# Clinical Acronym/Abbreviation Normalization using a Hybrid Approach

Yonghui Wu[1], Buzhou Tang[1], Min Jiang[1], Sungrim Moon[1], Joshua C. Denny[2], Hua Xu[1,2,*]

[1]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, Texas, USA
[2]Department of Biomedical Informatics, Vanderbilt University, Nashville, TN, USA
{yonghui.wu, buzhou.tang, min.jiang, sungrim.moon, hua.xu}@
uth.tmc.edu, josh.denny@vanderbilt.edu

**Abstract.** A unique characteristic of clinical text is the pervasive use of acronyms and abbreviations, which are often ambiguous. The ShARe/CLEF eHealth Evaluation Lab organized three shared tasks on clinical natural language processing (NLP) and information retrieval (IR) in 2013 and one of them was to normalize acronyms/abbreviations to UMLS concept unique identifiers (CUIs). This paper describes a hybrid system, which combines different Word Sense Disambiguation (WSD) methods and existing knowledge bases to normalize and encode clinical abbreviations. Our system achieved the best accuracy of 0.719 on the independent test set, which was ranked first in the challenge.

**Keywords:** clinical abbreviation, word sense disambiguation, support vector machines, vector space model.

## 1 Introduction

Abbreviations are widely used in clinical texts and often contain important clinical meanings. Any clinical NLP systems attempting to extract clinical information from free texts have to interpret abbreviations correctly. However, identification and normalization of abbreviations remains a challenging task, as clinical abbreviations are highly dynamic and ambiguous. Researchers have applied different methods to detect abbreviations from clinical texts [1][2][3], construct clinical abbreviation knowledge bases[4][5][6][7], and disambiguate ambiguous abbreviations [8][9][10]. Many clinical NLP systems such as MedLEE[11], MetaMap[12], Knowledgemap[13], and cTAKEs[14] have been developed to extract medical concepts from the clinical texts to facilitate health care and clinical research. However, recent studies showed that performances of these clinical NLP systems on identifying abbreviations were still not very satisfactory[15].

[*] corresponding author

The ShARe/CLEF eHealth Evaluation Lab organized three shared tasks in 2013, with a focus on Natural Language Processing (NLP) and Information Retrieval (IR) in the medical domain.[16] The Task 2 - Normalization of acronyms/abbreviations (we will use abbreviations for short in the rest of this paper) was to map acronyms and abbreviations to the UMLS (Unified Medical Language System) CUIs (Concept Unique Identifiers). The organizers highlighted abbreviations in clinical text and all participants were asked to develop a system to map these abbreviations into appropriate UMLS CUIs. This paper presents a detailed description of the abbreviation normalization system developed by the University of Texas School of Biomedical Informatics team. By combining different state-of-the-art word sense disambiguation (WSD) methods and knowledge bases, our system achieved the best accuracy of 0.719, ranked first in the challenge.

## 2 Method

Normalizing abbreviations to UMLS CUIs is an encoding problem. However, many abbreviations are ambiguous (has multiple meanings). Therefore we have to determine the correct sense/meaning of an abbreviation (a Word Sense Disambiguation (WSD)[17] problem), before mapping it to an UMLS CUI. Different WSD methods have been developed, including supervised machine learning methods, knowledge-based methods, and hybrid approaches [8][18][19]. Different methods have their own pros and cons. For example, supervised machine learning based methods often show good performance; but they require annotated samples for each abbreviation[20]. To maximize the performance of our system, we developed a hybrid strategy that combines a machine learning based method using Support Vector Machines (SVMs), a profile-based method using Vector Space Model, and a majority-sense method to resolve ambiguous abbreviations occurring at different frequency levels. Figure 1 shows an overview of our strategy.

