A Hybrid Normalization Method for Medical Concepts in Clinical Narrative using Semantic Matching

Yen-Fu Luo, MS¹, Weiyi Sun, PhD², Anna Rumshisky, PhD¹
¹University of Massachusetts Lowell, Lowell, MA, USA ; ²Nuance Communication Inc., Burlington, MA, USA

Abstract Normalization maps clinical terms in medical notes to standardized medical vocabularies. In order to capture semantic similarity between different surface expressions of the same clinical concept, we develop a hybrid normalization system that incorporates a deep learning model to complement the traditional dictionary lookup approach. We evaluate our system against the ShARe/CLEF 2013 challenge data in which 30% of the mentions have no concept mapping. When evaluating against the mentions which may be normalized to existing concepts, our hybrid system achieves 90.6% accuracy, obtaining a statistically significant improvement of 2.6% over a strong edit-distance and dictionary lookup combined baseline. Our analysis of semantic similarity between concepts and mentions reveals existing inconsistencies in ShARe/CLEF data, as well as problematic ambiguities in the UMLS. Our results suggest the potential of the proposed deep learning approach to further improve the performance of normalization by utilizing semantic similarity.

Introduction

The information about findings, symptoms, diseases, diagnoses, and medications recorded in the medical notes is invaluable for clinical applications. The information extracted from medical notes has been used in clinical decision-making¹⁻³, mortality prediction⁴,⁵, adverse drug effect analysis⁶,⁷ among others. A lot of work has focused on named entity recognition (NER) for clinical notes⁸⁻¹² which identifies text spans (“mentions”) of the clinically-relevant concepts. However, without linking such information to standardized medical vocabularies, the proliferation of similar-sounding acronyms, and conversely, the presence of synonyms that can be used interchangeably to refer to the same clinical concept, coupled with misspellings, reduce the usefulness of this information for generalization. For example, one may use “heart attack”, “MI”, and “myocardial infarction” to refer to the same general concept, and if different words are used to express this concept in patient records, we would fail to identify similarity between such patients, unless we can correctly identify the underlying concept in each case. In this paper, we focus specifically on the entity normalization task. As distinct from NER, the normalization task involves mapping individual mentions of clinically relevant entities to the entries in the standardized medical vocabulary.

There are very few annotated datasets available for the task of clinical entity normalization. One such dataset was provided by task 1 in ShARe/CLEF eHealth 2013 challenge* which defined two subtasks: (a) named entity recognition, and (b) normalization of disorders. In this research, we focus on the normalization subtask, defined as mapping a given mention to a concept unique identifier (CUI) from the Unified Medical Language System (UMLS)¹³. Following ShARe/CLEF 2013, SemEval-2014 Task 7¹⁴ and SemEval-2015 Task 14¹⁵ included additional annotated medical notes for the same task. Subtask (a), the named entity recognition, has been well explored in the Clinical NLP research community, with a number of NER systems, including MetaMap⁸, cTAKES⁹, CLAMP¹⁰, CliNER¹¹, and RapTAT¹², available. Although such well-known tools as MetaMap, cTAKES and CLAMP contain normalization modules, they are based on morphological variation. Such approaches have obvious drawbacks, as previously pointed out by Kate¹⁴ and elsewhere, specifically, (1) the systems based on morphological variation may not capture the semantic similarity between medical mentions and concepts, and (2) TF-IDF¹⁵ representation, commonly used in the normalization modules, may be inappropriate for short medical terms.

Amongst the top-ranked teams in subtask (b) of ShARe/CLEF 2013, DNorm¹⁶ learned a normalization weight matrix using pairwise ranking¹⁷ over TF-IDF representation and UTHeach¹⁸ normalized clinical terms based on cosine similarity over TF-IDF representation. In SemEval 2014 and 2015, UTH-CCB¹⁹,²⁰ used a similar TF-IDF representation.

---

*ShARe/CLEF eHealth 2013. https://sites.google.com/site/shareclefehealth/
and cosine similarity method for normalization. ezDI\textsuperscript{21,22} adopted dictionary lookup over lexical variants generated by Lexical Variants Generation\textsuperscript{3}, with Levenshtein edit distance\textsuperscript{23} used to normalize the mentions that fail to be normalized by exact match. UWM\textsuperscript{24,25} learned normalization rules automatically based on edit distance at character level. Following UWM, D’Souza and Ng\textsuperscript{26} proposed a sieve-based approach which includes a series of lexical transformations combined with exact match and rule-based partial match. In summary, most of the existing systems use either a vector representation, typically TF-IDF or lexical transformations, in combination with similarity ranking for the normalization task. Because of the characteristics of the dataset, dictionary lookup and its variants may produce high performance. However, such normalization methods are challenged by the remaining mentions which can not be resolved in this way and require semantic analysis\textsuperscript{14}.

