Informatics in Medicine Unlocked 13 (2018) 122–127

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

# Learning for clinical named entity recognition without manual annotations

Omid Ghiasvand, Rohit J. Kate*

*Department of Health Informatics and Administration, University of Wisconsin-Milwaukee, Milwaukee, WI, USA*

ARTICLE INFO

*Keywords:*
Named entity recognition
Clinical text
Machine learning
Natural language processing
Information extraction

ABSTRACT

*Background:* Named entity recognition (NER) systems are commonly built using supervised methods that use machine learning to learn from corpora manually annotated with named entities. However, manually annotating corpora is very expensive and laborious.

*Materials and methods:* In this paper, a novel method is presented for training clinical NER systems that does not require any manual annotations. It only requires a raw text corpus and a resource like UMLS that can give a list of named entities along with their semantic types. Using these two resources, annotations are automatically obtained to train machine learning methods. The method was evaluated on the NER shared-task datasets of i2b2 2010 and SemEval 2014.

*Results:* On the SemEval 2014 dataset for recognizing diseases and disorders, the method obtained F-measure of 0.693 for exact matching and of 0.773 allowing overlaps. This is comparable to many supervised systems in the past that had used manual annotations for training. On the i2b2 2010 dataset for recognizing problems, tests and treatments, the method obtained F-measures of 0.451, 0.338 and 0.204 respectively for exact matching and of 0.721, 0.587 and 0.475 respectively allowing overlaps. These results are better than an existing unsupervised method.

*Conclusions:* Experiments on standard datasets showed that the new method performed well. The method is general and could be applied to recognize entities of other types on other genres of text without needing manual annotations.

## 1. Introduction

Named entity recognition (NER) is an important task and is often an essential step for many downstream natural language processing (NLP) applications [1,2]. Many early systems were rule-based that required a lot of manual effort and expertise to build and were often brittle and not very accurate, hence most successful NER systems are currently built using supervised methods [3–5]. To build these systems, first a corpus is manually annotated with named entities of the desired type. Then machine learning methods are trained using the annotated corpus to automatically recognize named entities in new text. However, annotating corpora manually is laborious and expensive, particularly so in the clinical domain in which expensive clinical expertise is required for annotating clinical text [6]. Another disadvantage of manual annotation is that one requires new or additional manual annotations every time one wants to build an NER system for a new genre of text or for a new named entity type.

As an alternate to supervised methods, researchers have developed unsupervised methods for NER in the general NLP domain. Such methods typically use existing dictionaries or gazetteers of known named entities to match in text along with mechanisms for disambiguation [7,8] and/or bootstrapping [9–12]. Although several unsupervised NER systems have been developed in the general NLP domain, those methods do not directly apply to the clinical domain because NER in the clinical domain differs from NER in the general domain in two important ways. First, in the general domain, named entities, for example, locations, company names, or person names, typically do not have linguistic variations. In contrast, in the clinical domain there are usually several ways to mention the same named entity, for example, “left ventricular hypertrophy”, “left ventricle is hypertrophic”, “hypertrophy of left ventricle” all refer to the same named entity. Such variability makes it difficult for matching-based unsupervised methods to work well in the clinical domain. Second, clinical named entities are often multi-token terms with nested structures that include other named entities inside them, for example, “pleuropericardial chest pain” is an entity of the class disease/disorder and includes within it entities “chest pain” and “pain” of the same class as well as includes entities “pleuropericardial” and “chest” of the body

* Corresponding author. Department of Health Informatics and Administration, University of Wisconsin-Milwaukee, 2025 E Newport Ave, Milwaukee, WI 53211, USA.

*E-mail address:* katerj@uwm.edu (R.J. Kate).

https://doi.org/10.1016/j.imu.2018.10.011
Received 7 October 2018; Received in revised form 29 October 2018; Accepted 29 October 2018
Available online 30 October 2018

structure class. This makes determining the right boundaries of clinical named entities challenging and any unsupervised NER method needs to address correct boundary determination which is not needed in the general NLP domain. As a result, building unsupervised NER systems is more difficult in the clinical domain and very few researchers have built such systems.