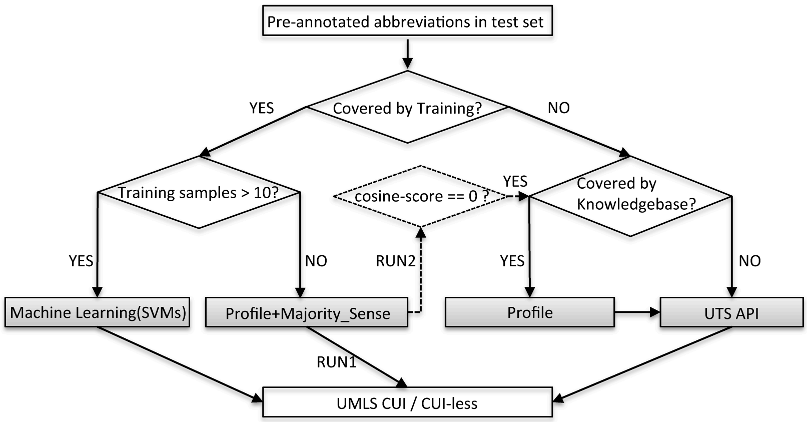

**Figure 1.** An overview of our system.
**UTS API**: UMLS Terminology Services API

### 2.1 Datasets

The organizer annotated abbreviations in 298 clinical notes, which were broken into a training set (199 notes) and a test set (99 notes). The detailed numbers of annotated abbreviations with their corresponding CUIs in the training and test sets are shown in Table 1. The reported numbers of abbreviations were counted after variation normalization - annotated abbreviations were normalized by removing special symbols (“.” , “-”, “_” and “+/-”) and converting into lowercases.

**Table 1.** Characteristics of the dataset.

| Dataset | Type | #Note | #Abbreviation<br>Total (Unique) | #CUI<br>Total (Unique) | #CUI-less |
|---|---|---|---|---|---|
| Training | ALL | 199 | 3805 (679) | 3624 (696) | 181 |
| | ECH | 42 | 684 (57) | 680 (58) | 4 |
| | RAD | 42 | 543 (156) | 505 (137) | 38 |
| | DIS | 61 | 2514 (576) | 2375 (597) | 139 |
| | ECG | 54 | 64 (3) | 64 (3) | 0 |
| Test | ALL | 99 | 3774 (714) | 3553 (706) | 221 |
| | ECH | 12 | 207 (39) | 207 (39) | 0 |
| | RAD | 12 | 134 (53) | 129 (55) | 5 |
| | DIS | 75 | 3433 (678) | 3217 (673) | 216 |
| | ECG | 0 | 0 (0) | 0 (0) | 0 |

### 2.2 Abbreviation Normalization

Our abbreviation normalization system consisted three steps: 1) find possible senses (also called sense inventory) for each abbreviation; 2) assign a correct sense from sense inventories to each occurrence of an abbreviation – the WSD step; and 3) map the assigned sense string to an UMLS CUI – the encoding step.

#### 2.2.1 Sense Inventory.

A straightforward method of building the sense inventory is to use all annotated abbreviations and their senses from the training set. However, there is no guarantee that the training data cover all abbreviations and all possible senses in the test data set. In this study, we constructed two sense inventories. One is a “limited” sense inventory that was built from the training corpus only by collecting all abbreviations and their senses annotated in the training set. The other one is a “broad” sense inventory built on existing clinical abbreviation knowledge bases. We included knowledge sources such as the UMLS LRABR, the ADAM [6] and the Bermen’s pathology abbreviation list[4]. For an abbreviation in test set, if it was occurred in training set, we used the “limited” sense inventory; otherwise the “broad” sense inventory was used.

### 2.2.2 WSD methods

Our system adopted three different WSD methods, including the machine learning based method using SVMs, the profile based method, and the majority-sense method.

