Medical Named Entity Recognition for Indonesian Language Using Word Representations

Arief Rahman

1School of Electrical Engineering and Informatics (SEEI), Institut Teknologi Bandung, Indonesia

Email: 123516008@std.stei.itb.ac.id

Abstract. Nowadays, Named Entity Recognition (NER) system is used in medical texts to obtain important medical information, like diseases, symptoms, and drugs. While most NER systems are applied to formal medical texts, informal ones like those from social media (also called semi-formal texts) are starting to get recognition as a gold mine for medical information. We propose a theoretical Named Entity Recognition (NER) model for semi-formal medical texts in our medical knowledge management system by comparing two kinds of word representations: cluster-based word representation and distributed representation.

1. Introduction

The study of using NLP in the biomedical domain has been in the spotlight for the last few years. Its popularity soars mainly because of the abundance of information and the computing power available for processing information. Named Entity Recognition (NER) is an important NLP task that allows us to get important medical information from medical texts easily, which can be useful to get new knowledge (e.g. in indigenous systems [1]).

NER is a task of finding named entities (NE), which can be words and phrases, which belong to the same semantic classes. In the biomedical domain, the widely accepted named entities are diseases, signs/symptoms, anatomical signs, and drugs. Several NER tools have been developed for getting medical named entities, such as MedLEE [2], MetaMap [3] and cTAKES [4]. These tools use a rule-based system, which relies on comprehensive medical knowledge and vocabularies.

Currently, most state-of-the-art medical NER systems use machine learning models. NER in particular uses supervised machine learning, in which labeled data are used to train models and then used to recognize new, unlabeled data. In machine learning, NER is approached as a sequence labeling problem that aims to find the best label sequence (e.g. IO, BIO, and BMESWO format label) in a given sentence (sequence of words). Three of the most-used machine learning algorithms in medical NER are Conditional Random Field (CRF) [5], Maximum Entropy (ME), and Support Vector Machines (SVM) [6]. One big downside, however, is the fact that the supervised machine learning algorithms only works well when the features are manually extracted through feature engineering.

Typical NER systems mostly used orthographic features (e.g., capitalization of letters, prefix, and suffix), syntactic features (e.g. POS tags), and n-gram features. However, researchers found that the performance of supervised machine learning modeling can be improved by using “unsupervised” features, which usually replaces the word form feature as an alternative word representation. These alternative word representations are learned by using unlabeled corpus. Some of the most famous unsupervised features are Brown cluster [7] and word embedding [8][9]. Brown cluster is one of the word
representations based on clustering. Word embedding, on the other hand, is a type of distributed word representation.

Most medical NER systems focused on formal medical texts, which contain formal linguistic styles. This makes it easier to find named entities because it is guaranteed to follow certain grammar. However, medical texts nowadays can also come from social media. They are usually found in informal medical chats or groups. These texts most of the time have both formal and informal grammar, which can be a big challenge for most NER systems. cTAKES [4] tried to process medical free-texts, with a satisfying F1-score (81.24%). However, the NER model in this system used a terminology-agnostic dictionary look-up algorithm.

In this research, we propose a theoretical model to our medical NER system by using two kinds of unsupervised features, which are cluster-based word representation and distributed word representation. In addition to its easy-to-use property, both of these features are also reliable to use on texts with informal language styles. The usage of word representation is expected to help in making a NER model that's easy to train, yet good at classifying named entities. In the following sections, we summarize the basic concepts of the Brown cluster and word embedding. We also explain our proposed theoretical model and the proposed experiments to be done. This study is conducted as a part of the Knowledge Management System (KMS) study conducted in [1] and [10].

2. Word Representations
In this section, we discuss two types of word representations: cluster-based word representation and distributed word representation. We do not discuss the distributional word representations because of its lack of development in the recent years.