One such unsupervised NER system by Zhang and Elhadad [13] was based on the observation that named entities in the same class tend to have similar vocabulary and occur in similar contexts. Their method first begins with a set of known terms for the target entity class which are obtained from a resource like UMLS [14] and are called seed terms. These are then matched wherever they occur as noun phrase chunks in a biomedical corpus. A signature is then created for each seed term in the form of a vector representation based on the inverse document frequencies (IDF) of the words occurring within the term and the words occurring in the contexts in which the term occurred in the corpus. The context of a term occurrence is defined as the previous and the next two words. A signature of the target entity class is then computed by averaging the signature of all its seed terms. During testing, the method first computes signature for a candidate entity and then computes its cosine similarity with the signatures of all the entity classes. The candidate entity is assigned the entity class with which it has the highest similarity provided the similarity is above an experimentally determined threshold.

In this paper, we present a novel method to build clinical NER systems which does not require any manual annotations, but it differs from the method by Zhang and Elhadad [13] in the following major ways. Unlike their method which did not employ machine learning, our method uses machine learning models which are trained on automatically annotated corpus followed by self-training iterations. While their method computes only one representative signature for an entire class of entities thus expecting that all entities of that class will conform to it, our method can learn to recognize named entities that could occur in varying contexts. Another major difference is that their method ignores the possibility that the seed terms could have multiple senses and may not be of the target entity class when found in a corpus, but our method makes sure that only unambiguous terms are used for creating automatic annotations. One more difference is that while their method considers only noun phrase chunks in order to determine candidate named entities, a limitation also pointed out by them [13], our method considers all noun phrases, including nested ones, as obtained through full parsing. We experimentally compare our method with theirs and demonstrate the advantages offered by our method.

A few unsupervised NER systems have been developed in languages other than English. Recently, Xu et al. [15] built an unsupervised NER system for Chinese medical text that leveraged syntactic knowledge, corpus statistics and lexical resources. However, their system is not built for doing NER for clinical text but is built for doing NER for online text of question and answering (Q&A) found on Chinese medical websites. Because of a different language as well as a different genre of text, their method addresses very different challenges than encountered while doing NER for clinical text in English. Oronoz et al. [16] built a system to automatically annotate medical records in Spanish. Their approach was mainly based on adding Spanish medical information in the form of medical terminologies enriched with Spanish terms to a standard linguistic analyzer [17]. Unlike our method, none of these methods had employed machine learning.

We note that for the task of extracting named entities from biological literature (as opposed to clinical text) there have been a few systems in which learning methods were trained using automatically generated annotations [18,19], but these were not designed for or applied to clinical text for extracting clinical named entities. Extracting named entities (such as gene names) from biological literature is very different from extracting named entities (such as disease names) from clinical text [3]. We point out that Zhang and Elhadad [13] had also applied their method for extracting named entities from biological literature but did not obtain as good results as they obtained for extracting named entities from clinical text due to lack of available seed terms as pointed out by them. In the biomedical domain, for the task of relation extraction (as opposed to the task of NER) there also have been systems that were trained using automatically generated annotations [20] or that did not use any supervision [21], but relation extraction is a different task from NER task which is the focus of this study. The latter work [21] was for clinical text in Italian and had used dictionaries and rules for the NER part of their system.

We want to point out the difference between unsupervised NER and learning for NER without manual annotations. In unsupervised NER, a system does not use supervised machine learning hence it does not need any annotated training data. In contrast, learning for NER without manual annotations means that the system uses supervised machine learning, however, it obtains training data automatically without requiring any manual effort. The system presented in this paper learns to do NER without manual annotations, while the system by Zhang and Elhadad [13] was an unsupervised NER system. Please note that rule-based systems are of neither type because although they do not require manual annotations, building them requires significant manual effort along with domain and linguistic expertise which is often more expensive than obtaining manual annotations.

In past, some techniques have been developed to reduce manual annotation effort, although they still require some amount of manual annotations. In active learning, which has been also applied to clinical NER [6,22], the learning process does not begin with a large annotated corpus but grows the corpus interactively and wisely by making the machine learning system itself ask for the most helpful examples that the human should annotate. This reduces the manual effort by avoiding annotating needless examples. Another technique is called pre-annotation [23] in which a system first automatically annotates a corpus then a human corrects the mistakes made by the system. This can save significant human annotation effort provided that the pre-annotated corpus is of good quality. However, both active learning and pre-annotation require some amount of manual annotations in contrast to the method presented in this paper which does not require any manual annotations.

