# Category Multi-Representation: A Unified Solution for Named Entity Recognition in Clinical Texts

Jiangtao Zhang[†], Juanzi Li[†], Shuai Wang[†], Yan Zhang[†], Yixin Cao[†], Lei Hou[†], and Xiao-Li Li[‡]

[†]Department of Computer Science and Technology,
Tsinghua University, Beijing 100084, China
[‡]Institute for Infocomm Research, A*STAR, Singapore 138632
{zhang-jt13@mails.,lijuanzi@}tsinghua.edu.cn, 18813129752@163.com,
zhangyan9988@qq.com, {caoyixin2011,greener2009}@gmail.com,
xlli@i2r.a-star.edu.sg

**Abstract.** Clinical Named Entity Recognition (CNER), the task of identifying the entity boundaries in clinical texts, is essential for many applications. Previous methods usually follow the traditional NER methods that heavily rely on language specific features (i.e. linguistics and lexicons) and high quality annotated data. However, due to the problem of *Limited Availability of Annotated Data* and *Informal Clinical Texts*, CNER becomes more challenging. In this paper, we propose a novel method that learn multiple representations for each category, namely *category-multi-representation* (CMR) that captures the semantic relatedness between words and clinical categories from different perspectives. CMR is learned based on a large scale unannotated corpus and a small set of annotated data, which greatly alleviates the burden of human effort. Instead of the language specific features, our proposed method uses more evidential features without any additional NLP tools, and enjoys a lightweight adaption among languages. We conduct a series of experiments to verify our new CMR features can further improve the performance of NER significantly without leveraging any external lexicons.

## 1 Introduction

Electronic Medical Records (EMR) contains valuable and detailed medical information of patients accessed and modified in a digital format [15]. Identifying the boundaries of clinically relevant entities in *clinical texts* from EMR and classifying them into predefined categories such as *disease*, *treatment* and *symptom*, namely *Clinical Named Entity Recognition* (CNER) is a fundamental task both in medical data mining and information extraction. CNER could benefit many applications in medical domain such as comorbidity analyses, syndromic surveillance, adverse drug event detection and the analysis of drug-drug interaction [12], as well as the NLP related tasks like information retrieval, relation extraction and question answering [18].

Most existing work of NER in medical domain [6, 10, 21, 4, 1] simply follows the conventional NER methods in general domain which focus on identifying

general named entities such as *person*, *location* and *organization*. They usually utilize linguistic features based on syntactics and lexicons[1] to feed a supervised model, e.g. SVM [24], CRF [21] or a hybrid of several classification models [6, 10]. However, these methods may achieve poor performance in realistic applications because (i) they heavily depend on linguistic features and lexicons, which varies greatly among different datasets or across various languages, and (ii) the annotated data for the supervised model is not always available.

Despite the success of traditional NER, CNER receives relatively few studies which has the following challenges:

**Limited Availability of Annotated Data** As mentioned above, previous works following traditional NER rely on a supervised model over a high quality training data. However, in the clinical domain, annotated data are not only expensive (usually requires domain expertise) but also often unavailable due to patient privacy and confidentiality requirements. Even though there are a few public available annotated datasets for CNER task, such as i2b2 2010 [25] and ShARe/CLEF eHealth 2013 [23], they are usually insufficient for training an applicable system. For example, ShARe dataset contains only 300 documents including 9,768 entity mentions annotated. On the other hand, the gap among different languages always requires new language-specific annotated data. Therefore, we need to use the unlabeled data, usually available in clinical domain, such as MIMC III [11], to alleviate the burden of human effort involved in creating annotated resources and improve the performance of CNER.

**Informal Clinical Texts** A clinical text is dictated by a doctor (and transcribed later by a third-party) to capture the proceedings of a doctor-patient interaction, or to document the results of a medical procedure or test [12]. It is usually far different from general texts and even scholarly medical literatures. Clinical texts have the following unique characteristics: 1) incomplete sentences, 2) informal grammar, and 3) littered with misspellings and non-standard shorthand, abbreviations and acronyms. All these characteristics result in the unreliability of the linguistic based features used in the traditional NER and the effectiveness of NLP tools (e.g. POS tagging). Therefore, we need to explore more evidential features with good generalization and independent of language to cope with characteristics of clinical texts.

To address these challenges, our solution is to learn semantic features by taking advantage of large scale unannotated corpora, instead of the language specific features, such as syntactic and lexicon. The semantic features will be trained in an unsupervised way, and measure the similarity between the words in clinical texts and CNER categories. Our solution doesn't rely on any additional NLP tools which can avoid the unreliable linguistic features, and alleviate the burden of language specific annotated data.

