Exploiting Task-Oriented Resources to Learn Word Embeddings for Clinical Abbreviation Expansion

Yue Liu\textsuperscript{1}, Tao Ge\textsuperscript{2}, Kusum Mathews\textsuperscript{3}, Heng Ji\textsuperscript{1}, Deborah L. McGuinness\textsuperscript{1}

\textsuperscript{1}Department of Computer Science, Rensselaer Polytechnic Institute
\{liuy30,jih,dlm\}@rpi.edu

\textsuperscript{2}School of Electronics Engineering and Computer Science, Peking University
getao@pku.edu.cn

\textsuperscript{3}Icahn School of Medicine, The Mount Sinai Hospital
kusum.mathews@mssm.edu

Abstract

In the medical domain, identifying and expanding abbreviations in clinical texts is a vital task for both better human and machine understanding. It is a challenging task because many abbreviations are ambiguous especially for intensive care medicine texts, in which phrase abbreviations are frequently used. Besides the fact that there is no universal dictionary of clinical abbreviations and no universal rule of abbreviation writing, such texts are difficult to acquire, expensive to annotate and even sometimes, confusing to domain experts. This paper proposes a novel and effective approach – exploit task-oriented resources to learn word embeddings for expanding abbreviations in clinical notes. We achieved 82.27% accuracy, close to human performance.

1 Introduction

Abbreviations appear frequently in the medical domain. Based on a popular online knowledge base, among the 3,096,346 stored abbreviations, 197,787 records are Medical abbreviations, ranked 1st among all 10 domains.\textsuperscript{1} An abbreviation can have over 100 possible explanations\textsuperscript{2} even within the medical domain. The authors of clinical texts, which are mainly doctors, health professionals and domain experts, usually write messages under time pressure and therefore frequently compress some words or phrases. Regarding intensive care medicine texts, the critical setting requires information to be expressed in an efficient way, resulting in code-like messages with poor readability. For example, given a sentence written by a fellow with training in critical care medicine, “STAT TTE c/w RVS. AKI - no CTA. .. etc”, it is difficult for non-experts to understand all abbreviations without specific knowledge. But when a doctor reads this, he/she would know that although “STAT” is widely used as the abbreviation of “statistic”, “statistics” and “statistical” in most domains, in hospital emergency rooms, it is often used to represent “Immediately”. In the medical research, abbreviation expansion allows a natural language processing system automatically analyze the texts of clinical notes, which may enable knowledge discovery (e.g., relations between diseases) in clinical texts.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig1.png}
\caption{Sample Input and Output of the task and intuition of distributional similarity}
\end{figure}

In this paper, we study the task of abbreviation expansion in clinical texts. As shown in Fig-
ure 1, our goal is to normalize all the abbreviations in the intensive care medicine texts to make the texts accessible to general readers. For accurately capturing the semantics of an abbreviation in its context, we adopt word embedding, which can be seen as a distributional semantic representation and is proved to be effective to compute the semantic similarity between words based on the context without any labeled data. The intuition of distributional semantics (Harris, 1954) is that if two words share similar contexts, they should have highly similar semantics. For example, in figure 1, “RF” and “respiratory failure” have very similar contexts so that their semantics should be similar. If we know “respiratory failure” is a possible candidate expansion of “RF” and its semantics is similar to the “RF” in the intensive care medicine texts, we can determine that it should be the correct expansion of “RF”. Due to the limited resource of intensive care medicine texts where full expansions rarely appear, we exploit abundant and easily-accessible task-oriented resources to enrich our dataset for the embedding training. To the best of our knowledge, we are the first to apply word embeddings to this task. Experimental results show that the embeddings trained on the task-oriented corpus are much more useful than those trained on other corpora. By combining the embeddings with domain-specific knowledge, we achieve 82.27% accuracy, which outperforms baselines and is close to human’s performance.

2 Related Work

The task of abbreviation disambiguation in biomedical documents has been studied by various researchers using supervised machine learning algorithms (Liu et al., 2004; Gaudan et al., 2005; Yu et al., 2006; Ucgun et al., 2006; Stevenson et al., 2009). However, the performance of these supervised methods mainly depends on a large amount of labeled data which is extremely difficult to obtain for our task since intensive care medicine texts are very rare resources in clinical domain due to the high cost of de-identification and annotation. Tengstrand et al. (2014) proposed a distributional semantic based approach for abbreviation expansion in Swedish but they focused on only single-word expansion and cannot handle multi-word phrases. In contrast, we use word embeddings combined with task-oriented resources and knowledge for expanding abbreviation, which can handle multiword expressions.

3 Approach

3.1 Overview

The overview of our approach is shown in Figure 2.