The contribution of this work is significant because the presented method could greatly reduce the manual effort and the cost of building clinical NER systems. It is especially relevant if encountering new genre or style of text (say, unique to a particular medical center) or a new named entity type for which new manual annotations will be otherwise needed to employ supervised methods or new rules will have to be manually written to build rule-based systems. The presented method is not language-specific and given that UMLS also includes terms of many other languages besides English, the method is also directly applicable for building clinical NER systems in other languages Finally, if manual annotations are available for an NER task but is not sufficient, then the method can be used to further improve clinical NER systems in a semi-supervised framework by providing more annotations automatically.

## 2. Materials and methods

Our method requires two resources – a raw corpus and a list of named entities along with their semantic types which can be obtained from a resource like UMLS. The method has two components: detecting presence of named entities in text and determining their correct boundaries. For each part, we train a machine learning classifier using automatic annotations as described in the following subsections. Fig. 1 gives an overview of the training process and Fig. 2 shows how the trained system is applied for NER.

### 2.1. Named entity detection

For the purpose of describing our method, we will assume that the NER task is for recognizing named entities of disease/disorder semantic

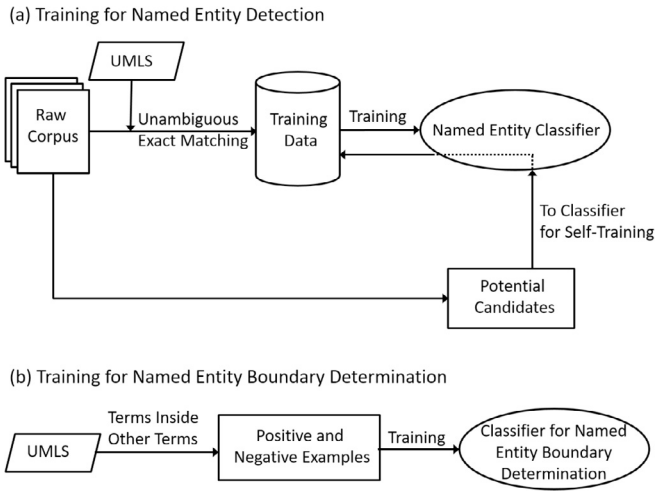


**Fig. 1.** Overview of the training process for (a) Named Entity Detection (b) Named Entity Boundary Determination. The training requires a raw corpus and a resource like UMLS but does not require any manual annotations.

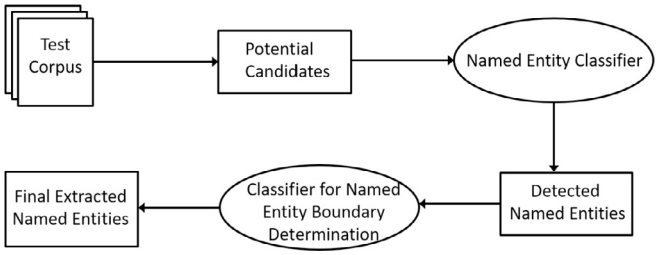


**Fig. 2.** Overview of applying the trained system for the named entity recognition task.

type [24], the same method can be otherwise applied for named entities of any other semantic type. Our method first automatically annotates a raw corpus with *unambiguous* named entities from UMLS of disease/disorder semantic type wherever they match exactly in the raw corpus. For example, the disease name "meningitis" is unambiguous because that name does not have any other semantic type in UMLS, hence if it occurs anywhere in the raw text we can be sure that it means only the disease (it is possible that it could be inside a larger term but this part of our method is only for detecting entities; the next part is for determining the correct boundaries). It will be thus automatically annotated as a positive example of disease/disorder named entity. Please note that the method at this stage will miss the diseases/disorders which either have multiple semantic types (for example, "distress" which could be a disease or a physiological process) or which do not match exactly (for example, UMLS lists "left atrium dilation" but the text may have "left atrium dilated"). With such incomplete annotated data, one cannot use sequence labeling based machine learning methods for NER, such as conditional random fields [25] (CRF), which label every word in the text as part of a named entity or not. This is because sequence labeling methods implicitly treat all the unannotated words as negative examples during training. Hence the named entities missed getting annotated will be incorrectly taken as negative examples which would adversely affect the training. Therefore instead of using a sequence labeling method, we use a classification method for named entity detection that is trained to classify a term found in the text as either of the required semantic type or not. In order to train such a classifier we also need to give it explicit negative examples.

