AMIA
INFORMATICS PROFESSIONALS. LEADING THE WAY.

OXFORD

Research and Applications

A long journey to short abbreviations: developing an open-source framework for clinical abbreviation recognition and disambiguation (CARD)

Yonghui Wu,[1] Joshua C Denny,[2,3] S Trent Rosenbloom,[2,3] Randolph A Miller,[2,3] Dario A Giuse,[2] Lulu Wang,[3] Carmelo Blanquicett,[4] Ergin Soysal,[1] Jun Xu,[1] and Hua Xu[1]

[1]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, [2]Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, Tennessee, [3]Department of Medicine, Vanderbilt University School of Medicine, and [4]Department of Medicine, University of Alabama at Birmingham, Birmingham

Corresponding Author: Hua Xu, PhD, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, 7000 Fannin St., Suite 600, Houston, TX 77030 USA. E-mail: hua.xu@uth.tmc.edu; Tel: 713-500-3924

ABSTRACT

Objective: The goal of this study was to develop a practical framework for recognizing and disambiguating clinical abbreviations, thereby improving current clinical natural language processing (NLP) systems' capability to handle abbreviations in clinical narratives.

Methods: We developed an open-source framework for clinical abbreviation recognition and disambiguation (CARD) that leverages our previously developed methods, including: (1) machine learning based approaches to recognize abbreviations from a clinical corpus, (2) clustering-based semiautomated methods to generate possible senses of abbreviations, and (3) profile-based word sense disambiguation methods for clinical abbreviations. We applied CARD to clinical corpora from Vanderbilt University Medical Center (VUMC) and generated 2 comprehensive sense inventories for abbreviations in discharge summaries and clinic visit notes. Furthermore, we developed a wrapper that integrates CARD with MetaMap, a widely used general clinical NLP system.

Results and Conclusion: CARD detected 27 317 and 107 303 distinct abbreviations from discharge summaries and clinic visit notes, respectively. Two sense inventories were constructed for the 1000 most frequent abbreviations in these 2 corpora. Using the sense inventories created from discharge summaries, CARD achieved an F1 score of 0.755 for identifying and disambiguating all abbreviations in a corpus from the VUMC discharge summaries, which is superior to MetaMap and Apache's clinical Text Analysis Knowledge Extraction System (cTAKES). Using additional external corpora, we also demonstrated that the MetaMap-CARD wrapper improved MetaMap's performance in recognizing disorder entities in clinical notes. The CARD framework, 2 sense inventories, and the wrapper for MetaMap are publicly available at https://sbmi.uth.edu/ccb/resources/abbreviation.htm. We believe the CARD framework can be a valuable resource for improving abbreviation identification in clinical NLP systems.

Key words: clinical abbreviation, sense clustering, machine learning, clinical natural language processing

INTRODUCTION

Health care providers frequently use abbreviations in clinical notes as a convenient way to represent long biomedical words and phrases.[1,2] These abbreviations often contain important clinical information (eg, names of diseases, drugs, or procedures) that must be recognizable and accurate in health records.[3,4] However,

understanding these clinical abbreviations is still a challenging task for current clinical natural language processing (NLP) systems, primarily for 2 reasons: a lack of comprehensive lists of abbreviations and their possible senses (sense inventory), and a lack of efficient methods to disambiguate abbreviations that have multiple senses.[5] Previous studies have explored different aspects of handling clinical abbreviations, such as building abbreviations and sense inventories from medical vocabularies or clinical corpora[2,6,7] and disambiguation methods.[8–11] Although these studies provided valuable insights for handling clinical abbreviations, few practical systems have been developed to improve the limited capability of current clinical NLP systems to identify abbreviations.

The goal of this study was to develop a practical, corpus-driven system to handle abbreviations in clinical documents. Previously, we developed a machine learning–based method to detect abbreviations in a clinical corpus,[12] a semisupervised method to build abbreviation sense inventories based on clustering and manual review,[7] and a profile-based method for abbreviation sense disambiguation.[10] In this study, we reimplemented the above algorithms and methods in Java and integrated them into a single open-source framework, called clinical abbreviation recognition and disambiguation (CARD). To demonstrate the utility of CARD, we applied it to Vanderbilt University Medical Center (VUMC)'s discharge summaries and clinic visit notes and constructed 2 comprehensive sense inventories that contain frequent abbreviations and their possible senses in each corpus. Using annotated abbreviation corpora from VUMC and the 2013 Shared Annotated Resources (ShARe)/Conference and Labs of the Evaluation Forum (CLEF) challenge, we demonstrated that CARD could improve the performance (F1 score) of Meta-Map[13] on handling abbreviations (F1 score 0.338–0.755 on the VUMC corpus). In addition, we applied CARD to a common task of disease named entity recognition (NER), and it improved the performance of NER on both external corpora, indicating the generalizability of the CARD framework. To the best of our knowledge, this is the first open-source practical system for clinical abbreviation recognition and disambiguation.