Li et al.\textsuperscript{27} recently applied convolutional neural network (CNN) to rank the candidate concepts in the normalization task. Their rule-based system first generates candidate concepts which are morphologically similar to a given mention. Then, CNN-based ranking is used to find the optimal normalization from the candidate concepts. However, this approach does not generalize to the cases when a concept is semantically similar but morphologically different. Inspired by the recent success of Bidirectional Long Short-Term Memory (Bi-LSTM) models\textsuperscript{28} in sequence modeling tasks, and specifically, in NER\textsuperscript{29,30}, we apply a LSTM-based model to the normalization task.

In this work, we propose a hybrid method which combines the traditional dictionary lookup with a deep learning approach to the normalization task and investigate its capability for matching concept mentions to the standardized clinical concepts at semantic level. We represent concept mentions and their context using dense lexical embeddings that capture semantic similarity between words as input to a deep learning model. Concept mentions and their context are processed by three parallel Bi-LSTM layers in a recurrent neural network model, with threshold-based concept assignment. The model is trained on a combination of labeled data and concept synonyms provided by the UMLS, ensuring wider coverage and better generalization. Evaluated on the normalized mentions from the ShARe/CLEF dataset, our deep learning hybrid system achieves 90.6% accuracy, obtaining a statistically significant (0.05 alpha) improvement of 2.6% over the accuracy of 88.0% for the edit distance and dictionary lookup combined.

Dataset

**ShARe/CLEF eHealth 2013**

The publicly available ShARe/CLEF 2013 dataset contains de-identified clinical free-text notes from the MIMIC II\textsuperscript{31} database. The dataset contains 199 notes in the training set and 99 notes in the test set. A disorder mention is defined as any span of text which can be mapped to a CUI in the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT)\textsuperscript{¶} terminology and which belongs to one of the 11 UMLS semantic types including “Congenital Abnormality”, “Acquired Abnormality”, “Injury or Poisoning”, “Pathologic Function”, “Disease or Syndrome”, “Mental or Behavioral Dysfunction”, “Cell or Molecular Dysfunction”, “Experimental Model of Disease”, “Anatomical Abnormality”, “Neoplastic Process”, and “Signs and Symptoms”. If no appropriate CUI can be found for a disorder mention, it is assigned to the “CUI-less” category. Due to limited access to the expanded dataset from SemEval 2014 Task 7 and SemEval 2015 Task 14, the development and evaluation of our work is based on ShARe/CLEF dataset.

**UMLS and SNOMED CT**

In order to minimize the variance caused by different UMLS releases, we used sources in English included in the active UMLS 2013 AA release which was when the ShARe/CLEF data was released. Following the specification of the task, we retrieved CUIs belonging to one of the 11 semantic types that occur in SNOMED CT. We also collected the concept synonyms for each retrieved CUI from all UMLS sources, and normalized them to lowercase. UMLS provides a list of concept synonyms for each CUI. We describe how these are used in our normalization approach in the methods section.

\textsuperscript{3}Lexical Variants Generation. \url{https://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/lvg/2014/docs/userDoc/tools/lvg.html}

\textsuperscript{¶}SNOMED CT. \url{https://www.snomed.org/snomed-ct}
Medical Information Mart for Intensive Care III (MIMIC III)

MIMIC III database was used to train word2vec word embeddings. This data were collected between 2001 and 2012 at Boston’s Beth Israel Deaconess Medical Center. The database contains over 2 million notes from 40K patients. The following pre-processing steps were applied to the notes: (1) using regular expressions to replace 64 Protected Health Information (PHI) patterns with 56 unique tokens; (2) removing punctuation; (3) replacing floating numbers with a unique placeholder token; (4) lowercasing.

Methods

We created two baseline systems as comparison to our hybrid deep learning approach. The first baseline is an exact-match system. The second baseline is the edit-distance system which is the top-ranked team in the SemEval 2014 and 2015 clinical tasks. Our hybrid normalization system utilizing deep learning approach is built on the state-of-the-art edit-distance system. We describe the details of two baseline systems and the architecture of our hybrid normalization method which combines the exact match and edit distance in the following sections.