2.1. Cluster-based Word Representation
Cluster-based word representation is a word representation that is obtained by inducing a clustering over words. There are two types of cluster-based word representation: hard clustering (Brown cluster) and soft clustering (HMM word representation).

Brown et al. [7] proposed a greedy agglomerative hierarchical clustering procedure that groups words to maximize the mutual information of bigrams. In this procedure, clusters are initialized with a single word in each cluster. After that, the clusters are merged greedily according to the bi-gram mutual information criterion. The resulting clusters are merged until only one big cluster remains.

The result of Brown cluster procedure is a low-dimensional representation of the vocabulary in the form of a hierarchical, binary tree. The low-dimensional representation mitigates the feature sparseness problem. The procedure runs in time complexity of $O(VK^2)$, with $V$ being the size of the vocabulary and $K$ being the number of clusters. Figure 1 shows an example of the Brown clustering result.

![Figure 1](image-url) Proposed named-entity recognition system process [11].

Several studies have proved that this unsupervised clustering procedure has helped in increasing NER performances [12]. The hierarchical nature of the Brown clusters allows words to be represented...
at different levels of the hierarchy. The Brown cluster has one major downside: it does not consider wider contexts and word usage in general, using only bigram statistics.

2.2. Distributed Word Representation
Distributed word representation is dense, low-dimensional, and real-valued vector that represents its syntactic (and potentially semantic) properties. This representation is also called word embedding. Each of its dimension represents a latent feature of the word, which captures syntactic and semantic properties for the computer to understand. A distributed representation is compact and can represent an exponential number of clusters in a limited number of dimensions.

Word embeddings are usually made using neural language model, which uses neural networks as the underlying predictive model [13]. It is different from clustering-based word representations, in which they are usually induced under n-gram language models. Historically, training and testing of word embedding models have been slow, scaling to the size of the vocabulary [14]. Many studies to allow word embedding scaling to very large training corpora were done in recent years. The biggest breakthrough comes from Mikolov [8][9], which induces word representations using a neural network that has a log-linear complexity. Mikolov’s design allows researchers to induce distributed word representation with much more efficiency without significant accuracy defect.

Mikolov proposed two architectures for efficient neural network language model (NNLM). The first one is called the Continuous Bag-of-Words model (CBOW) [8]. It uses a feedforward NNLM, with no non-linear hidden layer and shared projection layer for all words. The architecture uses both words from the history and the future (previous \( n \) tokens and next \( n \) tokens). The bag-of-words name comes from the fact that the position of the words in the window does not matter in the long run to determine the embedding.

The second architecture is called the skip gram model [8]. The skip gram architecture uses a different approach from CBOW, by maximizing the classification probability of a word based on another word in the same sentence. It does this by using each current word as an input to a log-linear classifier with continuous projection layer and predict words within a certain range before and after the current word. Increasing the context range improves both the quality of the resulting word vectors and the computational complexity. A weight decay penalty is given to distant words by doing less sampling in the training examples because they are usually less related to the current word.

Figure 2 shows both the CBOW and skip gram neural network architecture.

![Figure 2. CBOW and skip gram architecture [8.](image_url)
3. Proposed Named Entity Recognition Model
We propose the following NER model. First, we will crawl one or more medical-oriented social media
groups to obtain medical semi-formal texts. After that, we use a language formalization tool to reduce
the number of features and allow easier processing in the next steps. In the next step, we use will the
final dataset for two purposes: train a word representation and create a NER dataset. For the first purpose,
we use an unsupervised word representation training algorithm to train the word representation. For the
second purpose, currently, we plan to annotate the dataset manually with named medical entities.
Finally, we use both the NER dataset and the trained word representations to train a model using the
artificial neural network algorithm. The whole process can be seen in Figure 3.

In the next subsections, we explain the important parts in our NER system, in which we will discuss
the dataset, the word representations, and the model learning.