BACKGROUND

Large quantities of clinical documentation are available in electronic health record (EHR) systems, and they typically contain valuable patient information. NLP technologies have been applied to the medical domain to identify useful information in clinical notes, with the goals of improving patient care and facilitating clinical research.[1] Many clinical NLP systems have been developed to extract various medical concepts from clinical narratives, such as MedLEE,[14] MetaMap,[13,15] cTAKES,[16] KnowledgeMap,[17] MedTagger,[18] MedEx,[19] etc. Despite the success of these systems in extracting general clinical concepts, one of the remaining challenges is to accurately identify abbreviations in clinical text. Clinical abbreviations are shortened terms in clinical text, and may include acronyms and other shortened words or symbols. A recent study comparing 3 widely used clinical NLP systems, cTAKES, MetaMap, and MedLEE, on recognizing and normalizing abbreviations in hospital discharge summaries showed limited performance (ie, with F measures ranging from 0.17 to 0.60).[5] The ShARe/CLEF eHealth Evaluation Lab in 2013 organized a shared task on identifying abbreviations in clinical text, and the best performing team achieved an accuracy of 71.9%, indicating that abbreviation recognition and normalization is still a challenge in clinical NLP research.[20]

Two issues contribute to the challenges associated with handling clinical abbreviations. First, no complete list covering all clinical abbreviations and their possible senses currently exists. This is in part because many clinical abbreviations are invented in an ad hoc manner by health care providers during practice. Existing knowledge bases, such as the Unified Medical Language System (UMLS)[21] Metathesaurus and the SPECIALIST Lexicon, have low coverage rates for abbreviations used in clinical texts. Pakhomov et al. reported that only 26% of senses from 8 randomly selected acronyms were covered by a Mayo Clinic–approved acronym list, and only 25% were covered by the UMLS 2005AA[45] Biomedical researchers have manually constructed useful abbreviation lists, such as a list of over 12 000 abbreviations and their senses from pathology reports by Berman[2] and a list of over 3000 clinical abbreviations extracted from physicians' sign-out notes by Stetson et al.[22] It is time- and resource-consuming to build such knowledge bases of clinical abbreviations manually. Recently, we proposed a more feasible approach to building clinical abbreviation sense inventories, which first detects candidate abbreviations using machine learning methods[12] and then finds possible senses of an abbreviation using clustering-based methods followed by manual review.[23] Using a set of 70 hospital discharge summary notes, we demonstrated that the machine learning–based abbreviation detection method could achieve an F1 score of 95.7% (with a precision of 97% and recall of 94.5%). For abbreviation sense detection, our clustering-based approach showed that we could potentially save up to 50% in annotation cost, as well as detect more rare senses of abbreviations in a data set containing 13 frequent clinical abbreviations.[7] However, despite the promise of these approaches, they were only evaluated on small data sets and have not been used to generate comprehensive abbreviation and sense lists from large clinical corpora.

The second challenge for handling clinical abbreviations is that they can be ambiguous (eg, "AA" can mean African American, abdominal aorta, Alcoholics Anonymous, etc.).[4] Resolving word ambiguity is a fundamental task of NLP research, called word sense disambiguation (WSD).[24] Investigators have developed different methods to address this problem, including the majority sense approach, literature concept co-occurrence methods, semantic and linguistic rules, knowledge-based methods, and supervised machine learning methods to disambiguate biomedical words.[15,17,25–27] A number of studies have focused on clinical abbreviation disambiguation.[28–31] Many studies have shown that supervised machine learning–based WSD methods can achieve good performance. Most recently, there have been studies exploring more advanced features, such as word embeddings trained from neural networks.[32] However, supervised machine learning approaches are not very practical because they require an annotated training set for each ambiguous abbreviation.[11] To reduce the cost of annotating training corpora, Pakhomov et al.[29] proposed a semisupervised method to automatically create pseudo-annotated data sets by replacing long forms with corresponding abbreviations, which showed promising results. We and others also demonstrated that profile-based approaches that build on the vector space model work better on such pseudo-corpora.[10,29] However, all these approaches to clinical abbreviation disambiguation have been evaluated using only small data sets that contain limited numbers of clinical abbreviations.