Exact Match Baseline

The exact-match method contains two components: (1) exact match against the mentions in the training set, and (2) exact match against the UMLS concept synonyms mentioned in the dataset section above. We consider a mention to be an exact match if it occurs in the same typographical form in one of the above sources.

If a mention can be matched exactly to a unique CUI, it is normalized to that CUI. Otherwise, the mention is considered ambiguous and is assigned to the CUI-less category. For example, the mention “lung cancer” is treated as CUI-less because there are two CUIs (C0684249 and C0242379) bound to the concept synonym “lung cancer”. One common approach to resolving ambiguity in such case is to randomly assign the first CUI to the mention. However, this approach produced worse accuracy on the development dataset, therefore we do not use it. If no exact match is found, we also treat the mention as CUI-less.

Edit Distance Baseline

Ghiasvand and Kate and Kate implemented automatically learned edit-distance method to the normalization task and were top-ranked teams in SemEval 2014 and 2015 clinical task. We followed their approach to build a similar exact-match baseline and used their open-source edit-distance system for comparison as well as a component in our hybrid method. We use the open-source implementation of the edit-distance baseline by Kate with the default settings, but we omit the abbreviation disambiguation step. If multiple full forms exist for a given abbreviation, their strategy is to arbitrarily pick the first full form in the dictionary. We found that this arbitrary disambiguation strategy did not have a substantial impact on the relative performance improvement obtained by our hybrid system.

Deep Learning Model

Our deep learning model contains four layers: (1) an embedding layer that converts input tokens into word embeddings, (2) three Bi-LSTMs which are used to process the concept mention, the left context, and the right context, (3) a dense layer of size 2048 neurons, and (4) a softmax output layer. The model architecture is illustrated in Figure 1. Occurrence contexts are extracted with a window size +/- 10 tokens. The outputs from the Bi-LSTMs are concatenated and fed into the dense layer. Because the number of possible classes (CUI assignments) is large, we use a sampled softmax loss during training and a regular softmax in prediction. Dropout with 20% probability is applied to avoid model overfitting.

The text spans for the concept mention and for the left and right contexts are represented as sequences of word2vec word vectors. We used Gensim to train 200-dimensional word embeddings over all MIMIC III notes, using skip-gram in order to consider word order. We used hierarchical softmax and the frequency cutoff 3. The embeddings were sampled using the TensorFlow API. The sampled softmax loss is defined as follows:

\[ \text{Sampled Softmax Loss. https://www.tensorflow.org/api_docs/python/tf/nn/sampled_softmax_loss} \]

**Gensim. https://radimrehurek.com/gensim/
tuned continuously during model training.

Importantly, we expand our training data by creating additional training instances using UMLS concept synonyms. Specifically, for the all valid disease CUIs, the retrieved concept synonyms are used at the training stage as mentions with zero left and right context, with the corresponding CUI as a label. We exclude from training the instances which contain out of vocabulary words for our word2vec model, i.e. words that occurred less than three times in MIMIC. We also exclude concept synonyms that map to multiple CUIs.

According to the released data statistics, 30% of mentions in the ShARe/CLEF data are not mapped to any UMLS CUI, because no appropriate concept could be found. These mentions are assigned to the “CUI-less” category, which serves as a catch-all label for a variety of different concepts. Although we treat the normalization task as a classification problem, we do not assign a mention to a CUI-less category when it has the highest probability, but instead use a threshold determined on the validation data. That is, the CUI-less label is assigned whenever the probability of the most likely CUI is below that threshold. The output of the model is a probability distribution over all 64798 CUIs, including the CUI-less category. We divide the training set into 72% training (dev-train), 8% validation (dev-val), and 20% development test (dev-test), and use dev-val for threshold selection.

**Hybrid Method**

Since the ShARe/CLEF data contains solely the mentions of disorders, the relatively unique and precise surface lexical expressions make the dictionary lookup a very strong baseline. We therefore propose a hybrid approach which combines exact-match module, edit-distance module, and a deep learning model as illustrated in Figure 2. Our hybrid method normalizes concept mentions using a sequential pipeline. If a given concept mention is not associated with a CUI at a given stage in the pipeline, the mention is passed on to the next stage. We compared the performance of two hybrid configurations: (1) exact match + edit distance (EM + ED), and (2) exact match + edit distance + deep learning (EM + ED + DL).