Given the intensive care medicine texts, which for example are texts in the top-left triangle in figure 2, we first identify all abbreviations using regular expressions and then try to find all possible expansions of these abbreviations from domain-specific knowledge base3 as candidates. We train word embeddings using the clinical notes data with task-oriented resources such as Wikipedia article of candidates and medical scientific papers and compute the semantic similarity between an abbreviation and its candidate expansions based on their embeddings(vector representations of words).

3.2 Embeddings training with task oriented resources

Given an abbreviation as input, we expect the correct expansion to be most semantically similar to the abbreviation, which requires the abbreviation

3http://www.allacronyms.com

Figure 2: Approach overview.
and the expansion share similar contexts. For this reason, we exploit rich task-oriented resources – Wikipedia articles of all the possible candidates, research papers and books written by the intensive care medicine fellows – with our clinical notes data as a corpus to train word embeddings since the expansions of abbreviations in the clinical notes are likely to appear in these resources and share the similar contexts to the abbreviation’s contexts.

3.3 Handling MultiWord Phrases

In majority cases, an abbreviation’s expansion is a multi-word phrase. In such case, we need to obtain the phrase’s embedding so that we can compute its semantic similarity to the abbreviation. It is proven that a phrase’s embedding can be effectively obtained by summing the embeddings of words contained in the phrase (Mikolov et al., 2013; Socher et al., 2013). For computing a phrase’s embedding, we formally define a candidate \( c_i \) as a list of the words contained in the candidate, for example: one of MICU’s candidate expansions medical intensive care unit=[medical,intensive,care,unit]. Then, \( c_i \)’s embedding can be computed as follows:

\[
x(c_i) = \sum_{t \in c_i} x(t)
\]

where \( t \) is a token in the candidate \( c_i \) and \( x(\cdot) \) denotes the embedding of a word/phrase, which is a vector of real-value entries.

3.4 Expansion Candidate Ranking

Even though embeddings are very helpful to compute the semantic similarity between an abbreviation and a candidate expansion, in some cases, context-independent information is also useful to identify the correct expansion. For example, CHF in the clinical notes usually refers to “congestive heart failure”. By using embedding-based semantic similarity, we can find two possible candidates – “congestive heart failure” (similarity=0.595) and “chronic heart failure”(similarity=0.621). These two candidates have close semantic similarity score but their popularity scores in the medical domain are quite different – the former has a rating score\(^4\) of 50 while the latter only has a rating score of 7. Therefore, we can see that the rating score, which can be seen as a kind of domain-specific knowledge, can also contribute to the candidate.

We combine semantic similarity with rating information. Formally, given an abbreviation \( b \)’s candidate list \( l(b) = \{c_1, c_2, ..., c_n\} \), we rank \( l(b) \) based on the following formula:

\[
score(c) = \lambda \sum_{c_i \in l(b)} \frac{\text{rating}(c_i)}{\sum_{c_j \in l(b)} \text{rating}(c_j)} + (1 - \lambda) \frac{x(b) \cdot x(c)}{|x(b)||x(c)|}
\]

where \( \text{rating}(c) \) denotes the rating of this candidate as an expansion of the abbreviation \( b \), which reflects this candidate’s popularity, \( x(\cdot) \) denotes the embedding of a word. The parameter \( \lambda \) serves to adjust the weights of similarity and popularity.\(^5\)

4 Experiment Results

4.1 Data and Evaluation Metrics

The clinical notes we used for the experiment are provided by domain experts, consisting of 1,160 Physician logs of Medical ICU admission requests at a single tertiary care center. The logs were prospectively collected over one year with free-text descriptions of clinical presentation and triage decisions. We identify 818 abbreviations and find 42,506 candidates using domain-specific knowledge (i.e., www.allacronym.com/_medical). The enriched corpus contains 42,506 Wikipedia articles, each of which corresponds to one candidate, 6 research papers and 2 textbooks about intensive care unit, besides our raw intensive care medicine data.

We use word2vec (Mikolov et al., 2013) to train the word embeddings. The dimension of embeddings is empirically set to 100.

Since the goal of our task is to find the correct expansion for an abbreviation, we use accuracy as a metric to evaluate the performance of our approach. For ground-truth, we have 100 physician logs which are manually expanded and normalized by our collaborator, a well-trained domain expert, and thus we use these 100 physician logs as the test set to evaluate our approach’s performance.

4.2 Baseline Models

For our task, it’s difficult to re-implement the supervised methods since we do not have sufficient

\(^4\)All the rating information in this paper is from http://www.allacronyms.com. On this website, users are free to rate expansions of an abbreviation if they like the expansions. In general, a popular expansion has a high rating score.

\(^5\)In the experiments, \( \lambda \) is empirically tuned to 0.2 on a separated development set.
training data. And a direct comparison is also impossible because previous work all used different data sets which are not publicly available. Alternatively, we use the following baselines to compare with our approach.