Currently, no robust abbreviation identification modules are widely used in existing clinical NLP systems. While some systems can identify abbreviations together with their in-line definitions,[17] few can accurately identify ad hoc abbreviations without in-document definitions. The goal of this study was to develop a
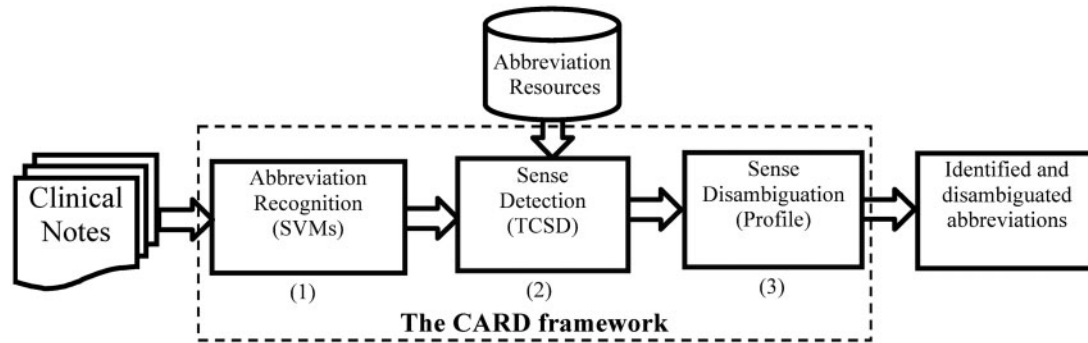
Figure 1. An overview of the CARD framework.

practical framework for clinical abbreviation recognition and disambiguation, by integrating state-of-the-art methods in handling clinical abbreviations. The contributions of this work are 3-fold. First, we developed CARD, an open-source software for handling clinical abbreviations. Users can apply it to their own corpora to detect abbreviations, create sense inventories, and build disambiguation methods. Second, by applying the CARD framework to Vanderbilt's clinical corpora, we created 2 comprehensive lists of frequent abbreviations and their senses for discharge summaries and clinic visit notes, which are valuable resources for clinical NLP research. Third, we integrated CARD with MetaMap to provide a convenient wrapper for end users of clinical NLP systems. We believe this work can improve the state-of-the-art practice of clinical NLP systems on handling abbreviations.

METHODS

Developing the CARD framework
As shown in Figure 1, the CARD framework consists of 3 components that build on proven methods for handling clinical abbreviations. The first is the abbreviation recognition module, which detects abbreviations from clinical corpora using machine learning algorithms.[12] The second is the abbreviation sense detection module, which finds possible senses of an abbreviation in a given corpus using a clustering-based approach.[7] The third is the abbreviation sense disambiguation module, which determines the correct meaning of an abbreviation in a given sentence based on sense profiles and the vector space model.[10] For a given clinical corpus, CARD will generate the following outputs: (1) a list of clinical abbreviations in the corpus, (2) a corpus-specific sense inventory of selected abbreviations, and (3) an abbreviation disambiguation tool, thus providing an end-to-end solution for handling abbreviations in the corpus. CARD is implemented in Java and available as an open-source package (https://sbmi.uth.edu/ccb/resources/abbreviation.htm). The following sections describe more details for each component of CARD.

Abbreviation recognition module
This module classifies whether a word in a given corpus is an abbreviation or not. In a previous study, we investigated 3 different machine-learning algorithms, support vector machines (SVMs), random forests (RFs), and decision trees, to detect abbreviations. Our results showed that the SVM classifier achieved the best performance.[12] Therefore, we implemented an SVM-based abbreviation recognizer in CARD using the libSVM package.[33] We also leveraged existing lists of known clinical abbreviations (eg, LRABR in the UMLS) to ensure that we did not miss important abbreviations already covered by existing abbreviation databases.

Sense detection module
This module finds all possible senses of each abbreviation (called a "sense inventory") according to the context and discourse information in the corpus. We have developed a semiautomatic method for building corpus-specific sense inventories for clinical abbreviations by combining a sense clustering algorithm with manual review of the centroid of each sense cluster.[7] For CARD, we reimplemented this method in Java and developed a web-based graphic interface to facilitate annotation of senses of clusters.

Sense disambiguation module
This module first recognizes abbreviations in a clinical document by looking up the abbreviation list generated by the first module. If an abbreviation has only 1 sense (according to the sense inventory generated by the second module), it assigns the sense to the abbreviation directly. Otherwise, it runs the sense disambiguation algorithm to determine the correct sense for the ambiguous abbreviation. Currently, CARD implements a profile-based WSD method[10] to disambiguate clinical abbreviations, because previous studies showed that such methods are robust.[10,29]

Applying CARD to VUMC's clinical corpora
We applied the CARD framework to 2 types of corpora at VUMC: discharge summaries and clinic visit notes. Using the standard CARD pipelines, we detected all abbreviations from each corpus, generated sense clusters and manually reviewed sense inventories for the top 1000 most frequent abbreviations, and built sense profiles for abbreviation sense disambiguation. We then evaluated the performance of the CARD framework on detecting and disambiguating abbreviations using a different annotated corpus from VUMC's.