**Results and Discussion**

**Normalization Task Evaluation**

In the ShARe/CLEF 2013 challenge, the tasks of NER and normalization were evaluated together, and only the combined score for these tasks was reported for participating systems. Therefore, we cannot directly compare our normalization system to the participating systems scores reported in the ShARe/CLEF 2013 challenge paper. Instead, we use the open-source implementation from Kate\(^{14}\), which reports high scores on SemEval 2014 and 2015 challenge data,
Figure 2: Architecture of the hybrid deep learning model.

as one of our baselines.

Following prior work, we evaluate our system under two settings: on the full dataset, including the CUI-less mentions, and only on the mentions which could be normalized to existing UMLS concepts. Table 1 shows system performance for the first setting, when the CUI-less category is included in evaluation. Table 2 shows the results of evaluation with the CUI-less category excluded, i.e. only on the mentions mapped to existing CUIs. We also report F-measure, precision, and recall following prior work. The reported weighted metrics consider label imbalance by weighting label metrics based on the number of true instances for each label.

Under the setting in which the CUI-less mentions are included, the EM + ED and EM + ED + DL configurations show comparable performance. However, when they are excluded from evaluation, the hybrid method with the deep learning component performs better on most of the metrics.

Table 1: Evaluation of Exact Match (EM), Exact Match + Edit Distance (EM + ED), and Exact Match + Edit Distance + Deep Learning (EM + ED + DL) models against all mentions including the CUI-less category.

<table>
<thead>
<tr>
<th></th>
<th>EM</th>
<th>EM+ED</th>
<th>EM+ED+DL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accuracy</strong></td>
<td>87.4</td>
<td>89.8</td>
<td>89.6</td>
</tr>
<tr>
<td><strong>Micro</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F-measure</td>
<td>87.9</td>
<td>90.7</td>
<td>91.2</td>
</tr>
<tr>
<td>Precision</td>
<td>88.5</td>
<td>91.6</td>
<td>92.9</td>
</tr>
<tr>
<td>Recall</td>
<td>87.4</td>
<td>89.8</td>
<td>89.6</td>
</tr>
<tr>
<td><strong>Macro</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F-measure</td>
<td>71.5</td>
<td>77.8</td>
<td>80.4</td>
</tr>
<tr>
<td>Precision</td>
<td>74.5</td>
<td>79.9</td>
<td>81.5</td>
</tr>
<tr>
<td>Recall</td>
<td>70.7</td>
<td>77.7</td>
<td>81.1</td>
</tr>
<tr>
<td><strong>Weighted</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F-measure</td>
<td>85.4</td>
<td>88.9</td>
<td>89.8</td>
</tr>
<tr>
<td>Precision</td>
<td>85.6</td>
<td>89.4</td>
<td>91.0</td>
</tr>
<tr>
<td>Recall</td>
<td>87.4</td>
<td>89.8</td>
<td>89.6</td>
</tr>
</tbody>
</table>

Semantic Similarity between Concepts and Concept Mentions

In order to understand the information captured by the deep learning model, we further examined the CUIs selected by the model on the dev-test data. Some examples of the concepts identified by the model for the mentions labeled as CUI-less in the annotated data are shown in Table 3. While not all examples selected by the model are equally

Table 2: Evaluation against mentions normalized to existing UMLS concepts.

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>EM</th>
<th>EM+ED</th>
<th>EM+ED+DL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>F-measure</strong></td>
<td>90.5</td>
<td>92.9</td>
<td>94.1</td>
</tr>
<tr>
<td><strong>Precision</strong></td>
<td>98.9</td>
<td>98.4</td>
<td>97.9</td>
</tr>
<tr>
<td><strong>Recall</strong></td>
<td>83.4</td>
<td>88.0</td>
<td>90.6</td>
</tr>
</tbody>
</table>

| Micro-F-measure | 72.1    | 78.5    | 81.4     |
| Precision       | 75.5    | 81.0    | 83.3     |
| Recall          | 70.7    | 77.7    | 81.1     |

| Weighted-F-measure | 86.0 | 89.7 | 91.5 |
| Precision         | 90.9 | 93.2 | 93.8 |
| Recall            | 83.4 | 88.0 | 90.6 |

reasonable, it is nonetheless clear that in many cases the model is able to capture semantic similarity between concepts and concept mentions, and ultimately provide a performance boost over the methods relying on dictionary lookup using lexical transformations. However, this also suggests that at least some of the potential improvement that could be provided by the deep learning model is lost due to underspecified labelling choices made during annotation. The latter may also be an artifact of the annotation methodology. For example, if the UMLS search interface is used to identify possible concept mappings for “suicidal ideation”, a human annotator would not be able to associate it with C0522178 “Thoughts of self harm”, since it will not be listed.