3.1. Data Sets
For this work, we will use the medical semi-formal texts fetched from several Facebook's Indonesian
medical discussion groups. These groups use semi-formal language style medical texts, with texts
ranging from medical information to medical questions. We will extract these texts using NodeXL. We
plan to use around one year worth of medical semi-formal texts to use.

To allow easier processing, we will formalize the texts obtained from the media social using our own
rule-based Indonesian formalization tool. We will also delete all texts that don't contain any medical
named entities. To detect these named entities, we will use a gazetteer that contains the most common
medical terms in medical texts in general. We will manually tag the named entities on the whole texts
after the whole processing steps are done.

3.2. Word Representation Features
In this work, we will use several word representations as features for our NER model. We will use the
Brown Cluster [7] and Mikolov’s word embedding [8]. To train these word representations, we use both
the Indonesian Wikipedia dump1 and a corpus containing formalized raw texts from several Facebook’s
Indonesian medical discussion groups. We plan to use around one year worth of medical semi-formal
texts to use. We use these corpora to train both the formal and informal text styles.

For the word representation generator, we will use Percy Liang’s Brown cluster generator2, which
is implemented in Python and is easy to use and parameterize. We will also use Mikolov’s word2vec3
to generate the word embedding and use both the CBOW and skip-gram models.

---
1 https://dumps.wikimedia.org/
2 https://github.com/percyliang/brown-cluster
3 https://code.google.com/archive/p/word2vec/
3.3. Neural Network Model and Word Representation Features

Our NER models will use the neural network, which has been used recently, with satisfying performance. Our network consists of an input layer, two hidden layers with ReLU activation function (with 100 nodes per layer), and a softmax output layer, which represents the probability of each named entity (including its BIO label) being the correct label. Figure 4 shows our proposed neural network architecture.

![Proposed neural network architecture for the NER model.](image)

We will train the network using the cross-entropy loss minimization as the training objective, with an addition of L2-regularization term. The neural network parameters correspond to the weight matrix $W$ and bias $b$.

The input layer consists of baseline features, Brown cluster encoding (in the form of n binary input nodes), and/or word representation of the current word to be tagged. We use Zhang and Johnson’s baseline features [15], which has been proven to be effective for NER tasks:

1. Previous two named entity predictions $y_{i-1}$ and $y_{i-2}$.
2. Current word $x_i$, with several word type information: all-capitalized, is-capitalized, all-digits, and alphanumeric.
3. Prefixes and suffixes of $x_i$. The tokens between hyphens (‘-‘) are used if at least one hyphen character is present in $x_i$.
4. Context $c = (x_{i-2}, x_{i-1}, x_i, x_{i+1}, x_{i+2})$.
5. Capitalization pattern in the context $c$.
6. Conjunction of $c$ and $y_{i-1}$

We normalize dates and numbers when using lexical features (e.g., 2017 becomes *DDDD*, and 0752-618201 becomes *DDDD*-*DDDDDDDD*). This process is performed separately from the word representation usage. For example, if we have induced an embedding for 10/18/2017, we will use the embedding of 10/18/2017 and *DD*/*DD*/*DDDD* in the baseline features listed above.
4. Test Scenarios
We will test our work by comparing both the word representations themselves and the generated models. We will test the word representations performances using perplexity, which is a cross-entropy based measure borrowed from language modeling. In particular, we will test the performance of Brown cluster and several distributed word representations, which are shown in Table 1. Because of the word embedding’s flexibility in its training parameters, we will also compare several word embeddings trained with different model types, window sizes, and embedding sizes.

Table 1. Test scenarios for word representations.

| Word Representation | Potential Parameter Changes |
|---------------------|-----------------------------|
| 1 Brown cluster     | -                           |
| 2 CBOW-word embedding| Embedding size (50, 100, 200, and 300)  |
| 3 Skip-Gram embedding| Embedding size (50, 100, 200, and 300)  |

We will use the Brown cluster and the best CBOW and Skip-gram word embeddings from the previous experiment for the model training. We will compare several models which use different features. The baseline features are based on Zhang and Johnson’s set of features [15]. Table 2 shows the feature combination used on each model. We will use the F1 measure as the main metric for our model performances. In order to mitigate the overfitting effect from potential lack of data for our model training, we will use the $k$-fold cross validation testing method, with $k$ of 5.

