

# Improving NER for Clinical Texts by Ensemble Approach using Segment Representations

**Hamada A. Nayel**

Department of Computer Science  
Benha University, Benha-Egypt  
Mangalore University, Mangalore-India  
hamada.ali@fci.bu.edu.eg

**H. L. Shashirekha**

Department of Computer Science  
Mangalore University,  
Mangalore-574199, India  
hlsrekha@gmail.com

## Abstract

Clinical Named Entity Recognition (Clinical-NER), which aims at identifying and classifying clinical named entities into predefined categories, is a critical pre-processing task in health information systems. Different machine learning approaches have been used to extract and classify clinical named entities. Each approach has its own strength as well as weakness when considered individually. Ensemble technique uses the strength of one approach to overcome the weakness of another approach by combining the outputs of different classifiers in order to make the decision thereby improving the results. Segment representation is a technique that is used to add a tag for each token in a given text. In this paper, we propose an ensemble approach to combine the outputs of four different base classifiers in two different ways, namely, majority voting and stacking. We have used support vector machines to train the base classifiers with different segment representation models namely IOB2, IOE2, IOBE and IOBES. The proposed algorithm is evaluated on a well-known clinical dataset i2b2 2010 corpus and results obtained illustrate that the proposed approach outperforms the performance of each of the base classifiers.

## 1 Introduction

Named Entity Recognition (NER) is a leading sub-task of information extraction originated from the Sixth Understanding Conference (MUC-6) (Grishman and Sundheim, 1996), which aims at identifying Named Entities (NEs) in a text and clas-

sifying them into predefined classes. Names of organizations, locations and persons are examples of NEs in general newswire domain, while DNA, RNA and protein are examples of NEs in biological domain. In clinical domain, terms representing problem, treatment and laboratory test are considered as NEs.

The exponential growth of health information systems produce a massive amount of Electronic Health Records (EHRs). It is vital to apply NER for health information systems because EHRs contain NEs representing laboratory test, problem and treatment in unstructured narrative documents (Friedman et al., 1994). Moreover, NER in clinical domain (Clinical-NER) is an important pre-processing task in health information systems where further tasks of health information systems depend essentially on the results of Clinical-NER. Clinical-NER is a challenging problem because, in addition to the general challenges of NER there are other challenges resulting from the nature of clinical NEs such as: -

1. Ambiguity:- the major sources of ambiguity are abbreviations and acronyms (Pakhomov et al., 2005), which are used routinely in clinical texts. Two different cases cause the ambiguity, (i) same abbreviation used for different entities such as "*EF (Ejection Fraction)*" which is used as a medical problem as well as a laboratory test, and (ii) an abbreviation conflicts with a word such as "*VS*" which is a laboratory test as well as abbreviation for the word "*versus*".
2. Multiple words entities:- most of clinical entities consist of multiple words such as "*lower abdominal pain*" and "*chest x-ray*".
3. Nested clinical entities:- some clinical entities occur as a part of longer entity such as

”BP (blood pressure)”, a laboratory test occurs in ”control BP” which is a treatment.

4. Polysemy:- same clinical term can represent different meanings based on the context, such as ”inflammation” may refer to skin problem, a cellular level problem as well as non-medical activity.
5. Synonymy:- a single medical concept can be expressed as multiple words (Dehghan et al., 2013) such as ”baby” and ”foetus” which means the same in many medical contexts.

In addition to these challenges, there is no standard nomenclature for clinical entities of same class.

### 1.1 NER approaches

The commonly used approaches for NER are dictionary based approach, rule based approach, Machine Learning (ML) approach and hybrid approach (Keretna et al., 2015). In dictionary based approach, a dictionary or lexicon, which contains a finite set of named entities is used to look up for the entities in texts. Rule based approach uses well-designed domain specific hand crafted rules by experts to match the entities. In ML approach, ML algorithms such as Support Vector Machines (SVMs), Conditional Random Fields (CRFs) and Maximum Entropy (ME) are used to create a learning model using training set to detect the boundaries of entities and classify them into one of the predefined classes. Hybrid approach combines two or more approaches to identify NERs. ML approach either solo or hybrid with another approach is preferable to use as they can easily adopt to new domains as well as identify unseen entities. The major requirement of an ML approach is an annotated data set tagged by experts (*training data*) to train the learning model.