Table 3: UMLS concepts predicted by the deep learning model for selected CUI-less mentions.

<table>
<thead>
<tr>
<th>CUI Annotation</th>
<th>CUI Prediction</th>
<th>Mention</th>
<th>Concept Synonym</th>
</tr>
</thead>
<tbody>
<tr>
<td>CUI-less</td>
<td>C0522178</td>
<td>suicidal ideation</td>
<td>Thoughts of self harm</td>
</tr>
<tr>
<td>CUI-less</td>
<td>C0023380</td>
<td>unarousable</td>
<td>Lethargy</td>
</tr>
<tr>
<td>CUI-less</td>
<td>C0575090</td>
<td>poor balance</td>
<td>Impairment of balance / Problem with balance</td>
</tr>
<tr>
<td>CUI-less</td>
<td>C0013404</td>
<td>trouble breathing</td>
<td>Difficulty breathing</td>
</tr>
<tr>
<td>CUI-less</td>
<td>C0234246</td>
<td>Abdomen rebound</td>
<td>Rebound tenderness</td>
</tr>
<tr>
<td>CUI-less</td>
<td>C0576690</td>
<td>Position sense impaired</td>
<td>Impaired body position sense</td>
</tr>
<tr>
<td>CUI-less</td>
<td>C0333133</td>
<td>mucus plugging event</td>
<td>Mucus cast / Mucus plug</td>
</tr>
<tr>
<td>CUI-less</td>
<td>C0238997</td>
<td>dullness to percussion</td>
<td>Chest dull to percussion</td>
</tr>
</tbody>
</table>

Another source of discrepancies between the gold standard annotation and possible semantically matching concepts is the presence of multiple similar concepts in the UMLS Metathesaurus itself. For example, the mention “depression” may be appropriately normalized to either C0011581 “Depressive disorder” or C0011570 “Mental Depression”. The same synonym can often be given for multiple CUIs, making them essentially equivalent. Thus, “chronic renal insufficiency” may be associated with either C0022661 “Kidney Failure, Chronic” or C0403447 “Chronic Kidney Insufficiency” both of which are listed with the synonym, “chronic renal insufficiency”.

Some of such inconsistencies have been or are being handled in the later UMLS releases, so in the future, one can expect improved performance of the models that rely on this information. For example, in the 2011 UMLS release used for the ShARe/CLEF data annotation, “infectious source” could be associated with either C0021311 “Infection”
or with C0009450 “Communicable disease” and these two concepts were merged in later UMLS releases.

Note that in this work, we have not attempted to address the disambiguation of abbreviations and acronyms, which we believe should be treated as a separate task. Effectively, this means that our system represents them in the same way as regular tokens, that is, using dense word2vec embeddings. However, the context representation used by the deep learning model is nonetheless able to identify the appropriate concepts for some of the acronyms and abbreviations that have multiple disambiguations. For example, our deep learning model was able to correctly normalize “Breast CA” to C0006142 “Malignant neoplasm of breast”, “pelvic fx” to C0149531 “Fracture of pelvis”, and “Schizoaffective D/O” to C0036337 “Schizoaffective Disorder”, and “C. difficile” to C0343386 “Clostridium difficile infection”. This suggests that it may be possible to include the disambiguation of acronyms and abbreviations in our hybrid deep learning system as a separate component in the future.

Conclusion

In this work, we have proposed a novel hybrid method for linking clinically relevant entity mentions in free-text medical notes to standardized medical vocabularies. This method combines traditional dictionary lookup with a deep learning model, showing an improvement over existing state-of-the-art and the potential of the deep learning model to better capture semantic similarity between concepts and concept mentions, compared to a strong edit distance and dictionary lookup-based baseline. We believe that the two main factors contributed to the relatively modest improvement the deep learning model was able to achieve: (1) the presence of the semantically diverse catch-all CUI-less category in the data, for which no consistent representation could be learned, and (2) the relatively small training dataset size, which is a known impediment to improved performance by deep learning techniques. The qualitative examination of the CUIs predicted by the deep learning model also highlighted existing inconsistencies and ambiguities in the ShARe/CLEF data, as well as in the UMLS, which will need to be addressed for normalization tasks in the future.

Acknowledgment

This work was supported in part by a research grant from Philips HealthCare.

References