Data sets
We used clinical documents taken from VUMC's synthetic derivative[34] database, which contains deidentified copies of EHR documents. From the synthetic derivative, we collected 3 years' worth of discharge summaries (2007–2009; a total of 123 064 documents) and clinic visit notes (2 628 169 documents) and used these 2 corpora to generate abbreviation lists and sense inventories. This study was approved by the Vanderbilt Institutional Review Board.

Figure 2. Web-based interface for sense annotation.

Abbreviation detection

To detect abbreviations from discharge summaries and clinic visit notes, we adopted an SVM classifier developed in one of our previous studies.[12] The classifier was trained using 70 discharge notes, and it showed an F1 score of 95.7% on detecting abbreviations from discharge summaries. However, when applying this classifier to the clinic visit notes, we found that there were new abbreviation patterns. Thus, we further annotated 30 additional clinic visit notes and retrained the model.

Sense clustering

After detecting the abbreviations, we identified the top 1000 most frequent abbreviations to build sense inventories. For each of these, we randomly sampled up to 1000 sentences from the corpus. The TCSD algorithm[7] was used to group these sentences into clusters, where sentences containing the same sense of an abbreviation were assigned together. We controlled the parameters to generate up to 20 sense clusters for each abbreviation.

Sense annotation

We collected the centroid sentences of each cluster and displayed them in a web-based annotation interface with the target abbreviation highlighted. Figure 2 shows the annotation interface. The experts then manually reviewed each sentence and annotated the sense according to annotation guidelines. We adopted a 3-round annotation workflow to speed up the annotation. In the first round, we recruited 2 third-year medical school students (L.W. and C.B.) to pre-annotate the abbreviations that they could understand with high confidence; they left blank the abbreviations they were not confident about. Three physicians with different clinical expertise (J.D., S.T.R., and R.A.M.) served as the second-round annotators. They manually reviewed pre-annotated abbreviations to resolve incomplete ones and corrected the pre-annotations as needed. For the instances that remained unsolved after individual review by the second-round annotators, we discussed their meanings in the third round in a group meeting among all annotators.

Annotation normalization and sense frequency calculation

We collected all the annotated senses and manually reviewed them to normalize the following variations: (1) plural/singular: normalize all plurals to singular, eg, milligrams to milligram; (2) synonyms: normalize each synonym to a commonly used one, eg, computed tomography angiogram to computed tomographic angiography; (3) tense: normalize different word tenses to the word's lemma, eg, discontinued to discontinue. Then we estimated the sense frequency based on the annotated senses. Finally, we mapped the normalized senses into the most specific UMLS concept unique identifiers (CUIs) by leveraging existing clinical NLP systems (MetaMap and KnowledgeMap[17]) and with help from a domain expert (E.S.).

Integrating CARD with MetaMap

The CARD framework is an independent package for clinical abbreviation recognition and disambiguation. To promote its use with other existing general clinical NLP systems, we developed a wrapper of CARD for the widely used MetaMap system. The wrapper reads MetaMap outputs and updates them using CARD's results. In this way, researchers can take advantage of the CARD framework to recognize and disambiguate clinical abbreviations without modifying their existing pipelines that use MetaMap.

Evaluation

To assess CARD's end-to-end performance on detecting and disambiguating clinical abbreviations, we evaluated it using 2 abbreviation corpora: an existing VUMC corpus that was built in our previous research, which consists of 32 discharge summaries, with abbreviations and their senses annotated by 3 domain experts,[5] and one from a different institution, the abbreviation corpus from the 2013 SHARe/CLEF challenge,[20] which was constructed from the Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC II) corpus.[35] There is no overlap between VUMC's evaluation data set and the data sets used to train the CARD system. In this evaluation, we reported the performances of 4 systems: MetaMap, cTAKES, CARD, and the MetaMap-CARD wrapper (we did not include the MedLEE system due to license expiration). Standard measurements (micro-averages) including precision (defined as the ratio between the number of correctly detected abbreviations and the total number of detected abbreviations by the system), recall (defined as the ratio between the number of abbreviations correctly detected by the system and the total number of abbreviations in the gold standard), and F1 score (calculated as 2*precision*recall/(pre-

cision + recall)) were reported. Table 1 shows the statistics of the 2 abbreviation corpora.