The method automatically annotates negative examples as well which are terms in the raw corpus which one can be sure are not diseases/disorders. These are the named entities that match exactly in UMLS and are of semantic types other than disease/disorder. These

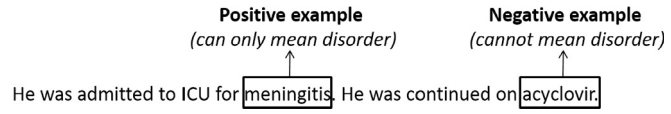


**Fig. 3.** Examples of automatic annotation to train a classifier for detecting named entities of disease/disorder semantic type. Meningitis matches in UMLS and is of only one semantic type of disease/disorder hence it is automatically annotated as a positive example. On the other hand, acyclovir also matches in UMLS but is not of the semantic type of disease/disorder hence it is automatically annotated as a negative example.

terms could be ambiguous and have multiple semantic types but none of them can be of disease/disorder semantic type. Fig. 3 shows an example of automatically annotating a given text with positive and negative examples. With this preliminarily annotated corpus of positive and negative examples, the method trains a classifier that learns to classify a given term in the corpus to be a disease/disorder or not from its surrounding context. We used three words within the sentence before and after the entity, their lemmatized forms, their stemmed forms, their part-of-speech (POS) tags and their UMLS semantic types as features. We used Weka software's [26] ensemble of decision trees obtained using the random-subspace method [27] as our classifier which we found to work best through pilot studies done within the training data. Next, as part of the self-training process, the trained classifier is applied back to the raw corpus to obtain more annotations and is retrained. This is done a few times until no new examples are gathered. In our experiments six iterations were found to be sufficient. The new annotations will now also include terms which could have multiple semantic types in UMLS (for example, "distress") but the method always makes sure during retraining that the desired semantic type is one of them in order to avoid generating incorrect annotations.

In order to recognize named entities, unlike a trained sequence labeling based NER method, it is not an efficient option for a classifier based NER method to be applied to entire text because it will need to be applied to all possible substrings of words. Hence we restrict applying the classifier only to potential named entity candidates which are noun phrases with at least one biomedical word. A word is deemed to be biomedical if it is present in any term in UMLS. In our method, POS tags and noun phrases were obtained using the Stanford parser [28].

### 2.2. Named entity boundary determination

While the above trained classifier is good at detecting presence of named entities of the desired semantic type in text, it is not good at determining their exact boundaries. For example, if "congenital heart disease" is mentioned in the text, it is possible that the above method may detect only "heart disease" as the disease/disorder. The second part of our method is designed to determine the correct boundaries of the detected named entities. It is also trained leveraging UMLS and without requiring any manual annotations. This part does not even require a raw corpus. The terms in UMLS of the desired semantic type that include another term within them of the same semantic type are deemed as positive examples. For example, "acute duodenal ulcer" is a disease/disorder and includes "ulcer" which is also a disease/disorder. Conversely, terms in UMLS of other semantic types that include a term within them of the desired semantic type are deemed as negative examples. For example, "excision of ulcer of stomach" is a procedure but includes "ulcer" which is a diseases/disorder, hence it will be a negative examples (of note, "ulcer of stomach" will be a positive example). Fig. 4 illustrates these examples that are automatically obtained from UMLS.

A classifier (using same machine learning method [27] as before) is then trained using these automatically collected positive and negative examples with the words along with their positions in the named entity, their lemmatized forms, their stemmed forms, their POS tags and their UMLS semantic types as features. The classifier thus learns what type of

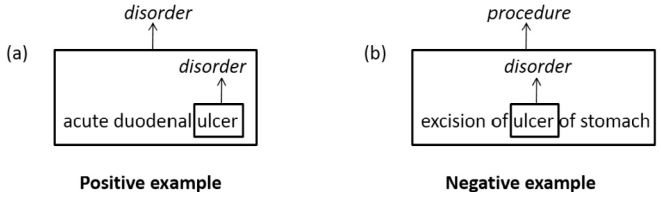


**Fig. 4.** Automatically obtained examples from UMLS to train the method for named entity boundary determination for disease/disorder semantic type. The UMLS term "acute duodenal ulcer" is of the type disease/disorder and includes another UMLS term "ulcer" of the same type hence it is deemed as a positive example for training the method to extend boundary of a disease/disorder term. The UMLS term "excision of ulcer of stomach" is of a different semantic type but includes term "ulcer" of disease/disorder semantic type hence it is deemed as a negative example for training the method to extend boundary of a disease/disorder term.

words can expand boundary of a disease/disorder. For example, it may learn that the presence of the words "acute" and "duodenal" extends the boundary of a disease/disorder on its left side. The method can thus also recognize named entities not already in UMLS. The method may also learn that the presence of the word "excision" cannot extend the boundary of a disease/disorder on its left side. The classifier is applied to all noun phrases in which the detected named entity occurs and the noun phrase that gives the highest classification score is selected as a recognized named entity if it is above an experimentally determined threshold.