- Rating: This baseline model chooses the highest rating candidate expansion in the domain specific knowledge base as the expansion of an abbreviation.
- Raw Input embeddings: We trained word embeddings only from the 1,160 raw intensive care medicine texts and choose the most semantically related candidate as the answer.
- General embeddings: Different from the Raw Input embeddings baseline, we use the embedding trained from huge biomedical data which includes knowledge bases like PubMed and PMC and wikipedia dump of biomedical related articles (Pyysalo et al., 2013) for semantic similarity computation.

4.3 Results
Table 1 shows the performance of abbreviation expansion. Our approach significantly outperforms the baseline methods – achieves 82.27% accuracy.

<table>
<thead>
<tr>
<th>Approaches</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rating</td>
<td>21.32%</td>
</tr>
<tr>
<td>Raw Input embeddings</td>
<td>26.45%</td>
</tr>
<tr>
<td>General Embeddings</td>
<td>28.06%</td>
</tr>
<tr>
<td>Our Approach</td>
<td>82.27%</td>
</tr>
</tbody>
</table>

Table 1: Overall performance

Figure 3 shows how our approach improves the performance of rating-based approach. By using embeddings, we can learn the meaning of “OD” used in the clinical notes should be “overdose” rather than “out-of-date” and this semantic information largely benefit the abbreviation expansion model.

Compared with our approach, embeddings trained only from the intensive care medicine texts does not significantly contribute to the performance over the rating baseline. The reason is that the size of data for training the embeddings is so small that many candidate expansions of abbreviations do not appear in the corpus, which results in the poor performance. It is notable that General embeddings trained from large biomedical data are


- ‘OD’- our approach: [‘overdose’, ‘osteo-chondritis dissecans’, ‘optic disc’ ... ... etc.]

Figure 3: Ranking lists of expansions of “OD” by the rating-based method, our approach not effective for this task because many abbreviations in the intensive care medicine texts appear in the biomedical corpus with a different sense.


Figure 4: The output of general embeddings trained on large biomedical texts

For example, “OD” in intensive care medicine texts refers to “overdose” while in PubMed corpus it usually refers to “optical density”, as shown in figure 4. Therefore, the embeddings trained from the PubMed corpus do not benefit the expansion of abbreviations in the intensive care medicine texts.

Moreover, we estimated human performance for this task, shown in table 2. Note that the performance is estimated by our collaborator who is a well-trained fellow in intensive care medicine based on her experience and the human’s performance estimated in table 2 is under the condition that the participants can not use any other external resources. We can see that our approach can achieve a performance close to domain experts and thus it is promising to tackle this challenge.

4.4 Analysis of Errors
The distribution of errors is shown in table 3. There are mainly 3 reasons that cause the incorrect expansion. In some cases, the certain abbreviation is hardly used thus does not exist in the knowledge base. In this case we would not be able to populate the corresponding candidate list, resulting in “Out of vocabulary”. Secondly, in many cases although we have the correct expansion in the candidate list, it’s not ranked as the top one due to the lower semantic similarity because there are not enough samples in the
### Groups | Accuracy
---|---
General readers | <40%  
Nurses | 40%  
Mid-level provider (nurse practitioner or physician associate) | 70%  
Trained physician in Medicine | 80%  
Domain experts strictly trained in Emergency Department or Intensive Care Unit | >90%

Table 2: Estimated human performance for abbreviation expansion

Among all the incorrect expansions in our test set, such kind of error accounted for about 54%. The possible solution may be adding more helpful data to the embedding training, which means discovering more task-oriented resources. In a few cases, we failed to identify some abbreviations because of their complicated representations. For example, we have the following sentence in the patient’s texts: “No n/v/f/c.” and the correct expansion should be “No nausea/vomiting/fever/chills/diarrhea/dysuria.” Such abbreviation is by far the most difficult one to expand in our task because it does not exist in any knowledge base and usually only occurs once in the training data.

### Type of error | Percentage
---|---
Out of Vocabulary | 27%  
Lack of training samples | 54%  
Unidentified representation | 19%

Table 3: Error distribution

## 5 Conclusion and Future Work

This paper proposes a simple but novel approach for automatic expansion of abbreviations. It achieves very good performance without any manually labeled data. Experiments demonstrate that using task-oriented resources to train word embeddings is much more effective than using general or arbitrary corpus.

In future work, we plan to collectively expand semantically related abbreviations co-occurring in a sentence. In addition, we expect to integrate our work into a natural language processing system for processing the clinical notes for discovering knowledge, which will largely benefit the medical research.

## Acknowledgements

We thank the anonymous reviewers for their thought-provoking comments. We also thank the Icahn School of Medicine at Mount Sinai Hospital for providing the experiment data and evaluation.

## References


Hong Yu, Won Kim, Vasilios Hatzivassiloglou, and John Wilbur. 2006. A large scale, corpus-based approach for automatically disambiguating biomedical...