In addition, we also evaluated the utility of CARD in a common NER task, recognizing disorder entities in clinical documents, using 2 external NER data sets. The first is the NER corpus from the 2014 Semantic Evaluation (SemEval) challenge Task 7,[36] which annotated all disorder entities and their corresponding UMLS CUIs. The other is a newly created NER data set that contains 45 clinic visit notes from University of Texas Health Science Center at Houston (UTHealth) EHRs, with disorder entities and their UMLS CUIs annotated by domain experts. Please refer to Table 2 for detailed information about the 2 NER data sets. Using the 2014 SemEval challenge evaluation script, we reported the standard precision, recall, and F1 score for NER, as well as the accuracy for encoding to UMLS CUIs. For baselines, we included the results on all entities and abbreviations only from the MetaMap 2012 and Apache cTAKES 3.2.2 in this study.

RESULTS

We constructed 2 comprehensive abbreviation sense inventories from the VUMC discharge summaries and clinic visit notes. Table 3 shows a comparison of the 2 sense inventories. The CARD framework identified more potential abbreviations in the clinic visit notes than in the discharge summaries (107 303 compared to 27 317). For the top 1000 most frequent abbreviations, the sense clustering algorithm detected 20 101 clusters from clinic visit notes and 15 536 clusters from discharge summaries. The annotation time varied among abbreviations and annotators. It took approximately 2–3 weeks to finish the first-round annotation. After normalizing the annotations, the sense inventory from discharge summaries contained 915 abbreviations with 1299 unique senses, and the sense inventory from clinic visit notes contains 954 abbreviations with 1499 unique senses, showing that the abbreviations in clinic visit notes may have more diverse meanings. The abbreviations from the 2 sense inventories accounted for 92.5% and 93.4%, respectively, of the total abbreviation occurrences in both types of notes. Table 4 shows some examples of the sense inventory files, which contain not only the senses, but also the CUIs of the senses and the estimated sense frequencies.

Table 5 compares the performances of MetaMap, cTAKES, CARD, and MetaMap-CARD wrapper on the VUMC corpus and the SHARe/CLEF abbreviation corpus. CARD outperformed MetaMap and cTAKES for recognition and disambiguation of clinical abbreviations on both corpora. After integrating with CARD, the performance of MetaMap was significantly improved, with F scores increasing from 0.338 to 0.761 on the VUMC corpus and from 0.210 to 0.332 on the SHARe/CLEF corpus. Table 6 compares the performances of MetaMap, cTAKES, and MetaMap-CARD wrapper on extracting disorder entities from the SemEval corpus and the UTHealth corpus. After integrating with CARD, the performance of MetaMap for identifying disorder mentions (encoding accuracy) improved 1.1% on the SemEval corpus and 2.9% on the UTHealth

Table 1. Statistics on 2 abbreviation corpora used in the evaluation

Data set name	# Notes	Note types	Unique[a]	Occurrences[b]
VUMC	32	Discharge summary	332	1112
SHARe/CLEF	300	Discharge Radiology ECG ECHO	1016	7579

[a]Unique: the number of unique abbreviations.
[b]Occurrences: the total number of occurrences of all abbreviations.

Table 2. Statistics on 2 disorder NER data sets

Dataset	No. of Notes	Note types	Total no. of entities	No. of Abbreviations
SemEval	431	Discharge Radiology ECG ECHO	19 115	1701
UTHealth	45	Clinic visit	3391	428

Table 3. Sense inventories from discharge notes and clinic visit notes

Source	Notes	Potential abbreviation	Selected for inventory	Occurrence coverage (%)	Sense cluster	Valid abbreviation	Identified sense
DS	123 067	27 317	1000	93.4	15 536	915	1299
CV	2 628 169	107 303	1,000	92.5	20 101	954	1499

corpus. The accuracy of recognizing and disambiguating abbreviations improved only 10.2% and 18.5% on the SemEval corpus and UTHealth corpus, respectively.

DISCUSSION

In this paper, we discuss the development of CARD, an open-source framework for abbreviation recognition and disambigua-

Table 4. Examples from the sense inventories

	Sense	CUI	Prevalence (%)
LL	Lithotripsy	C0023878	55.2
	Lower lobe	C0225758	15.1
	Left leg	C1279606	13.1
	Left lower	C0442068	6.30
	Balloon	C1704777	3.33
	Lower lumbar	NA	2.28
	Two	C0205448	1.93
	Lower leg	C1140621	1.40
	Left lateral	NA	1.05
	Overall	C1561607	0.35
MAC	Monitored anesthesia care	C0497677	68.8
	Mycobacterium avium complex	C0026914	16.6
	Macular	C0332574	8.60
	Macaroni	C0452696	6.00

NA: There is no single UMLS CUI for this sense.