In this paper, we propose *a unified solution* for CNER without leveraging any language specific features. It induces multiple representations for each category, namely *category-multi-representation* (CMR) that is used to measure the

[1] These methods extract lexicons from UMLS [3] or MeSH: https://www.nlm.nih.gov/mesh/meshhome.html.

semantic similarity between words and categories. Specifically, we first construct a semantic space of clinical texts by employing a model of distributed representation (word embedding) over a large unannotated clinical corpus (e.g. MIMC III). As each entity mention has been classified into a certain predefined category in the annotated dataset, each category could be regarded as a *vector cluster* in the semantic space. Then we learn multiple representations for each category from 4 different aspects by leveraging the statistics and context information derived from the large unlabeled data to holistically capture the meaning of each category. That is, CMR shares a common semantic space with words in clinical texts which could easily be used to measure the semantic similarity between words and categories. Based on these representations, our proposed model only requires a small annotated dataset for training a sequence labeling model due to the good generalization ability of CMR. For inference, we adopt a heuristic method to assign a threshold for each CMR, which aims to filter out irrelevant noise (words) belonging to the corresponding category.

**Contributions** Our main contributions are summarized as follows.

- To the best of our knowledge, this is the first work for CNER that represents category from multiple perspectives, which is based on unlabeled clinical corpus without any additional NLP tools.
- Our CNER model is a *united* method, which is independent of language-specific features (i.e. lexicons and linguistic features), and lightweight for adaption to identify clinical entities in another language and another dataset.
- Extensive experiments are conducted on two public datasets and the results demonstrate that our new CMR features can further improve the performance of CNER by 2.05% in terms of F1 score.

## 2 Problem Definition

Given a clinical text $\mathbf{s} = \langle w_1, w_2, ...w_{|\mathbf{s}|}\rangle$ and a set of predefined categories $C = \{c_1, ...c_{|C|}\}$, the output of our task is to generate a list of tags $t_i$ for each word $w_i \in \mathbf{s}$. $t_i \in \mathcal{T} = \{cp|c \in C, p \in \mathcal{P} - \{O\}\} \cup \{O\}$ is a *category-position* combinatorial tag for $w_i$, where $\mathcal{P} = \{B, I, O\}$ is a set of *position* tags indicating the position information of a word located in an entity mention. $B$ and $I$ stand for beginning, intermediate positions of a *multi-word* entity respectively and $O$ denotes outside of any entity mention. In short, our task is to identify every entity mention $m = \langle w_i, \cdots, w_j\rangle, i \leq j$ (perhaps including multiple words) occurring in the clinical text $\mathbf{s}$ and then classify it into a predefined category $c_i \in C$. Fig.1

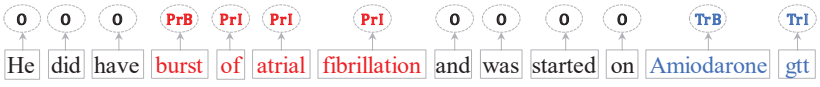


**Fig. 1.** An example of labeling process for CNER.

gives an example of sequence labeling for CNER, in which $C = \{Pr, Tr, Te\}$, represents *Problem*, *Treatment* and *Test* respectively.

## 3 Our Proposed Approach

Our proposed method presupposes the existence of two resources: (1) an annotated corpus $\mathcal{L}$ in which each word has been annotated as a predefined category $c \in C$; (2) a much larger unannotated clinical corpus $\mathcal{U}$. The main steps of our method are as follows. Firstly, we construct a semantic clinical space by training a word embedding model over $\mathcal{U}$. Each predefined category can be seen as a word cluster in this space. Secondly, we learn abstract representations for each category from many different perspectives (CMR) derived from $\mathcal{U}$. Thirdly, we generate a bundle of novel features for the target word based on its distance to each of CMR. Lastly, an appropriate learning algorithm is applied to $\mathcal{L}$ with the generated new features to evaluate our method. The focus of this paper is primarily on the first three steps.

### 3.1 Generating Semantic Space

We first construct a semantic space by learning word embeddings (e.g. GloVe [17] and Word2vec [16]) on $\mathcal{U}$ to obtain low-dimensional, real-valued vector representation for each word in clinical texts. Each word $w \in \mathcal{U} \cup \mathcal{L}$ is represented as a point (vector) $\mathbf{v}_w$ in this semantic space. If an entity mention $m = \langle w_i, \cdots, w_j \rangle$ contains more than one word ($i < j$), we simply represent it as the mean vector of its component words, i.e. $\mathbf{v}_m = (\frac{1}{j-i+1}) \sum_{k=i}^{j} \mathbf{v}_{w_k}$.