To build SVM-based classifiers for the ambiguous abbreviations, different types of features were extracted from the training data, including: *1) Word features* - stemmed words within a window size of 5 of the annotated abbreviation. The Snowball Stemmer from python NLTK (Natural Language Tookit) package was used in this research. *2) Word feature with direction* - The relative direction (left side or right side) of stemmed words in feature set 1) towards the annotated abbreviation. *3) Position feature* - The distance between the feature word and the target abbreviation. *4) Word formation features from the annotated abbreviation* - include: a) special characters such as "-" and "."; b) features derived from the different combination of numbers and letters; c) features derived from the number of uppercase letters. *5) Note type feature* - types of notes derived from the file name. There are four types of notes as shown in Table 1. *6) Section feature* - we developed a program to automatically extract candidate section headers and manually reviewed them to remove false positives and aggregate the variations. A list of 38 unique section headers were constructed and used to extract section information. The parameters of SVM were optimized and determined based on a 5-fold cross validation using the training set.

In a previous study[8], we have developed a profile-based WSD method that used dictated discharge summaries as an external source to build sense profiles and applied them to disambiguate abbreviations in admission notes via a vector space model. The method starts with a given sense inventory. For each sense of an ambiguous abbreviation, it searches for the sense string (long form of an abbreviation) in a corpus to automatically create the pseudo training samples, from which a profile vector is created for each sense. During disambiguation, a context vector of the testing sample is created and compared with each sense profile vector to calculate the cosine-similarities. The sense corresponding to the highest similarity score will be selected as the correct sense. In this study, if a testing abbreviation occurred in the training corpus, we built sense profiles using the training corpus based on the "limited" sense inventory. Otherwise, we built sense profiles by using 3 years of clinical notes from Vanderbilt University Hospital (2007-2009) based on the "broad" sense inventory. All the features used in the machine learning based method (except the group 4) were used to build sense profiles.

The majority-sense based method is very simple. It always takes the majority sense of an ambiguous abbreviation as the correct sense. The challenge is to find the majority sense. We applied this method to abbreviations occurred in the training corpus only because we could estimate the majority senses of these abbreviations based on the training set.

### 2.2.3 Encoding senses to UMLS CUIs

Once we determine the sense of an ambiguous abbreviation, we need to map the sense string to an appropriate UMLS CUI in this task. For abbreviations occurred in the training set, we used the "limited" sense inventories, where UMLS were already associated with each sense based on the annotation. Therefore it is straightforward to assign UMLS CUIs for these abbreviations. For abbreviations that were not covered by the training corpus, we used the UMLS Terminology Service (UTS) API to assign a CUI to a sense string.

### 2.2.4 Strategies

As shown in Figure 1, we divided abbreviations to following different groups based on their frequency in the training corpus and applied different sense inventories and WSD methods to these different groups.

*1) High Frequency Ambiguous Abbreviation (HFA) –* These abbreviations occurred in the training corpus no less than 10 times and had more than one sense according to the annotation. The machine learning based approach was applied to these abbreviations.

*2) High Frequency Non-Ambiguous abbreviation (HFNA)* - These abbreviations occurred in the training corpus no less than 10 times, but had only one sense according to the annotation. Although it is possible that these abbreviations could have other senses in the testing set. We assumed the sense based on the training set was the majority sense and simple applied the majority-sense method to this group.

*3) Low Frequency abbreviation (LF)* - These abbreviations occurred in the training corpus no less than 10 times. Because of the limited sample sizes, machine learning based methods were not appropriate for this group. We primarily applied the profile-based method to this group of abbreviations, at two settings 1) "limited" sense inventory + profiles based on training corpus, and 2) "broad" sense inventory + profiles based on Vanderbilt corpus. To further improve the performance, we combined majority sense with the profile-based method, as our previous study showed beneficial performance with this approach [19]. Here we did not discriminate between ambiguous and non-ambiguous abbreviations. The top ranked sense was selected. Apparently, the rationale behind setting 1 was to trust the Training data, whereas the setting 2 was to trust the knowledge base. Our submission 1 used the setting 1 to normalize the low frequency abbreviations. In submission 2, we further collected the samples with a zero similarity-score from submission 1, and use setting 2 to process these zero scored samples. If the samples were not covered by the knowledge-based profiles, we simply moved them into the uncovered abbreviation group (see below).