### 1.2 General Framework of NER using ML approach

Figure 1 shows the general framework of NER using ML approach. In this model, a training data set is used to train the classifier and a set of untagged data (*testing data*) is used to evaluate the performance of the classifier. In tokenization phase, data sets are tokenized into set of tokens or words. In feature extraction phase, a set of features are extracted. Feature extraction is a very important phase as the performance of the model depends essentially on features. Then the features of the

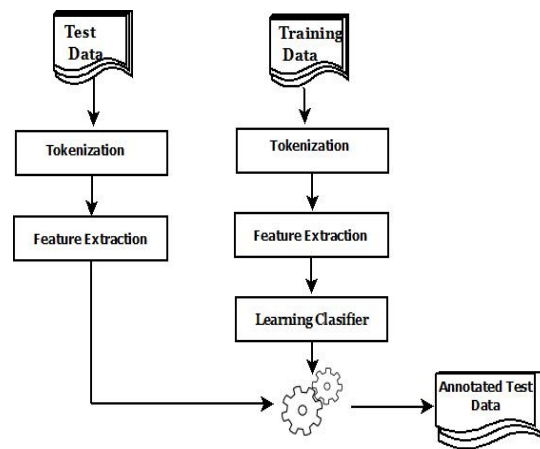


Figure 1: Machine learning framework for NER

training data are used to learn the model and that of testing data are used for evaluation. The success of ML approach depends on the quality of annotated training data, quality of the features extracted as well as the algorithm used for creating the classification model. Each classification algorithm has its strength as well as weakness when used individually. Some classifiers give good results on some datasets whereas the same classifier perform very bad on some other datasets. So, instead of considering a single classifier, it will be beneficially to pool the classifiers and then take the collective decision similar to the decision taken by a committee rather than an individual. This technique which overcomes the weakness of some classifiers using the strength of other classifiers is termed as ”ensemble” and is gaining importance for various applications. Ensemble classification uses a set of classifiers preferably weak, diverse and heterogenous classifiers as base classifiers and combines the output of these base classifiers in different ways to get the final output. To achieve the diversity of base classifiers, researchers are using different feature sets, different training sets and/or different classification algorithms. There are different approaches to create ensemble classifiers such as bagging, boosting and stacking (Polikar, 2006). In bagging, different training subsets are drawn with replacement from the entire training data and each training data subset is used to train each base classifier. The outputs of base classifiers are combined using majority voting. Boosting is similar to bagging, but the selection process of training subsets subsequently gives more weight to misclassified samples. Stacking uses outputs

	<b>IO</b>	<b>IOB1</b>	<b>IOB2</b>	<b>IOE1</b>	<b>IOE2</b>	<b>IOBE</b>	<b>IOBES</b>
Treatment	O	O	O	O	O	O	O
/	O	O	O	O	O	O	O
stay	O	O	O	O	O	O	O
IHSS	I-problem	B-problem	B-problem	E-problem	E-problem	B-problem	S-problem
AF	I-problem	B-problem	B-problem	E-problem	E-problem	B-problem	S-problem
ESRD	I-problem	B-problem	B-problem	E-problem	E-problem	B-problem	S-problem
on	O	O	O	O	O	O	O
HD	I-treatment	I-treatment	B-treatment	I-treatment	E-treatment	B-treatment	S-treatment
,	O	O	O	O	O	O	O
IgA	I-problem	I-problem	B-problem	I-problem	I-problem	B-problem	B-problem
nephropathy	I-problem	I-problem	I-problem	I-problem	E-problem	E-problem	E-problem
on	O	O	O	O	O	O	O

Table 1: An example of using different Segment Representation models

of base classifiers to train a new model, which is known as meta-classifier (Wolpert, 1992) and the meta-classifier is used for final classification.