### 3.2 Category Multi-Representation

We first build a *category-words* set for each predefined category based on $\mathcal{L}$ and $\mathcal{U}$. That is, $\forall c_i \in C$, we get $\mathbf{c}_i = \{w_{i1}, \cdots, w_{ij}, \cdots, w_{i|\mathbf{c}_i|}\}$ where each $w_{ij}$ has been annotated as $c_i$ in training dataset of $\mathcal{L}$ and occurs at least 100 times in $\mathcal{U}$. Then each predefined category can be regarded as a word cluster in the semantic space. The key point is how to represent the cluster of each category in order to more holistically capture the meaning of it.

**One-center Representation** Since the distance between vectors indicates the strength of the semantic relatedness of their corresponding words in the semantic space, we regard each category as a *hypersphere* constructed by the vectors in its category-words set. Each word located in the hypersphere of a category is more likely classified into it without considering any orthographic and syntactic features. In another word, the closer a word $w$ is to the centre of the hypersphere of a category $c_i$, the more likely the word $w$ belongs to the category $c_i$. Then we represent the category $c_i$ as the *centroid vector* of the semantic vectors of its category-words set $\mathbf{c}_i$ as follows:

$$\mathbf{R_o}(c_i) = centroid(\mathbf{c}_i) = \langle median_{i1}, \cdots, median_{in} \rangle, i = 1, \cdots, |C| \qquad (1)$$

where the centroid vector is defined as the median value of each dimension of the semantic vectors of words in $\mathbf{c}_i$ and $n$ is the dimension size of the embedding vectors.

**Multi-sub-center Representation** Each predefined category usually can be subdivided into several sub-categories in clinical texts. For example, category *Disease* can be classified as *Mental or Behavioral Dysfunction* and *Neoplastic*

*Process.* Words in the same sub-category are more similar (closer in semantic space) to each other than to those in other sub-categories. In other words, the category may not be a normal hypersphere, it could be represented as several smaller sub-hyperspheres. Therefore, we use a clustering algorithm (Affinity propagation used in this paper which does not predefine the number of clusters) to group all words in each category $\mathbf{c}_i$ into $K_i$ clusters $\{\mathbf{s}_{i1}, \cdots, \mathbf{s}_{ij}, \cdots, \mathbf{s}_{iK_i}\}$ where $\mathbf{s}_{ij}$ is a subset of words in $\mathbf{c}_i$. Then, we represent each category $c_i$ as the *set of centroids* of its sub-hyperspheres. The premise of this representation is that some words which are a bit far from the centroid of the category are probably close to the centroids of some sub-hyperspheres.

$$\mathbf{R_m}(c_i) = \{centroid(\mathbf{s}_{i1}), \cdots, centroid(\mathbf{s}_{iK_i})\}, i = 1, \cdots, |C| \tag{2}$$

**Influence Representation** The first two representations do not consider the importance of component words of categories. However, different words belonging to a certain category may have different influence on the category. Those mentions occurring more frequently in $\mathcal{U}$ generally are more prominent and representative for their categories. For example, since mentions *cancer* and *tumor* representing certain diseases occur in $\mathcal{U}$ frequently, we consider they are more representative for category *Disease* and those mentions related to them closely such as *Carcinoma* are more likely be recognized as *Disease*. We define the influence factor $if(w_{ij})$ of each word $w_{ij} \in \mathbf{c}_i$ as the normalized frequency of the mention that it belongs to[2] occurring in $\mathcal{U}$. Then we represent each category as the *weighted mean vector* of word embeddings of its category-words set:

$$\mathbf{R_i}(c_i) = \frac{1}{|\mathbf{c}_i|} \sum_{j=1}^{|\mathbf{c}_i|} if(w_{ij}) \cdot \mathbf{v}_{w_{ij}}, i = 1, \cdots, |C| \tag{3}$$

**Context Representation** Our last category representation bases on following assumption: contexts of each mention occurring in $\mathcal{U}$ embrace rich information and patterns which are helpful to recognize the entity mention. For example, "*the effect of* $\cdots$" is always followed by a *drug name*. Therefore, adding context information into category representation will be useful. We consider a fixed length of window for each mention: two previous words and two following words in $\mathcal{U}$. Then we construct a set of context words for each category $\mathbf{cw}_i = \{cw_{i1}, \cdots, cw_{i|\mathbf{cw}_i|}\}$ where $cw_{ij}$ denotes a bigram or unigram context word occurring over a certain number of times in $\mathcal{U}$ (e.g. 50). Then we represent each category $c_i$ as the *mean vector* of the set of its context words:

$$\mathbf{R_c}(c_i) = \frac{1}{|\mathbf{cw}_i|} \sum_{j=1}^{|\mathbf{cw}_i|} \mathbf{v}_{cw_{ij}}, i = 1, \cdots, |C| \tag{4}$$

where $\mathbf{v}_{cw_{ij}}$ denotes the embedding vector of a context word $cw_{ij}$.