## 3. Experimental methodology

We evaluated our method on two benchmark datasets that had been previously used for clinical NER shared-tasks. The first dataset was from the SemEval 2014 NER shared-task (Task 7A) [24] for recognizing disease/disorder semantic type. This semantic type was composed of total eleven UMLS semantic types that included disease or syndrome, signs and symptoms, and pathologic function. This dataset contained a mix of four types of clinical reports – discharge summaries, echocardiogram reports, electrocardiograph reports and radiology reports. There were total 199 clinical reports for training and 133 clinical reports for testing. Given that our method does not require any manual annotations, the clinical reports meant for training were used only in their raw unannotated forms along with 5000 additional unannotated clinical reports randomly selected from a very large set of unannotated clinical reports that were provided to the participants of the shared-task. We had experimentally found that around 5000 unannotated clinical reports were sufficient for training our method. The second dataset was from i2b2 2010 shared-task [29] for recognizing named entities of problems, treatments and tests semantic types. This dataset contained 349 discharge summaries for training and 477 discharge summaries for testing. The training data was again used only in its raw unannotated form in addition to the 5000 unannotated clinical reports mentioned before. The threshold mentioned in the previous section was set to 0.7 for the disease/disorder named entity type for the SemEval 2014 dataset, to 0.6 for problems and tests named entity types for the i2b2 2010 datatset, and to 0.2 for its treatment named entity type. These were found through pilot studies using a small portion of the training dataset.

The performance was evaluated using the standard measures of precision (fraction of extracted named entities that are correct), recall (fraction of named entities extracted out of all the gold-standard named entities) and F-measure (harmonic mean of precision and recall) [30]. Recognizing that an NER system could sometimes extract only portion of a named entity from text and miss the rest, two evaluation scoring schemes were used – strict and relaxed, as was also done in the shared-tasks. The strict scoring scheme only counts perfect matches, so a system gets zero credit if the extracted entity has any extra token or is missing even a single token when compared to the gold-standard entity. The relaxed scheme, on the other hand, gives credit for partial matching according to the span of the extracted entity that overlaps with the span of the gold-standard entity.

**Table 2**
Comparison of overall performance across all entity classes on the i2b2 2010 NER dataset. Our method is compared with MetaMap system [31], system by Zhang and Elhadad [13] (ZE'13), and the best performing supervised system in i2b2 2010 shared-task [29].

| *Method* | Overall Strict F-score | Overall Relaxed F-score |
|---|---|---|
| *Our method* | 0.331 | 0.594 |
| *MetaMap* | 0.113 | 0.279 |
| *ZE'13* | 0.265 | 0.531 |
| *Best Supervised* | 0.852 | 0.924 |

## 4. Results and discussion

Table 1 shows our results obtained on the i2b2 2010 dataset. For comparison, we have also shown the results that were obtained by Zhang and Elhadad [13] denoted as ZE'13 in the table. It can be observed that our method obtains better F-scores than their method on all the three entity classes for both strict and relaxed scoring schemes. Table 2 shows the average performance across all entity classes obtained by our system compared with the system by Zhang and Elhadad [13] (ZE'13), the MetaMap system (results replicated from Ref. [13]) and the best performing supervised system from i2b2 2010 shared-task [29]. It can be seen that our system also obtains much better results than the rule-based MetaMap system [31]. However, it is far behind the best supervised system which had used other types of features and deep knowledge resources as was also noted by Zhang and Elhadad [13].

Table 3 shows our results obtained on the SemEval 2014 dataset. We do not know of results obtained on this dataset by any system that did not use manual annotations for training that we could have used for a direct comparison. Hence for the sake of comparison, we have shown results of our own supervised system [32,33] from SemEval 2014 which had ranked 3rd in this shared-task. It used CRF as the machine learning

**Table 1**
Results obtained on i2b2 2010 NER dataset. ZE'13 denotes the results obtained by Zhang and Elhadad [13] which are shown for comparison.