### 1.3 Segment Representation

Segment Representation (SR) (Cho et al., 2013) involves the process of assigning suitable class label to the words in a given text. SR models have been applied for different tasks such as Part of Speech (PoS) tagging and Noun Phrase chunking (NP-chunking) (Wu, 2014). SR model comprises set of tags, which determine the position of a token in NE, combined with the class label that NE belongs to. The tags used in SR techniques are Begin (**B**), End (**E**), Inside (**I**), Single (**S**) and Outside (**O**). For example, SR for a token is B-XXX means that word is the first word of a NE of class XXX. SR can represent multiple word NEs and nested NEs. Different models are used for segment representation by different researchers. The primary SR model **IO** (Béchet et al., 2000) assigns the tag **I** for the tokens inside the entity and the tag **O** for the tokens outside the entity, but is not able to represent the boundaries of two consecutive entities of the same class. **IOB1** model has been introduced to solve this problem (Ramshaw and Marcus, 1995), by assigning the tag **B** to the first token of consecutive NEs of same class, while **IOB2** model assigns the tag **B** for the first word of each NE (Ratnaparkhi, 1998). **IOE1** and **IOE2** models use same concepts of **IOB1** and **IOB2** respectively, but assigns the tag **E** to the last token of NEs (Kudo and Matsumoto, 2001). Sun et al. (2010) introduced **IOBE** model which concerns with the beginning and end of the NE. **IOBE** model assigns the tags **B** and **E** for the first and last word of all NEs re-

spectively. **IOBES** model is a modified version of **IOBE** model that is concerned with single word NEs. In addition to **IOBE** tags, the **IOBES** model assigns the tag **S** to the NEs of a single word. This model differentiates between the single word and multiple words NEs. Example of tagging the text fragment "Treatment / stay IHSS AF ESRD on HD , IgA nephropathy on .." with different SR models is shown in Table 1.

## 2 Related Work

The research works carried out in ensemble approach uses different training data sets or different learning algorithms to create the base classifiers. Different ML algorithms such as SVM and CRF have been used for Clinical-NER (Li et al., 2008). Keretna et al. (2014), have introduced a hybrid approach using rule-based and dictionary-based approaches to identify drug names in unstructured and informal texts. The system was evaluated on i2b2 2009 medication challenge dataset and reported 66.97% f-score. Dictionaries and rule-based approaches have been extensively used to extract clinical entities in clinical information systems such as MedLEE developed by Friedman et al. (1994), MetaMap developed by Arnsion and Lang (2010) and cTAKES developed by Savova et al. (2010). Gurulingappa et al. (2010) trained CRFs on textual features enhanced with the output of a rule-based NER system. They evaluated their work using i2b2/VA 2010 medical challenge dataset and reported 81.2% f-measure. Halgrim et al., (2010) designed a hybrid approach that comprised of CRF and Rule-based approach for Clinical-NER. Zhang and Elhadad (2013) de-