In summary, we learn 4 representations $\mathbf{R}_*(c_i), * \in \{\mathbf{o}, \mathbf{m}, \mathbf{i}, \mathbf{c}\}$ for each predefined category $c_i$ which capture the four different semantic information of it.

[2] If one word belongs to multiple mentions, we simply choose the one with highest frequency.

### 3.3 Generating CMR Features

We first calculate 4 kinds of semantic relatedness between target word $w_j$ and a category $c_i$ based on CMR by leveraging a *distance function* such as *Euclidean distance* as follows.

$$
\begin{aligned}
&d_{\mathbf{o}}(w_j, \mathbf{R_o}(c_i)) = dist(\mathbf{v}_{w_j}, centroid(\mathbf{s}_j)) \\
&d_{\mathbf{m}}(w_j, \mathbf{R_m}(c_i)) = \min_{k \in [1,\cdots,K_i]} dist(\mathbf{v}_{w_j}, centroid(\mathbf{s}_{ik})) \\
&d_{\mathbf{i}}(w_j, \mathbf{R_i}(c_i)) = \frac{1}{|\mathbf{c}_i|} \sum_{k=1}^{|\mathbf{c}_i|} if(w_{ik}) \cdot dist(\mathbf{v}_{w_j}, \mathbf{v}_{w_{ik}}) \\
&d_{\mathbf{c}}(w_j, \mathbf{R_c}(c_i)) = \frac{1}{|\mathbf{cw}_i|} \sum_{k=1}^{|\mathbf{cw}_i|} dist(\mathbf{v}_{w_j}, \mathbf{v}_{cw_{ik}})
\end{aligned}
\tag{5}
$$

Then we define a *threshold* of each category for each CMR based on the distances between the annotated word and each representation of its corresponding category, which is selected with the optimization objective to maximize $F_\beta$-score.

$$
\tau_*(c_i) = \arg\max_{t^* \in V}((1+\beta^2)\frac{P^{(t^*)} \cdot R^{(t^*)}}{(\beta^2 \cdot P^{(t^*)}) + R^{(t^*)}}), * \in \{\mathbf{o}, \mathbf{m}, \mathbf{i}, \mathbf{c}\} \tag{6}
$$

where $P$ is precision and $R$ is recall; $V = (0, 0.01, 0.02, \cdots, 1)$; $\beta$ determines the weight that should be given to recall relative to precision. The lowest threshold $\tau_*(c_i)$ is chosen that optimizes the $F_\beta$-score.

Finally, we generate one feature per representation of each predefined category. The value of the feature is either *True* or *False* depending on whether the calculated *distance* is above the threshold $\tau_*(c_i)$ or not.

$$
f_*^{c_i}(w_j) = \begin{cases} 0 \text{ if } f_*(w_j, \mathbf{R}_*(c_i)) > \tau_*(c_i) \\ 1 \text{ if } f_*(w_j, \mathbf{R}_*(c_i)) \leq \tau_*(c_i) \end{cases}, * \in \{\mathbf{o}, \mathbf{m}, \mathbf{i}, \mathbf{c}\} \tag{7}
$$

## 4 Experiments

### 4.1 Data Sets

To the best of our knowledge, the annotated corpora of the i2b2/VA 2010 shared task (i2b2) and ShARe/CLEF eHealth 2013 Shared Task (ShARe) are the only two public available datasets for CNER. Table 1 and 2 show the statistics of these two datasets respectively. In i2b2, 3 different categories have been annotated: *Problem* (Pr), *Treatment* (Tr), *Test* (Te) from discharge summaries and progress notes. ShARe involves annotation of disorder mentions in a set of narrative clinical reports. Since ShARe does not provide the exact category of each disorder mention, we map each disorder mention into a category by ourselves according to its linking UMLS CUI (Concept Unique Identifier)[3]. Then we get 11 different

[3] In UMLS, each concept (entity) is represented by its CUI and is semantically classified into one of semantic types.