| Entity Class | *Method* | Strict | | | Relaxed | | |
|---|---|---|---|---|---|---|---|
| | | *Precision* | *Recall* | *F-score* | *Precision* | *Recall* | *F-score* |
| Problem | *Our method* | 0.391 | 0.533 | 0.451 | 0.623 | 0.858 | 0.721 |
| | *ZE'13* | 0.267 | 0.317 | 0.291 | 0.492 | 0.715 | 0.583 |
| Test | *Our method* | 0.326 | 0.351 | 0.338 | 0.561 | 0.615 | 0.587 |
| | *ZE'13* | 0.369 | 0.221 | 0.277 | 0.546 | 0.526 | 0.536 |
| Treatment | *Our method* | 0.191 | 0.219 | 0.204 | 0.396 | 0.593 | 0.475 |
| | *ZE'13* | 0.286 | 0.159 | 0.204 | 0.454 | 0.379 | 0.413 |

**Table 3**
Results obtained on SemEval 2014 Task 7A dataset. For comparison, results of a supervised system from our prior work [32] that used the same type of features are also shown.

| Entity Class | *Method* | Strict | | | Relaxed | | |
|---|---|---|---|---|---|---|---|
| | | *Precision* | *Recall* | *F-score* | *Precision* | *Recall* | *F-score* |
| Disease/Disorder | *Our method* | 0.783 | 0.622 | 0.693 | 0.881 | 0.69 | 0.773 |
| | *Supervised Method* | 0.787 | 0.726 | 0.755 | 0.911 | 0.856 | 0.883 |

**Table 4**
Results comparing our NER system with and without the boundary determination component.

| Entity Class | *Boundary Determination* | Strict | | | Relaxed | | |
|---|---|---|---|---|---|---|---|
| | | *Precision* | *Recall* | *F-score* | *Precision* | *Recall* | *F-score* |
| **Problem** (i2b2 2010) | Without | 0.367 | 0.501 | 0.424 | 0.622 | 0.858 | 0.721 |
| | With | 0.391 | 0.533 | 0.451 | 0.623 | 0.858 | 0.721 |
| **Test** (i2b2 2010) | Without | 0.294 | 0.334 | 0.313 | 0.503 | 0.607 | 0.55 |
| | With | 0.326 | 0.351 | 0.338 | 0.561 | 0.615 | 0.587 |
| **Treatment** (i2b2 2010) | Without | 0.174 | 0.22 | 0.194 | 0.381 | 0.588 | 0.462 |
| | With | 0.191 | 0.219 | 0.204 | 0.396 | 0.593 | 0.475 |
| **Disease/Disorder** (SemEval 2014) | Without | 0.78 | 0.571 | 0.659 | 0.884 | 0.65 | 0.749 |
| | With | 0.783 | 0.622 | 0.693 | 0.881 | 0.69 | 0.773 |

method that was trained using the manually annotated training data using the same type of features that we used in our current method which makes it more directly comparable. It can be seen that although lower, the performance of our system that requires no manual annotations is competitive to the supervised system (strict F-score 0.693 vs. 0.755). The best system [34] in SemEval 2014 task had obtained strict F-score of 0.813 and had used different types of features and resources. We note that our system actually did better than 12 out of the 21 team systems [24] that had participated in SemEval 2014 task and had used manual annotations in their supervised methods (their strict F-scores ranged from 0.153 to 0.694). Hence even without expensive manual annotations our method obtained performance competitive to some supervised methods on this dataset.

We note that the performance of our system on the SemEval dataset was superior to that obtained on the i2b2 dataset. One likely reason for this is that we had used the same 5000 unannotated documents provided with the SemEval dataset for also building model for the i2b2 dataset because no such large unannotated corpus was available to us for the i2b2 dataset. However, the two datasets were from different sources, although both datasets included data from Beth Israel Deaconess Medical Cente, i2b2 dataset also included data from University of Pittsburgh Medical Center and Partner's Healthcare. Hence what the system learns about NER task from one dataset may not be applicable to the other dataset. We believe the performance on the i2b2 dataset would improve if our system was given a large unannotated corpus from the same source as the i2b2 dataset.

As was described in the Methods section, our method has two components – detecting named entities followed by their boundary determination. In order to evaluate the contribution of the boundary determination component, we did an ablation study in which the method did not use this component. Without that component, the output entities are simply the ones that the named entity detection classifier classified as positive out of the potential candidates. It may, for example, extract “heart disease” but miss “congenital heart disease” which the boundary determination component may be able to extract. The results are shown in Table 4 for both the datasets. It can be observed that the boundary determination component always improved results. It generally improved both precision and recall which indicate that it improved the results by correcting the incorrectly extracted entities. But it is still remarkable how well the named entity detection component alone worked.