*4) Uncovered abbreviation (UN)* - These abbreviations never occurred in the training corpus. We applied the "broad" sense inventory and the profile-based method built on Vanderbilt corpus. Any abbreviations that were not covered by the sense inventory or profile-based WSD method was directly processed by the UTS API to assign a CUI. If nothing returns from the UTS API, "CUI-less" was be assigned to the abbreviation.

### 2.3 Evaluation

The accuracy was used to evaluate all participating systems. For each submitted system run, the evaluation will generate a "strict accuracy score", defined as the proportion of correctly normalized abbreviations with the top code selected by the annotators (one best), and a "relaxed accuracy score", defined as the proportion of correctly normalized abbreviations based on a list of possibly matching codes generated by the annotators (n-best).

## 3 Result

By combining existing abbreviation knowledge bases, we were able to construct a comprehensive sense inventory composed of 42,613 unique abbreviations with 102,150 possible senses. We were able to build profiles for 7,503 abbreviations from the Vanderbilt Discharge Summaries and 5,345 abbreviations from the Vanderbilt Radiology notes. Table 2 shows the final scores of our system reported by the organizer, where the best results are in bold. Our best run (#1) achieved the best strict score and relaxed score of 0.719 and 0.725, respectively, which was ranked No. 1 in the challenge.

**Table 2.** Performance of our system on the test set.

| | Strict | Relaxed |
|---|---|---|
| Run1 | **0.719** | **0.725** |
| Run2 | 0.683 | 0.689 |

Table 3 shows the numbers of abbreviations as well as our system's performance in different frequency groups. There were 50 ambiguous high frequency abbreviations, which accounted for 40.21% of total abbreviation occurrences in the training set. The average accuracy of these 50 ambiguous abbreviations was 88.26% on the training set using 5-fold cross validation. There were 43 high frequency non-ambiguous abbreviations, which contributed 847 occurrences (22.26% of total occurrences) in the training set. The low frequency abbreviation group composed of 586 abbreviations with a total occurrence of 1,428 (37.53% of total occurrences).

**Table 3.** Strict score for each divided group.

| | HFA | HFNA | LF | UN |
|---|---|---|---|---|
| #Abbr (#instance) in Training | 50 (1530) | 43 (847) | 586 (1,428) | - |
| #Abbr (#instance) in Test | 50 (1341) | 43 (644) | 284 (1,226) | 337 (563) |
| Run1 accuracy | 82.03% | 96.58% | **75.69%** | 11.20% |
| Run2 accuracy | 82.03% | 96.58% | 64.52% | 11.20% |

**HFA**: High Frequency Ambiguous; **HFNA**: High Frequency Non-Ambiguous; **LF**: Low Frequency; **UN**: Uncovered

Table 3 also tells us that all 50 ambiguous high frequency abbreviations and 43 non-ambiguous high frequency abbreviations occurred in the training set also appeared in the test set. However, only 284 out of 586 low frequency abbreviations in the training also appeared in the test set. There were 337 abbreviations occurred in the test set but not in the training set. For the Low Frequency abbreviations, Run 1 achieved the best accuracy of 75.69%, which outperformed run 2 for more than 11%. The scores for other groups were the same as they adopted the same strategies.

## 4 Discussion

In this study, we developed a hybrid system to normalize abbreviations in clinical text to UMLS CUIs and applied it to the 2013 ShARe/CLEF eHealth shared task 2. Although ranked first in the challenge, the best performance of our system was an accuracy of 0.719, indicating abbreviation normalization is still a challenging task in clinical NLP research.