semantic types for this dataset and merge them into 5 categories: *Anatomical Abnormality* (AA), *Pathologic Function* (PF), *Injury or Poisoning* (IP), *Signs and Symptoms* (SS) and *Others*[4](O) according to hierarchies of semantic types in UMLS. Notice these two datasets have totally different categories and our proposed method could work well on both of them which will be demonstrated in following subsections. Two public available corpora are used as unannotated

**Table 1.** The statistics of i2b2

| Dataset | Pr | Tr | Te | All |
|---|---|---|---|---|
| Training | 11968 | 8500 | 7369 | 27837 |
| Test | 18550 | 13560 | 12899 | 45009 |

**Table 2.** The statistics of ShARe

| Dataset | AA | PF | IP | SS | O | All |
|---|---|---|---|---|---|---|
| Training | 250 | 2304 | 221 | 838 | 1525 | 5138 |
| Test | 157 | 2107 | 96 | 735 | 1535 | 4630 |

clinical data: the 378,000 *Medline abstracts* that are indexed as pertaining to clinical trials and *MIMIC III* that comprises deidentified health data associated with 40,000 critical care patients. Then we build a semantic space by training a word embedding model — GloVe [17] used in this paper — on these two corpora (merged).

### 4.2 Our Models and Parameter Settings

In our experiments, We apply two state-of-the-art sequence labeling models: CRF and BLSTM+CRF (BLSTM for short) with the generated new CMR features to evaluate our method. We implement CRF employing *CRFsuite*[5] and BLSTM using *theano* library.[6] The parameter settings of these two models are showed in Table 3 and 4 respectively.

**Table 3.** CRF settings

| C-value | context window | regularization |
|---|---|---|
| 5 | 2+2 | L1&L2 |

**Table 4.** BLSTM settings

| Layers | Layer size | batch size | activation function | learning rate | drop out | epochs | optimizer |
|---|---|---|---|---|---|---|---|
| 2 | 100 | 64 | RELU | 1E-04 | 0.5 | 100 | adam |

The considered performance metrics are precision, recall and F1-score and we adopt the *strict metrics* for evaluation used in both tasks. Performance scores are macro-averaged over classes, giving equal weight to all classes.

### 4.3 Threshold Settings for Determining CMR Features

We first investigate the impact of providing threshold of CMR that determine the feature values on NER performance. Fig.2 shows the threshold setting procedure for different CMR in which threshold is set by finding the distance that maximizes F1-score on i2b2. It can be seen that the thresholds are generally lower and the F1-scores higher for Multi-sub-center Representation and One-center Representation (also observed on ShARe). It indicates these two representations

[4] For those mentions mapping to unknown CUI, i.e. CUI-less.
[5] http://www.chokkan.org/software/crfsuite/
[6] http://deeplearning.net/software/theano/

are better to separate the categories and important to capture the meaning of a category. This is confirmed in the subsequent experiments, the results of which show that the highest performance is obtained with these two representations.

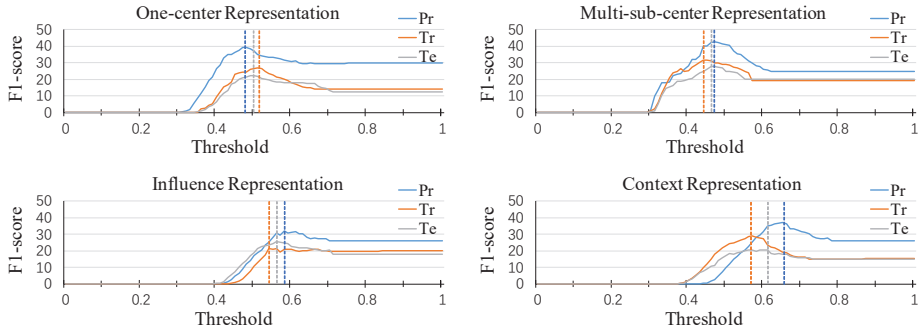


**Fig. 2.** Threshold Setting Procedure for CMR on i2b2.

To study the impact of changing the optimization objective to various $F_\beta$-scores on NER performance, experiments are conducted with the following $\beta$ values: 0.5, 1.0, 2.0 and 5.0. The highest F1-score are observed when $\beta$ is set to 1.0 in i2b2 and 2.0 in ShARe. Then in our following experiments, we set $\beta$=1 for i2b2 and $\beta$=2 for ShARe.

### 4.4 Comparison with Different CMR Features

In order to study and verify the effectiveness of the proposed new CMR features to the learning algorithms, our four groups of CMR features are evaluated and compared one by one. For CRF model, we combine our CMR features with a set of traditional features — orthographic and syntactic features[7] — as our baseline which is the most traditional method in CNER. For BLSTM, we take the word embedding concatenating character embedding as the baseline — which is state-of-the-art in general domain of NER task. Table 5 and 6 show the comparison of different feature combinations on two datasets respectively.