Table 2. Test scenarios for NER model.

| Feature Combination |
|---------------------|
| 1 Baseline features |
| 2 Brown cluster     |
| 3 CBOW-word embedding|
| 4 Skip-Gram embedding|
| 5 Brown cluster + CBOW word embedding |
| 6 Brown cluster + Skip-gram word embedding |
| 7 Baseline + Brown cluster + CBOW word embedding |
| 8 Baseline + Brown cluster + Skip-gram word embedding |

5. Conclusions
In this study, we proposed our theoretical NER models for identifying medical named entities in medical semi-formal texts. The word representations cited here can hopefully be used to enhance Indonesian medical named entity recognizer in semi-formal texts. We also hope that this research can be used to spur interests in natural language processing research in the larger scale.

Acknowledgments
This study is financially supported by the Institut Teknologi Bandung (ITB) under Program Penelitian, Pengabdian kepada Masyarakat, dan Inovasi (P3MI). We would like to acknowledge Kridanto Surendro (Information System Laboratory – SEEI ITB) for supporting the writing of this paper.

References
[1] Atrinawati L H and Surendro K 2009 Assessment for knowledge management readiness *Electrical Eng. and Informatics 2009 (ICEEI’09) Int. Conf.* on 2 pp 399–404.
[2] Friedman C 1997 Towards a comprehensive medical language processing system: methods and issues Proc. of the AMIA Annual Fall Symp. p 595.

[3] Aronson A R and Lang F M 2010 An overview of MetaMap: historical perspective and recent advances J. Am. Med. Informatics Assoc. 17(3) pp 229–36.

[4] Savova G K et al. 2010 Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications J. Am. Med. Informatics Assoc. 17(5) pp 507–13.

[5] Lafferty J, McCallum A, and Pereira F C N 2001 Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

[6] Tsochantaridis I, Joachims T, Hofmann T, and Altun Y 2005 Large margin methods for structured and interdependent output variables J. Mach. Learn. Res. 6 pp 1453–84.

[7] Brown P F, Desouza P V, Mercer R L, Pietra V J D, and Lai J C 1992 Class-based n-gram models of natural language Comput. Linguist. 18(4) pp 467–79.

[8] Mikolov T, Chen K, Corrado G, and Dean J 2013 Efficient estimation of word representations in vector space arXiv (Preprint arXiv1301.3781).

[9] Mikolov T, Sutskever I, Chen K, Corrado G, and Dean J 2013 Distributed Representations of Words and Phrases and their Compositionality.

[10] Aradea S I and Surendro K 2014 An overview of multi agent system approach in knowledge management model Int. Conf. on Information Technol. Syst. and Innovation (STEI ITB).

[11] Tang B, Cao H, Wang X, Chen Q, and Xu H 2014 Evaluating word representation features in biomedical named entity recognition tasks Biomed Res. Int. 2014.

[12] Turian J, Ratinov L, Bengio Y, and Roth D 2009 A preliminary evaluation of word representations for named-entity recognition NIPS Workshop on Grammar Induction, Representation of Language and Language Learning pp 1–8.

[13] Bengio Y 2008 Neural net language models Scholarpedia.

[14] Bengio Y, Ducharme R, Vincent P, and Jauvin C 2003 A neural probabilistic language model JMLR J. Mach. Learn. pp. 1137–55.

[15] Zhang T and Johnson D 2003 A robust risk minimization based named entity recognition system Proc. seventh Conf. Nat. Lang. Learn. HLT-NAACL 2003 4 pp 204–7.