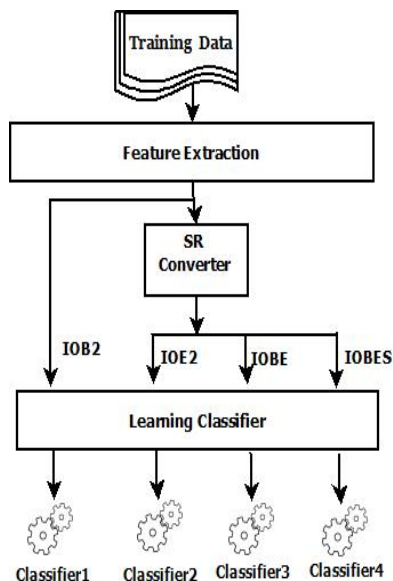


Figure 2: Learning base classifiers

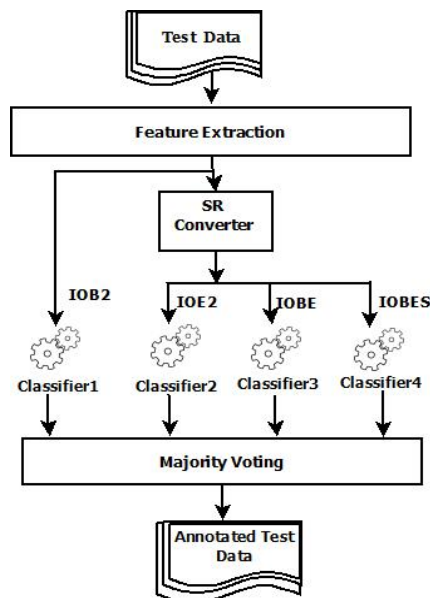


Figure 3: Combining base classifiers using Majority Voting

veloped an unsupervised approach for extracting clinical entities from free text. They used inverse document frequency as a base to filter candidate clinical NEs. Ekbal and Saha (2013) used stacked ensemble approach to extract biomedical NEs. Shashirekha and Nayel (2016) studied the performance of biomedical NER using different SR models. Keretna et al. (2015) introduced a technique for boosting clinical-NER by extending IOBES model, and have introduced a new tag to resolve the problem of ambiguity. They evaluated the proposed technique on i2b2/VA 2010 medical challenge dataset. There is a growing interest in studying Clinical-NER for non-English texts (Wu et al., 2015; Spat et al., 2008). Wu et al. (2015) trained a deep neural network model to extract clinical entities from Chinese texts .

In this paper, we have proposed an ensemble algorithm for Clinical-NER. Up to our knowledge, this is the first work that uses SR models to achieve diversity of base classifiers. Our approach is a two-stage ensemble algorithm. In the first stage, we have used SVM algorithm to create four base classifiers with different SR models namely IOB2, IOE2, IOBE and IOBES. Stacking using CRF as a meta-classifier and Majority Voting have been used separately to combine the results of base classifiers in the second stage.

### 3 Methodology

We propose a two-stage ensemble approach for clinical-NER. Figure 2 shows the framework of first phase, where training data is used to learn the base classifiers. We have used SVM algorithm to learn four different base classifiers using different SRs models namely, IOB2, IOE2, IOBE and IOBES. In second phase, we have combined the outputs of the base classifiers created in the first phase using Majority Voting and Stacking separately which form the result of ensemble technique. Figures 3 and 4 show the framework of second phase using Majority Voting and Stacking respectively. We designed a SR converter module to convert the dataset which is available in IOB2 model into other SR module.

#### 3.1 Feature extraction

Features, the properties of tokens or words, are the keystones of ML algorithms. The following features were extracted for our system:-

1. Word length:- This is a numeric value that determines the length of the current token.
2. Context words:- These are the words surrounding the current word. The context window of size  $n$  means  $n$  words before the current word and  $n$  words after the current word, e.g. context window of size 3

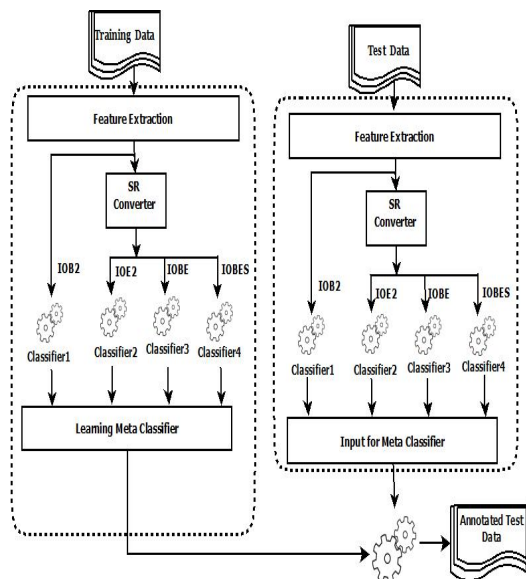


Figure 4: Combining base classifiers using Stacking

is  $w_{i-3}...w_i...w_{i+3}$  where  $w_i$  is the current word.