Two traditional state-of-the-art models without leveraging lexicons (baselines) perform not well on both datasets. When we add our CMR features to these models one by one, the experiment results show each group of CMR features achieves improvement on both datasets. We can see the Multi-sub-center Representation features achieve the best improvement among all CMR features while the improvement obtained from Context Representation and Influence Reresentation features are relatively small. This indicates that Multi-sub-center Representation is more representative than other CMR and could more holistically capture the meaning of the category. When we combine all CMR features,

[7] The same as the ones used in [10] except lexical features extracted from existing annotated tools.

**Table 5.** Comparison with different CMR features on i2b2

| CMR | CRF | | | | BLSTM | | | |
|---|---|---|---|---|---|---|---|---|
| | P | R | F1 | ΔF1 | P | R | F1 | ΔF1 |
| baseline | 82.05 | 78.86 | 80.42 | | 82.20 | 81.57 | 81.88 | |
| $+f_{\mathbf{o}}$ | 84.69 | 78.35 | 81.40 | +0.97 | 84.29 | 81.70 | 82.97 | +1.09 |
| $+f_{\mathbf{m}}$ | 83.45 | 80.01 | 81.69 | +1.27 | 83.92 | 83.07 | 83.50 | +1.61 |
| $+f_{\mathbf{i}}$ | 82.77 | 79.11 | 80.90 | +0.48 | 83.40 | 81.67 | 82.53 | +0.65 |
| $+f_{\mathbf{c}}$ | 82.35 | 79.54 | 80.92 | +0.50 | 83.26 | 82.48 | 82.87 | +0.98 |
| $+f_{all}$ | 83.92 | 80.12 | 81.98 | +1.55 | **84.48** | **83.39** | **83.93** | +2.05 |

**Table 6.** Comparison with different CMR features on ShARe

| CMR | CRF | | | | BLSTM | | | |
|---|---|---|---|---|---|---|---|---|
| | P | R | F1 | ΔF1 | P | R | F1 | ΔF1 |
| baseline | 74.22 | 61.16 | 67.06 | | 73.93 | 66.11 | 69.80 | |
| $+f_{\mathbf{o}}$ | 75.95 | 61.06 | 67.70 | +0.64 | 75.32 | 66.31 | 70.53 | +0.73 |
| $+f_{\mathbf{m}}$ | 75.34 | 62.12 | 68.09 | +1.04 | 75.31 | 67.35 | 71.11 | +1.31 |
| $+f_{\mathbf{i}}$ | 74.99 | 61.34 | 67.48 | +0.42 | 74.83 | 66.36 | 70.34 | +0.54 |
| $+f_{\mathbf{c}}$ | 74.79 | 61.58 | 67.55 | +0.49 | 74.25 | 66.95 | 70.41 | +0.61 |
| $+f_{all}$ | 75.71 | 62.32 | 68.37 | +1.31 | **75.74** | **67.93** | **71.62** | +1.82 |

we achieve further significant improvement on both datasets (1.55% improvement of CRF and 2.05% improvement of BLSTM on i2b2 as well as 1.31% and 1.82% on ShARe) that indicates the four groups of CMR features could compensate each other and combination of them could further improve the performance. We also find our CMR features get more improvement for BLSTM (2.05% on i2b2 and 1.82% on SHARe) than CRF and achieve the best performance on both datasets. The possible reason is that our CMR features are derived from word embedding and could work better when combining with it. Furthermore, in addition to powerful capability of BLSTM model, features used in BLSTM including word embedding and CMR features are semantic, without considering orthographic and syntactic features, which could potentially more effectively address the challenge of informal clinical texts.

### 4.5 Comparison with Previous Systems

Our evaluation show that the performance of NER significantly improves after adding our new CMR features. However, how much it contributes toward improving the state-of-the-art determines the practical significance of the improvement. Thus, we compare the performance of our method to the top systems in the i2b2/VA 2010 concept extraction task and ShARe/CLEF eHealth 2013 Shared Task.

