medinfo*, vol. 10, pp. 371-375, 2001.

[4] M. Krauthammer and G. Nenadic, "Term identification in the biomedical literature," *Journal of Biomedical Informatics*, vol. 37, pp. 512-526, 2004.
[5] O. Bodenreider, T. C. Rindflesh, and A. Burgun, "Unsupervised, Corpus-Based Method for Extending a Biomedical Terminology," presented at ACL workshop on NLP in biomedical domain, 2002.
[6] M. Torii and H. Liu, "Headwords and Suffixes in Biomedical Names," presented at KDLL, 2006.
[7] M. Torii, H. Liu, Z. Hu, and C. Wu, "A comparison study of biomedical short form definition detection algorithms," presented at TMBio, 2006.
[8] H. Ao and T. Takagi, "ALICE: an algorithm to extract abbreviations from MEDLINE," *J Am Med Inform Assoc*, vol. 12, pp. 576-86, 2005.
[9] J. T. Chang, H. Schutze, and R. B. Altman, "Creating an online dictionary of abbreviations from MEDLINE," *J Am Med Inform Assoc*, vol. 9, pp. 612-20, 2002.
[10] A. S. Schwartz and M. A. Hearst, "A simple algorithm for identifying abbreviation definitions in biomedical text," *Pac Symp Biocomput*, pp. 451-62, 2003.

Address for correspondence

Manabu Torii, Department of Biostatistics, Bioinformatics, and Biomathematics, Georgetown University Medical Center, Washington, DC 20007, USA.mt352@georgetown.edu