Table 5. Results of MetaMap, cTAKES, CARD, and MetaMap-CARD wrapper on identifying abbreviations in the corpora of VUMC and SHARe/CLEF

System	VUMC corpus			SHARe/CLEF corpus		
	Precision	Recall	F1 score	Precision	Recall	F1 score
MetaMap	0.482	0.260	0.338	0.402	0.142	0.210
Ctakes	0.285	0.116	0.165	0.368	0.107	0.166
CARD	0.910	0.645	0.755	0.503	0.204	0.291
MetaMap-CARD wrapper	0.819	0.710	0.761	0.434	0.269	0.332

tion. By applying CARD to VUMC discharge summaries and clinic visit notes, we constructed 2 sense inventories of the top 1000 most frequent abbreviations in each corpus, with each sense mapped to UMLS CUIs and its frequency estimated by sense clustering analysis. Our evaluation using the VUMC corpus and the SHARe/CLEF corpus showed that CARD could achieve better performance on abbreviation identification compared with the general NLP systems MetaMap and cTAKES. After integrating CARD with MetaMap, we demonstrated improved performance of MetaMap on the disorder NER task. We believe CARD, as well as the resources generated by this study (eg, the sense inventories), may enable current clinical NLP systems to better interpret abbreviations.

Existing clinical NLP systems have limited capability to deal with abbreviations in clinical documents from specific institutions. The CARD framework provides a practical solution to develop local sense inventories and customized disambiguation methods for each institution. As shown in Table 4, when applying Meta-Map and cTAKES directly to VUMC discharge summaries, we achieved low performance on identification of abbreviations (F1 score 0.338 and 0.165, respectively). However, CARD could achieve an F1 score of 0.755 by using sense inventories and disambiguation profiles generated from local corpora. This finding indicates a difference in abbreviations among notes from different institutions and the need to develop local sense inventories and disambiguation methods for each institution. CARD improves performance in both recall and precision. The locally generated abbreviation and sense lists improved the recall of the system. Using a set of 70 discharge summaries manually annotated in a previous study,[12] we found that this sense inventory achieved a coverage of 83% in terms of abbreviation occurrences (2347/2827), and a coverage of 81% in terms of sense occurrences (2279/2827). If we consider only the abbreviations covered by the sense inventory, the sense coverage goes up to 97% (2279/2347). On the other hand, existing abbreviation databases covered only around 60% of abbreviations, as reported by previous research.[6] The improved precision of CARD is mainly from the profile-based disambiguation approach, which utilizes local corpora to generate sense profiles. Thus, it is not surprising that CARD worked better on the VUMC corpus than it did on the SHARe/CLEF corpus (Table 5), as it uses the sense inventories from VUMC. This result suggests that it would be ideal to develop local sense inventories and sense profiles, if possible.

Table 6. Results of MetaMap, cTAKES, and MetaMap-CARD wrapper on recognizing and encoding disorder entities using the annotated corpora from the 2014 SemEval Task 7 and UTHealth

System	SemEval		UTHealth	
	NER F1 (Pre/Rec)	Encoding Acc	NER F1 (Pre/Rec)	Encoding Acc
All entities				
MetaMap	0.566 (0.654/0.499)	0.442	0.515 (0.567/0.471)	0.384
cTAKES	0.411(0.323/ 0.564)	0.439	0.404(0.359/ 0.461)	0.385
MetaMap-CARD	0.576 (0.648/0.518)	0.453	0.535 (0.568/0.506)	0.413
Abbreviations only				
MetaMap	0.494 (0.741/0.371)	0.327	0.552 (0.683/0.463)	0.238
cTAKES	0.468(0.571/0.396)	0.259	0.295(0.619/0.194)	0.124
MetaMap-CARD	0.601 (0.541/0.677)	0.429	0.719 (0.715/0.722)	0.423

We also examined the majority sense strategy using the estimated sense frequency from the sense inventory constructed from discharge summaries. Using the majority sense method, the CARD system achieved an end-to-end F1 score of 0.744 on the VUMC corpus, suggesting that the estimated majority sense from CARD is very helpful in sense disambiguation. After examining the estimated sense distribution, we found that 83% of the ambiguous abbreviations (239) had a majority sense with a relative frequency over 70%, which explains the good performance of the majority sense strategy. This finding is consistent with our previous WSD study on a small set of clinical abbreviations.[31] However, the majority sense of an abbreviation is not always available, and it can vary in clinical notes from different subdomains of medicine or from different institutions, which prohibits wide use of this simple but effective method.