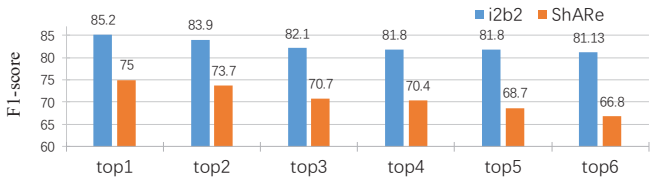


**Fig. 3.** Comparison with top 6 systems in two shared tasks.

Fig.3 shows the results of the top 6 systems in these two tasks. Almost all systems use hybrid models integrating several models such as CRF and SSVM with a set of rich features. Furthermore, all systems leverage the output of existing annotation tools such as cTAKEs, MetaMap and rely lexicons derived from UMLS to improve the NER performance. Our best method (BLSTM + 4 CMR) is better than system 3 and equal to system 2 in i2b2 and ranks the third in ShARe. The results suggest that by integrating CMR features derived from a large scaled unlabeled corpus into one single model our work can achieve state-of-the-art without leveraging any outside lexicons and any existing annotated tools. In addition, our CMR features can easily integrate with other models such as discriminative semi-Markov HMM Models used by the best system in i2b2. It may further improve the performance of these systems.

## 5 Related Works

Most early existing NER techniques in medical domain typically focus on traditional machine learning methods such as Support Vector Machine (SVM) [24], Hidden Markov Model (HMM) [22] and Conditional Random Fields (CRF) [21] integrating a set of complicated hand-crafted features. Some other methods leverage hybrid models [6, 10] to improve the performance of NER. However, their performance may be affected by some common drawbacks: 1) with the change of corpora and languages, the process to reconstruct feature set is difficult; 2) some complex features with syntactic information rely on the performance of other NLP modules; 3) these features with expert knowledge are expensive to acquire.

There also exist some well-known annotation tools in clinical texts such as cTAKEs [20], MetaMap [2] and ConText [7]. Most of them can extract various types of named entities from clinical texts and link them to concepts in UMLS. However, these tools heavily rely on external dictionaries such as SNOMED-CT [9] and are only suitable for English. A large amount of works [6, 10, 5] usually leverage the annotating results of these tools as a part of features to feed into their models and achieve further improvement of performance.

Another thread of NER in medical domain focuses on recognizing one single named entity, such as [27] finding anatomies from discharge summaries, [19, 14, 28] recognizing drug names and [26] extracting disease names from clinical texts. Different with these clinical NER works addressing single entity type, we are addressing a comprehensive set of challenges in identifying multiple named entities to analysis the clinical texts.

Recently, some attempts [29, 13, 8] focus on applying deep neural network to NER in clinical texts. Most of these concatenate word-level embedding, character-level embedding and lexicon embedding as input. Then multiple convolutional layers are stacked over the input to extract useful features automatically and then fed into RNN models. Although these methods claim no feature engineering, their performance are heavily rely on the training dataset (also rely on lexicons) and usually not satisfied when the training set is small. Since our proposed CMR features are derived from large scale unannotated corpus, our method reduce the

limitation of small training set and is easy to be adapted to new domains while large scale unannotated corpora are often readily available.

## 6 Conclusion and Future Work

The existing CNER systems simply follow the traditional NER methods used in general domain which usually leverage the linguistic features including syntactic and lexicon features. Compared with successful performance of NER in general domain, CNER achieves relatively pool performance due to the issues of *Limited Availability of Annotated Data* and *Informal clinical texts*. In this paper, we propose a novel unified method for CNER without considering any linguistic features. It learned multiple representations for each category to capture the semantic similarity between words and categories from 4 different perspectives In the future, we will evaluate our method in other domains, such as biomedical domain. In addition, we will explore new unsupervised methods that is useful when training dataset is not available.

**Acknowledgements** The work is supported by major national research and development projects (2017YFB1002101), NSFC key project (U1736204, 61661146007), Fund of Online Education Research Center, Ministry of Education (No. 2016ZD102), and THU-NUS NExT Co-Lab.