Different WSD methods have been developed to resolve ambiguity of clinical abbreviations. Each of them has its advantages and limitations. Majority-sense based method is often used as a baseline and a number of studies have reported reasonable performance of this approach in clinical abbreviations [8][20]. However, how to determine the majority sense of each abbreviation is not straightforward. Supervised machine learning-based methods often reach high performance in WSD tasks. But it requires annotated data for each ambiguous abbreviation, which is not very practical. Moreover, it requires that training and test data sets should be similar (e.g., same type of clinical notes). In our previous study[8], the profile-based method demonstrated reasonable performance and good transportability across different types of clinical notes. As shown in Table 2, most of these methods showed expected results. Machine learning methods achieved an accuracy of 0.8203 on high frequency ambiguous abbreviations and "profile + majority-sense" reached an accuracy of 0.7569 when they were built from the training corpus. However, we noticed that the profile-based method did not work well on uncovered abbreviations (accuracy 0.1120). Most likely this is due to that the profiles were created from clinical notes at a different institution (Vanderbilt). This finding indicates that the generalizability of the profile-based approach needs further investigation when it is applied across different institutions.

Another critical but often-neglected issue of clinical abbreviation normalization is the lack of comprehensive knowledge bases that contain abbreviations and their senses (also called sense inventory). In reality, a sense inventory is a prerequisite for any WSD methods. In this study, we combined multiple existing abbreviation/sense lists and generated a comprehensive sense inventory, with the hope to capture abbreviations that were not occurred in the training set. However, our analysis showed that only about 68.25% uncovered abbreviations were in our list and only about 36% of senses of uncovered abbreviations were in our knowledge-based sense inventory. This finding indicates the need of building sense inventories for clinical abbreviations. We are currently working on this problem, by developing clustering based methods for building clinical abbreviation sense inventories [7].

To further understand the errors of our system and explore opportunities for further improvements, we manually reviewed 50 incorrectly normalized abbreviations in the uncovered set. Among the 50 abbreviations, the knowledge-based profile covered the senses of 18 abbreviations only (36%), indicating insufficient sense inventories. We also found that for 25 incorrectly normalized abbreviations, the system predicted CUIs were closely related to the gold standard annotations (e.g., the system predicted CUI - "C2208743-serum BUN/creatinine ratio" versus the gold standard CUI - "C0010294-Creatinine"). This finding suggests that mapping an identified sense string to an appropriate UMLS CUI is still challenging.

## 5 Conclusion

In this paper, we introduced a system to normalize pre-annotated abbreviations developed for the task 2 of ShARe/CLEF eHealth 2013 challenge. Our system achieved the best strict score of 0.719 and relaxed score of 0.725, which ranked first in the challenge.

**Acknowledge**
This study is supported in part by grants from NLM R01LM010681, NIGMS 1R01GM102282, and NCI R01CA141307, and by grant No. 10510592 for Patient-Centered Cognitive Support under the Strategic Health IT Advanced Research Projects Program (SHARP) from the Office of the National Coordinator for Health Information Technology. The annotation corpus was provided by the Shared Annotated Resources (ShARe) project funded by the United States National Institutes of Health with grant number R01GM090187.

## References

[1] H. Liu, Y. A. Lussier, and C. Friedman, "A study of abbreviations in the UMLS," *Proc Amia Symp*, pp. 393–7, 2001.

[2] H. Xu, P. D. Stetson, and C. Friedman, "A study of abbreviations in clinical notes," *Amia Annu Symp Proc*, pp. 821–5, 2007.

[3] Y. Wu, S. T. Rosenbloom, J. C. Denny, R. A. Miller, S. Mani, D. A. Giuse, and H. Xu, "Detecting abbreviations in discharge summaries using machine learning methods," *Amia Annu Symp Proc*, vol. 2011, pp. 1541–9, 2011.

[4] J. J. Berman, "Pathology abbreviated: a long review of short terms," *Arch Pathol Lab Med*, vol. 128, pp. 347–52, Mar. 2004.