We note a few sources of errors that our method made. Application of both the named entity detection and boundary determination components of our system rely on correctly identified noun phrases. However, since the noun phrases were obtained automatically, they were not always perfect which sometimes led to errors. Sometimes the named entities were themselves not noun phrases, either because of variation in a word that would change the type of the phrase or due to discontinuity in the mention. For example, “left ventricle is moderately dilated” mentioned in text is not a noun phrase although the name of the entity to be extracted here is “left ventricle dilation” which would be a noun phrase if directly mentioned in text. In future, our method could be improved by removing its restriction to noun phrases.

Certain atypical annotation conventions adopted in the data used for evaluation was another source of error for our system. For example, while our method would recognize “ulcerative colitis” and “shortness of breath” as named entities which seem correct, the gold-standard annotation had “moderate ulcerative colitis” and “increased shortness of breath” as the correct entities. While supervised methods that are trained on such annotations get a chance to learn these annotation conventions, it is beyond the scope of methods like ours that do not use manual annotations for training to learn these conventions. These are some reasons why our method did worse than some supervised systems, particularly on the i2b2 2010 dataset [29,35,36] that were trained on manually labeled annotations and were also equipped with knowledge resources. We want to point out that our method still has the advantage of not requiring any manual effort. Inconsistency in annotations used for evaluation was another source of errors. For example, if the text mentioned “chest pain” sometimes only “pain” would be tagged as the disease/disorder in the gold-standard annotation but at other times the anatomical part would also be included in the gold-standard annotation.

## 5. Conclusions

We presented a novel method for doing clinical NER that does not require any manual annotations. It only requires a raw corpus and a resource like UMLS that can give a list of named entities along with

their semantic types. Using these two resources our method automatically obtains annotations for training machine learning methods for named entity detection and for their boundary determination. Our new method performed better than an existing unsupervised clinical NER method. It was also competitive to supervised methods on one dataset. The presented method provides an alternate to building NER systems for clinical domain without needing expensive manual annotations.

## Ethical statement

Not applicable.

## Conflicts of interest

Authors declare that they have no conflict of interest.

## Acknowledgement

We thank the organizers of SemEval 2014 Task 7 and i2b2 2010 shared-task for creating and providing the data which was used in this work.