3. **Word affixes:-** These are prefixes and suffixes of the current word. Prefix of length  $n$  is the first  $n$  characters of the word, while suffix of length  $n$  is the last  $n$  characters of the word. We have used all suffixes and prefixes up to length 5.
4. **Part-of-Speech (PoS) tags:-** PoS information is a very important feature, it determines the role of the word in the sentence. PoS tags are extracted using GENIA tagger V3.0.21<sup>1</sup>.
5. **Chunk Information:-** Chunk information is useful when determining the boundaries of NEs. chunk information is extracted using GENIA tagger V3.0.21.
6. **Word Normalization:-** Two types of normalization namely stemming feature and word shape feature are used. Word stem means the root of a word. GENIA tagger V3.0.21 is used to extract the stems. There are two types of word shapes, general word shape and summarized word shape. In a general word  $X$  is substituted for each capital letter,  $x$  for each small character and  $d$  for consecutive digits. In a summarized word shape, consecutive

capital letters are replaced by  $X$ , consecutive small letters by  $x$  and consecutive digits by  $d$ .

7. **Orthographic features:-** These features capture word formation information. The set of all orthographic features extracted are shown in Table 2.
8. **Dynamic Feature:-** It denotes the predicted tags of the words preceding the current word. This feature is calculated during running. An example of dynamic feature of size 4 are the tags  $t_{-4}, t_{-3}, t_{-2}, t_{-1}$  corresponding to the words  $w_{-4}, w_{-3}, w_{-2}, w_{-1}$ , where current word is  $w_0$ .
9. **Stop Words:-** This is a logical feature which fires only if the current word is a stop word.
10. **Non-Word:-** This is a binary value which fires only if the word exists in entire dictionary. We used Grady augmented dictionary in `qdapDictionaries` package in R software (R Core Team, 2017).
11. **Head Nouns:-** The noun phrase describes the functionality or property of a clinical NE called head noun (Ekbal and Saha, 2013). For example, examination is the head noun of "cardiovascular examination". Head nouns are very important as these play a key role for correct classification of a clinical NE class. Unigrams and bigrams are used as head nouns. For domain dependency, training data is used to extract head nouns.

### 3.2 Support Vector machines

SVM is a binary classifier, which creates a hyper-plane that discriminates between the two classes. SVM can be extended to multi-classes problems by combing several binary SVMs and combining using a one-vs-rest method or one-vs-one method (Hsu and Lin, 2002).

### 3.3 Evaluation Metrics

The performance of our system is reported in terms of f-measure (Hripcsak and Rothschild, 2005). F-measure is a harmonic mean of Precision (**P**) and Recall (**R**). Denoting **TP** as the number of true positives, **FP** number of false positives and **FN** as the number of false negatives, recall, precision and f-measure are calculated as follow:

$$P = \frac{TP}{TP + FP}$$

<sup>1</sup><http://www.nactem.ac.uk/GENIA/tagger/>

Feature	Example
INITCAPS	Tonsillectomy
ALLCAPS	MCV, RBC
ENDCAPS	pH, proBNP
INCAPS	freeCa
CAPSMIX	cTropnT
HASDIGIT	pO2,calHCO3
HASHYPHEN	hyper-CVAD
ALPHNUM	B12
GREEK	alpha
NUMBER	101.5
HASATGC	LACTATE
PUNCT	INR(PT)
ROMAN	IV, CD

Table 2: List of orthographic features and examples

$$R = \frac{TP}{TP + FN}$$

$$f - measure = \frac{2 * P * R}{P + R}$$

### 3.4 Dataset

Our model is eval on i2b2 dataset (Uzuner et al., 2011), which was originally created for entity and relation extraction purposes at i2b2/VA 2010 challenge. It includes 826 discharge summaries for real patients from the University of Pittsburgh Medical Centre, Partners Health Care and Beth Israel Deaconess Medical Centre. Pittsburgh discharge summaries was used as a test set in i2b2 challenge and other two sources used as a training set. Statistics of the dataset is shown in Table 3. Both testing and training sets are manually annotated with three different named entities namely, treatment, problem and test. It is important to note that, there is lack of data sets that used for Clinical-NER.