## References

1. Abacha, A.B., Zweigenbaum, P.: Medical entity recognition: A comparison of semantic and statistical methods. In: BioNLP. pp. 56–64 (2011)
2. Aronson, A.R., Lang, F.M.: An overview of metamap: historical perspective and recent advances. JAMIA 17, 229–236 (2010)
3. Bodenreider, O.: The unified medical language system (umls): Integrating biomedical terminology (2004)
4. Bodnari, A., Deléger, L., Lavergne, T., Névéol, A., Zweigenbaum, P.: A supervised named-entity extraction system for medical text. In: Working Notes for CLEF Conference (2013)
5. Bodnari, A., Deléger, L., Lavergne, T., Névéol, A., Zweigenbaum, P.: A supervised named-entity extraction system for medical text. In: CLEF (2013)
6. de Bruijn, B., Cherry, C., Kiritchenko, S., Martin, J., Zhu, X.: Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. Journal of the American Medical Informatics Association 18(5), 557 (2011)
7. Chapman, W.W., Chu, D., Dowling, J.N.: Context: An algorithm for identifying contextual features from clinical text. pp. 81–88. BioNLP '07 (2007)
8. Dernoncourt, F., Lee, J.Y., Uzuner, Ö., Szolovits, P.: De-identification of patient notes with recurrent neural networks. JAMIA 24(3), 596–606 (2017)
9. Donnelly, K.: Snomed-ct: The advanced terminology and coding system for ehealth. Studies in health technology and informatics 121, 279–90 (2006)
10. Jiang, M., Chen, Y., Liu, M., Rosenbloom, S.T., Mani, S., Denny, J.C., Xu, H.: A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. JAMIA 18, 601–606 (2011)
11. Johnson, A.E.W., Pollard, T.J., Shen, L., Lehman, L.w.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., Mark, R.G.: Mimic-iii, a freely accessible critical care database. Scientific Data 3 (2016)

12. Kundeti, S.R., Vijayananda, J., Mujjiga, S., Kalyan, M.: Clinical named entity recognition: Challenges and opportunities. In: 2016 IEEE International Conference on Big Data, 2016. pp. 1937–1945 (2016)
13. Li, L., Jin, L., Jiang, Z., Song, D., Huang, D.: Biomedical named entity recognition based on extended recurrent neural networks. In: BIBM. pp. 649–652 (2015)
14. Liu, S., Tang, B., Chen, Q., Wang, X.: Effects of semantic features on machine learning-based drug name recognition systems: Word embeddings vs. manually constructed dictionaries. Information 6(4), 848–865 (2015)
15. Meystre, S.M., Savova, G.K., Kipper-Schuler, K.C., Hurdle, J.F.: Extracting information from textual documents in the electronic health record: a review of recent research. Yearbook of medical informatics pp. 128–44 (Jan 2008)
16. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems 26, pp. 3111–3119 (2013)
17. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: EMNLP. vol. 14, pp. 1532–1543 (2014)
18. Ratinov, L., Roth, D.: Design challenges and misconceptions in named entity recognition. In: CoNLL (6 2009)
19. Sadikin, M., Fanany, M.I., Basaruddin, T.: A new data representation based on training data characteristics to extract drug name entity in medical text. Computational Intelligence and Neuroscience 2016, 16 (2016)
20. Savova, G.K., Masanz, J.J., Ogren, P.V., Zheng, J., Sohn, S., Kipper-Schuler, K.C., Chute, C.G.: Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. JAMIA 17(5), 507–513 (2010)
21. Settles, B.: Biomedical named entity recognition using conditional random fields and rich feature sets. pp. 104–107. JNLPBA (2004)
22. Shen, D., Zhang, J., Zhou, G., Su, J., Tan, C.L.: Effective adaptation of a hidden markov model-based named entity recognizer for biomedical domain. pp. 49–56. BioMed (2003)
23. Suominen, H., Salanterä, S., Velupillai, S., Chapman, W.W., Savova, G.K., Elhadad, N., Pradhan, S., South, B.R., Mowery, D., Jones, G.J.F., Leveling, J., Kelly, L., Goeuriot, L., Martínez, D., Zuccon, G.: Overview of the share/clef ehealth evaluation lab 2013. In: CLEF. vol. 8138, pp. 212–231 (2013)
24. Takeuchi, K., Collier, N.: Bio-medical entity extraction using support vector machines. pp. 57–64. BioMed (2003)
25. Uzuner, ., South, B.R., Shen, S., DuVall, S.L.: 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association 18(5), 552 (2011)
26. Wei, Q., Chen, T., Xu, R., He, Y., Gui, L.: Disease named entity recognition by combining conditional random fields and bidirectional recurrent neural networks. Database 2016 (2016)
27. Xu, Y., Hua, J., Ni, Z., Chen, Q., Fan, Y., Ananiadou, S., Chang, E.I.C., Tsujii, J.: Anatomical entity recognition with a hierarchical framework augmented by external resources. PLOS ONE 9, 1–13 (10 2014)
28. Zeng, D., Sun, C., Lin, L., Liu, B.: Enlarging drug dictionary with semi-supervised learning for drug entity recognition. In: BIBM. pp. 1929–1931 (2016)
29. Zhao, Z., Yang, Z., Luo, L., Zhang, Y., Wang, L., Lin, H., Wang, J.: Ml-cnn: A novel deep learning based disease named entity recognition architecture. In: BIBM. p. 794 (2016)