[5] P. D. Stetson, S. B. Johnson, M. Scotch, and G. Hripcsak, "The sublanguage of cross-coverage," *Proc Amia Symp*, pp. 742–6, 2002.

[6] W. Zhou, V. I. Torvik, and N. R. Smalheiser, "ADAM: another database of abbreviations in MEDLINE," *Bioinformatics*, vol. 22, pp. 2813–8, Nov. 2006.

[7] H. Xu, Y. Wu, N. Elhadad, P. D. Stetson, and C. Friedman, “A new clustering method for detecting rare senses of abbreviations in clinical notes,” *J. Biomed. Inform.*, vol. 45, no. 6, pp. 1075–1083, Dec. 2012.

[8] H. Xu, P. D. Stetson, and C. Friedman, “Combining corpus-derived sense profiles with estimated frequency information to disambiguate clinical abbreviations,” *Amia Annu. Symp. Proc. Amia Symp. Amia Symp.*, vol. 2012, pp. 1004–1013, 2012.

[9] S. Moon, S. Pakhomov, and G. B. Melton, “Automated disambiguation of acronyms and abbreviations in clinical texts: window and training size considerations,” *Amia Annu. Symp. Proc. Amia Symp. Amia Symp.*, vol. 2012, pp. 1310–1319, 2012.

[10] Y. Kim, J. Hurdle, and S. M. Meystre, “Using UMLS lexical resources to disambiguate abbreviations in clinical text,” *Amia Annu. Symp. Proc. Amia Symp. Amia Symp.*, vol. 2011, pp. 715–722, 2011.

[11] C. Friedman, P. O. Alderson, J. H. Austin, J. J. Cimino, and S. B. Johnson, “A general natural-language text processor for clinical radiology,” *J Am Med Inf. Assoc*, vol. 1, pp. 161–74, Mar. 1994.

[12] A. R. Aronson, “Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program,” 20020204.

[13] J. C. Denny, P. R. Irani, F. H. Wehbe, J. D. Smithers, and A. Spickard, “The KnowledgeMap project: development of a concept-based medical school curriculum database,” *Amia Annu Symp Proc*, pp. 195–9, 2003.

[14] J. J. Savova Gk Fau - Masanz, P. V. Masanz Jj Fau - Ogren, J. Ogren Pv Fau - Zheng, S. Zheng J Fau - Sohn, K. C. Sohn S Fau - Kipper-Schuler, C. G. Kipper-Schuler Kc Fau - Chute, and C. G. Chute, “Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications,” 20100907.

[15] Y. Wu, J. C. Denny, S. T. Rosenbloom, R. A. Miller, D. A. Giuse, and H. Xu, “A comparative study of current Clinical Natural Language Processing systems on handling abbreviations in discharge summaries,” *Amia Annu. Symp. Proc. Amia Symp. Amia Symp.*, vol. 2012, pp. 997–1003, 2012.

[16] H. Suominen, S. Salantera, and S. Velupillai, “Overview of the ShARe/CLEF eHealth Evaluation Lab 2013,” in *Proceedings of CLEF 2013*.

[17] M. J. Schuemie, J. A. Kors, and B. Mons, “Word sense disambiguation in the biomedical domain: an overview,” *J Comput Biol*, vol. 12, pp. 554–65, Jun. 2005.

[18] S. Pakhomov, T. Pedersen, and C. G. Chute, “Abbreviation and acronym disambiguation in clinical discourse,” *Amia Annu Symp Proc*, pp. 589–93, 2005.

[19] X. Hua, D. S. Peter, and F. Carol, “Combining Corpus-derivedSense Profiles with Estimated Frequency Information to Disambiguate Clinical Abbreviations,” *Submitt. Amia 2012*.

[20] H. Liu, V. Teller, and C. Friedman, “A multi-aspect comparison study of supervised word sense disambiguation,” *J Am Med Inf. Assoc*, vol. 11, pp. 320–31, Jul. 2004.