## 4 Experiments and results

The proposed method combines the outputs of base-classifiers using two different approaches namely Majority Voting and Stacked Generalization.

For training the base classifiers, YamCha<sup>2</sup> toolkit along with TinySVM-0.092<sup>3</sup> is used. While training, a context window of size 3 is used (i.e.  $w_{i-3}, w_{i-2}, w_{i-1}, w_i, w_{i+1}, w_{i+2}, w_{i+3}$ ,

<sup>2</sup><http://chason.org/taku/software/yamcha/>

<sup>3</sup><http://chason.org/taku/software/TinySVM/>

where  $w_i$  is the current word) and the dynamic features are set at three (i.e. the output tags  $t_{i-3}, t_{i-2}, t_{i-1}$  of the three words  $w_{i-3}, w_{i-2}, w_{i-1}$  preceding the current word  $w_i$  will be considered).

In Majority Voting, the out of all base classifiers are combined together and the output of final system is decided based on majority voting. If majority voting fail then the highest performance output of the base classifiers is considered on final output. For Stacked Generalization, an open source implementation of CRF, CRF++ package<sup>4</sup>, has been used for constructing a CRF-based meta classifier.

The results of base classifiers and ensemble classifiers using Majority Voting and Stacking are shown in Table 4. The results show that, the best base classifier is the classifier based on IOBE SR model and the worst is the classifier based on IOE2 SR model. Both ensemble classifiers outperform the base classifiers and ensemble using stacking approach reported the best f-score.

## 5 Conclusion

Clinical-NER is a key task in health information systems. Different approaches have been applied for Clinical-NER. Ensemble approach tries to overcome the weakness of one approach by the strength of another. In our paper, we have designed an ensemble approach using majority voting and stacking techniques to combine the results of base classifiers. We have used SVM for learning base classifiers using different SR models and CRF classifier for learning the meta-classifier. Up to our knowledge, it is the first work that uses SR models for learning the base classifiers. The performance of our approach outperforms the performance of each of base classifiers.

## References

- Alan R Aronson and François-Michel Lang. 2010. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.
- Frédéric Béchet, Alexis Nasr, and Franck Genet. 2000. Tagging unknown proper names using decision trees. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, ACL ’00, pages 77–84, Stroudsburg, PA, USA. Association for Computational Linguistics.

<sup>4</sup><https://taku910.github.io/crfpp/>

		Training set	Test set	Total
<b>No. of Documents</b>		349	477	826
	Problem	11968	18500	30468
<b>Named Entities</b>	Treatment	8500	13560	22060
	Test	7369	12899	20268

Table 3: Statistics of i2b2 dataset

Classifiers	SR Model	F-score
Base Classifier	IOB2	77.31
	IOE2	76.06
	IOBE	77.48
	IOBES	77.21
Ensemble Classifiers	Stacking	<b>77.63</b>
	Majority Voting	<b>77.53</b>