In this study, we found that different types of clinical documents demonstrated different abbreviation patterns, even when they were from the same institution. We compared the 2 abbreviation sense inventories from discharge summaries and clinic visit notes from VUMC and noticed that only 663 abbreviations occurred in both inventories. For abbreviations occurring in both types of notes, they often had different sense distributions. For the 663 overlapping abbreviations, there were 991 senses found in discharge summaries, whereas 1049 senses were found in the clinic visit notes. This finding indicates the need to develop corpus-specific sense inventories in order to achieve optimal performance.

Another interesting finding is about the comprehensibility of abbreviations in clinical documents among physicians. Out of the 15 536 sense clusters derived from VUMC discharge summaries, our physician reviewers labeled 344 of them (2.2%) as "unknown" after 3 rounds of annotation, denoting that some abbreviations are difficult to understand even for physicians in the same institution. We examined these unknowns and found that most of them were abbreviations with limited contextual information to indicate their meaning (eg, EOSIRE, NC: lgl). Additionally, these annotations were performed in the context of only a single note in front of the physician, and perhaps would be interpretable with access to the entire medical record. A potential solution for this problem would be to develop techniques that can correctly expand abbreviations while physicians type them in at the time of document entry.

CARD has limitations that can further guide development. For abbreviation detection, the long tail of less frequent abbreviations needs further investigation. One solution is to leverage the existing abbreviation knowledge base to help determine the inclusion of low-frequency abbreviations. Moreover, even with the clustering approach, manual annotation of the centroid of each sense cluster would still be time-consuming. Methods that can automatically link sense clusters to corresponding senses in a knowledge base (eg, CUIs in the UMLS) would be very helpful. In addition to MetaMap, we also plan to build CARD wrappers for other popular clinical NLP systems such as cTAKES.

CONCLUSION

In this study, we developed and implemented an open-source framework for clinical abbreviation recognition and disambiguation. We demonstrated how CARD can be used to generate corpus-specific sense inventories and how it could improve the performance of an existing NLP system (MetaMap) on recognition and disambiguation of clinical abbreviations, thereby improving its performance on the disorder NER task. The CARD framework, 2 comprehensive abbreviation sense inventories, and the wrapper for MetaMap are publicly available at https://sbmi.uth.edu/ccb/resources/abbreviation.htm. We believe this tool will benefit the NLP community by improving current practice on understanding clinical abbreviations and other related clinical NLP tasks.

COMPETING INTERESTS

None.

CONTRIBUTORS

Y.W. and H.X. were responsible for the overall design, development, and evaluation of this study. J.D., S.T.R., and R.A.M. worked on the annotation guideline and also served as second-round annotators as domain experts. L.W. and C.B. worked on the first-round annotation. D.G. contributed to the annotation guideline and third-round annotation. ES reviewed the sense inventory and mapped the senses to UMLS CUI. J.X. contributed to the evaluation of CARD. Y.W. and H.X. did the bulk of the writing; J.D., S.T.R., and D.G. also contributed to writing and editing this manuscript. All authors reviewed the manuscript critically for scientific content, and all authors gave final approval for publication.

REFERENCES

1. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform*. 2008;35:128–144.
2. Berman JJ. Pathology abbreviated: a long review of short terms. *Arch Pathol Lab Med*. 2004;128(3):347–352.
3. Xu H, Stetson PD, Friedman C. A study of abbreviations in clinical notes. *AMIA Annu Symp Proc*. 2007;821–5.
4. Sheppard JE, Weidner LC, Zakai S, Fountain-Polley S, Williams J. Ambiguous abbreviations: an audit of abbreviations in paediatric note keeping. *Arch Dis Childhood*. 2008;93(3):204–206.
5. Wu Y, Denny JC, Rosenbloom ST, Miller RA, Giuse DA, Xu H. A comparative study of current Clinical Natural Language Processing systems on handling abbreviations in discharge summaries. *AMIA Annu Symp Proc*. 2012;2012:997–1003.
6. Liu H, Lussier YA, Friedman C. A study of abbreviations in the UMLS. *Proc AMIA Symp*. 2001:393–397.
7. Xu H, Wu Y, Elhadad N, Stetson PD, Friedman C. A new clustering method for detecting rare senses of abbreviations in clinical notes. *J Biomed Inform*. 2012;45(6):1075–1083.
8. Wu Y, Tang B, Jiang M, Moon S, Denny C, Xu H. Clinical acronym/abbreviation normalization using a hybrid approach. *Proceedings of CLEF*. 2013;2013:2013.
9. Moon S, Berster B, Xu H, Cohen T. Word sense disambiguation of clinical abbreviations with hyperdimensional computing. *AMIA Annu Symp Proc*. 2013:1007–1016.