## References

[1] Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? J Biomed Inf 2009;42(5):760–72.
[2] Savova G, Pestian J, Connolly B, Miller T, Ni Y, Dexheimer JW. Natural language processing: applications in pediatric research. Pediatric biomedical informatics. Singapore: Springer; 2016. p. 231–50.
[3] Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. IMIA Yearb Med Inf 2008;35(8):128–44.
[4] Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, Liu S, Zeng Y, Mehrabi S, Sohn S, Liu H. Clinical information extraction applications: a literature review. J Biomed Inf 2018;77:34–49.
[5] Liu F, Chen J, Jagannatha A, Yu H. Learning for biomedical information extraction: methodological review of recent advances. 2016 Jun 26. arXiv preprint arXiv:1606.07993.
[6] Chen Y, Lasko TA, Mei Q, Denny JC, Xu H. A study of active learning methods for named entity recognition in clinical text. J Biomed Inf 2015;58:11–8.
[7] Alfonseca E, Manandhar S. An unsupervised method for general named entity recognition and automated concept discovery. Proceedings of the first international conference on general WordNet, Mysore, India. 2002. p. 34–43.
[8] Nadeau D, Turney PD, Matwin S. Unsupervised named-entity recognition: generating gazetteers and resolving ambiguity. Conference of the Canadian society for computational studies of intelligence. 2006. p. 266–77.
[9] Riloff E, Jones R. Learning dictionaries for information extraction by multi-level bootstrapping. Proceedings of the sixteenth national conference on artificial intelligence and the eleventh innovative applications of artificial intelligence conference (AAAI/IAAI). 1999 Jul 18. p. 474–9.
[10] Cucchiarelli A, Velardi P. Unsupervised named entity recognition using syntactic and semantic contextual evidence. Comput Ling 2001;27(1):123–31.
[11] Etzioni O, Cafarella M, Downey D, Popescu AM, Shaked T, Soderland S, Weld DS, Yates A. Unsupervised named-entity extraction from the web: an experimental study. Artif Intell 2005 Jun 1;165(1):91–134.
[12] Elsner M, Charniak E, Johnson M. Structured generative models for unsupervised named-entity clustering. Proceedings of the north american chapter of the association for computational linguistics - human language technologies (NAACL HLT) conference. 2009. p. 164–72.
[13] Zhang S, Elhadad N. Unsupervised biomedical named entity recognition: experiments with clinical and biological texts. J Biomed Inf 2013;46(6):1088–98.
[14] Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004;32(suppl_1):D267–70.
[15] Xu J, Gan L, Cheng M, Wu Q. Unsupervised medical entity recognition and linking in Chinese online medical text. J Healthc Eng 2018:2548537. 13 pages.
[16] Oronoz M, Casillas A, Gojenola K, Perez A. Automatic annotation of medical records in Sspanish with disease, drug and substance names. Ruiz-Shulcloper J, Sanniti di Baja G, editors. Progress in pattern recognition, image analysis, computer vision, and applications. CIARP, vol. 8259. Berlin, Heidelberg: Springer; 2013. Lecture notes in computer science.
[17] Padró L, Reese S, Agirre E, Soroa A. Semantic services in freeling 2.1: wordnet and UKB. Proceedings of 5th global WordNet conference. 2010. p. 99–105.
[18] Usami Y, Cho HC, Okazaki N, Tsujii JI. Automatic acquisition of huge training data for bio-medical named entity recognition. Proceedings of BioNLP Workshop. 2011. p. 65–73.
[19] Vlachos A, Gasperin C. Bootstrapping and evaluating named entity recognition in the biomedical domain. Proceedings of the HLT-NAACL BioNLP workshop on linking natural language and biology. 2006. p. 138–45.
[20] Peng Y, Wei CH, Lu Z. Improving chemical disease relation extraction with rich features and weakly labeled data. J Cheminf 2016;8:53.
[21] Alicante A, Corazza A, Isgrò F, Silvestri S. Unsupervised entity and relation extraction from clinical records in Italian. Comput Biol Med 2016;72:263–75.
[22] Kholghi M, Sitbon L, Zuccon G, Nguyen A. Active learning reduces annotation time for clinical concept extraction. Int J Med Inf 2017;106:25–31.
[23] Skeppstedt M. Annotating named entities in clinical text by combining pre-annotation and active learning. Proceedings of the 51st annual meeting of the association for computational linguistics student research workshop. 2013. p. 74–80.
[24] Pradhan S, Elhadad N, Chapman W, Manandhar S, Savova G. Semeval-2014 Task 7: analysis of clinical text. Proceedings of the 8th international workshop on semantic evaluation (SemEval). 2014. p. 54–62.
[25] Lafferty J, McCallum A, Pereira FC. Conditional random fields: probabilistic models for segmenting and labeling sequence data. Proceedings of the eighteenth international conference on machine learning (ICML). 2001. p. 282–9.
[26] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. ACM SIGKDD Explor Newsl 2009;11(1):10–8.
[27] Ho TK. The random subspace method for constructing decision forests. IEEE Trans Pattern Anal Mach Intell 1998;20(8):832–44.
[28] Manning C, Surdeanu M, Bauer J, Finkel J, Bethard S, McClosky D. The Stanford CoreNLP natural language processing toolkit. Proceedings of 52nd annual meeting of the association for computational linguistics (ACL): system demonstrations. 2014. p. 55–60.
[29] Uzuner Ö, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inf Assoc 2011;18(5):552–6.
[30] Japkowicz N, Shah M. Evaluating learning algorithms: a classification perspective. Cambridge University Press; 2011.
[31] Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. J Am Med Inf Assoc 2010;17(3):229–36.
[32] Ghiasvand O, Kate R. UWM: disorder mention extraction from clinical text using CRFs and normalization using learned edit distance patterns. Proceedings of the 8th international workshop on semantic evaluation (SemEval). 2014. p. 828–32.
[33] Ghiasvand O. Disease name extraction from clinical text using conditional random fields Master's Thesis Milwaukee, USA: University of Wisconsin-Milwaukee; May 2014
[34] Zhang Y, Wang J, Tang B, Wu Y, Jiang M, Chen Y, Xu H. UTH_CCB: a report for SemEval 2014–Task 7 analysis of clinical text. Proceedings of the 8th international workshop on semantic evaluation (SemEval). 2014. p. 802–6.
[35] 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. J Am Med Inf Assoc 2011;18(5):557–62.
[36] Jiang M, Chen Y, Liu M, Rosenbloom ST, Mani S, Denny JC, Xu H. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. J Am Med Inf Assoc 2011;18(5):601–6.