Table 4: Results of base and ensemble classifiers

- Han-Cheol Cho, Naoaki Okazaki, Makoto Miwa, and Junichi Tsujii. 2013. Named entity recognition with multiple segment representations. *Information Processing & Management*, 49(4):954 – 965.
- A. Dehghan, J. A. Keane, and G. Nenadic. 2013. Challenges in clinical named entity recognition for decision support. In *2013 IEEE International Conference on Systems, Man, and Cybernetics*, pages 947–951, Oct.
- Asif Ekbal and Sriparna Saha. 2013. Stacked ensemble coupled with feature selection for biomedical entity extraction. *Knowledge-Based Systems*, 46(0):22 – 32.
- Carol Friedman, Philip O Alderson, John HM Austin, James J Cimino, and Stephen B Johnson. 1994. A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association*, 1(2):161–174.
- R. Grishman and B. Sundheim. 1996. Message Understanding Conference-6: A Brief History. In *Proceedings of the 16th International Conference on Computational Linguistics*, Copenhagen, June.
- Gurulingappa H, Hofmann-Apitius M, and Fluck J. 2010. Concept identification and assertion classification in patient health records. In *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data*.
- Scott Halgrim, Fei Xia, Imre Solti, Eithon Cadag, and Özlem Uzuner. 2010. Extracting medication information from discharge summaries. In *Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents*, Louhi

'10, pages 61–67, Stroudsburg, PA, USA. Association for Computational Linguistics.

George Hripcsak and Adam S. Rothschild. 2005. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association*, 12(3):296–298.

Chih-Wei Hsu and Chih-Jen Lin. 2002. A comparison of methods for multiclass support vector machines. *Trans. Neur. Netw.*, 13(2):415–425, March.

S. Keretna, C. P. Lim, and D. Creighton. 2014. A hybrid model for named entity recognition using unstructured medical text. In *2014 9th International Conference on System of Systems Engineering (SOSE)*, pages 85–90, June.

Sara Keretna, Chee Peng Lim, Doug Creighton, and Khaled Bashir Shaban. 2015. Enhancing medical named entity recognition with an extended segment representation technique. *Computer Methods and Programs in Biomedicine*, 119(2):88 – 100.

Taku Kudo and Yuji Matsumoto. 2001. Chunking with support vector machines. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8. Association for Computational Linguistics.

Dingcheng Li, Karin Kipper-Schuler, and Guergana Savova. 2008. Conditional random fields and support vector machines for disorder named entity recognition in clinical texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, BioNLP '08, pages 94–95, Stroudsburg, PA, USA. Association for Computational Linguistics.

Serguei Pakhomov, Ted Pedersen, and Christopher G Chute. 2005. Abbreviation and acronym disambiguation in clinical discourse. In *AMIA Annual Symposium Proceedings*, volume 2005, page 589. American Medical Informatics Association.

R. Polikar. 2006. Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, 6(3):21–45, Third.

R Core Team, 2017. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

- Lance A. Ramshaw and Mitchell P. Marcus. 1995. Text chunking using transformation-based learning. *In proceeding of the Third ACL Workshop on Very Large Corpora*.
- Adwait Ratnaparkhi. 1998. *Maximum entropy models for natural language ambiguity resolution*. Ph.D. thesis, University of Pennsylvania, PA, USA.
- Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.
- H. L. Shashirekha and H. A. Nayel. 2016. A comparative study of segment representation for biomedical named entity recognition. In *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1046–1052, Sept.
- Stephan Spat, Bruno Cadonna, Ivo Rakovac, Christian Gütl, Hubert Leitner, Günther Stark, and Peter Beck. 2008. Enhanced information retrieval from narrative german-language clinical text documents using automated document classification. *Studies in health technology and informatics*, 136:473.
- Jiashen Sun, Tianmin Wang, Li Li, and Xing Wu. 2010. Person name disambiguation based on topic model. In *CIPS-SIGHAN Joint Conference on Chinese Language Processing*, page 391.
- Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.
- David H Wolpert. 1992. Stacked generalization. *Neural networks*, 5(2):241–259.
- Yonghui Wu, Min Jiang, Jianbo Lei, and Hua Xu. 2015. Named entity recognition in chinese clinical text using deep neural network. *Studies in health technology and informatics*, 216:624.
- Yu-Chieh Wu. 2014. A top-down information theoretic word clustering algorithm for phrase recognition. *Information Sciences*, 275:213 – 225.
- Shaodian Zhang and Noémie Elhadad. 2013. Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. *Journal of Biomedical Informatics*, 46(6):1088 – 1098.