10. Xu H, Stetson PD, Friedman C. Combining corpus-derived sense profiles with estimated frequency information to disambiguate clinical abbreviations. *AMIA Annu Symp Proc*. 2012;2012:1004–1013.

11. Liu H, Teller V, Friedman C. A multi-aspect comparison study of supervised word sense disambiguation. *J Am Med Inform Assoc*. 2004;11(4): 320–331.

12. Wu Y, Rosenbloom ST, Denny JC, et al. Detecting abbreviations in discharge summaries using machine learning methods. *AMIA Annu Symp Proc*. 2011;2011:1541–1549.

13. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp*. 2001:17–21.

14. Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. *J Am Med Inform Assoc*. 1994;1(2):161–174.

15. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc*. 2010;17(3):229–236.

16. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc*. 2010;17(5):507–513.

17. Denny JC, Irani PR, Wehbe FH, Smithers JD, Spickard A 3rd. The KnowledgeMap project: development of a concept-based medical school curriculum database. *AMIA Annu Symp Proc*. 2003:195–199.

18. Wagholikar KB, Torii M, Jonnalagadda SR, Liu H. Pooling annotated corpora for clinical concept extraction. *J Biomed Semantics*. 2013;4(1):3.

19. Jiang M, Wu Y, Shah A, Priyanka P, Denny C, Xu H. Extracting and standardizing medication information in clinical text: the MedEx-UIMA system. *2014 Summit on Clinical Research Informatics*. 2014;37–42.

20. Suominen H, Salanterä S, Velupillai S, et al. Overview of the ShARe/CLEF eHealth Evaluation Lab 2013. In: P Forner, H Müller, R Paredes, P Rosso, B Stein, eds. *Information Access Evaluation Multilinguality, Multimodality, and Visualization*. Berlin Heidelberg: Springer; 2013:212-231.

21. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res*. 2004;32(Database issue): D267–D270.

22. Stetson PD, Johnson SB, Scotch M, Hripcsak G. The sublanguage of cross-coverage. *Proc AMIA Symp*. 2002:742–746.

23. Xu H, Stetson PD, Friedman C. Methods for building sense inventories of abbreviations in clinical notes. *J Am Med Inform Assoc*. 2009;16(1):103–108.

24. Navigli R. Word sense disambiguation: a survey. *ACM Comput Surv*. 2009; 41(2):1–69.

25. Schuemie MJ, Kors JA, Mons B. Word sense disambiguation in the biomedical domain: an overview. *J Comput Biol*. 2005;12(5):554–565.

26. Xu H, Markatou M, Dimova R, Liu H, Friedman C. Machine learning and word sense disambiguation in the biomedical domain: design and evaluation issues. *BMC Bioinformatics*. 2006;7:334.

27. Stevenson M, Guo Y. Disambiguation in the biomedical domain: the role of ambiguity type. *J Biomed Inform*. 2010;43(6):972–981.

28. Liu H, Johnson SB, Friedman C. Automatic resolution of ambiguous terms based on machine learning and conceptual relations in the UMLS. *J Am Med Inform Assoc*. 2002;9(6):621–636.

29. Pakhomov S, Pedersen T, Chute CG. Abbreviation and acronym disambiguation in clinical discourse. *AMIA Annu Symp Proc*. 2005:589–593.

30. Moon S, Pakhomov S, Melton GB. Automated disambiguation of acronyms and abbreviations in clinical texts: window and training size considerations. *AMIA Annu Symp Proc*. 2012;2012:1310–1319.

31. Wu Y, Denny JC, Rosenbloom ST, et al. A preliminary study of clinical abbreviation disambiguation in real time. *Applied Clin Inform*. 2015;6(2): 364–374.

32. Wu Y, Xu J, Zhang Y, Xu H. Clinical abbreviation disambiguation using neural word embeddings. *ACL-IJCNLP* 2015;2015:171.

33. Chang C-JL. LIBSVM: a library for support vector machines. Available at: http://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf.

34. Roden DM, Pulley JM, Basford MA, et al. Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin Pharmacol Ther*. 2008;84(3):362–369.

35. Saeed M, Villarroel M, Reisner AT, et al. Multiparameter intelligent monitoring in intensive care II: a public-access intensive care unit database. *Crit Care Med*. 2011;39(5):952–960.

36. Pradhan S, Elhadad N, Chapman W, Manandhar S, Savova G. SemEval-2014 Task 7: analysis of clinical text. *SemEval*. 2014 2014;199(99):54.