Supervised Named Entity Recognition for Clinical Data

Devanshu Jain

Dhirubhai Ambani Institute of Information and Communication Technology,
Gandhinagar, Gujarat, India 382007
devanshu.jain919@gmail.com

Abstract. Clinical Named Entity Recognition is a part of Task 1b, organised by CLEF eHealth organisation in 2015. The aim is to automatically identify clinically relevant entities in medical text in French. A supervised learning approach has been used for training the tagger. For the purpose of training, Conditional Random Fields(CRF) has been used. An extensive set of features was used for training. Precision, recall and F1 Score were used as evaluation metrics. Ten fold cross validation technique was used to evaluate the system. The best precision obtained was 0.91 and the best recall obtained was 0.66. After the test results were announced, the best F1 score obtained for exact matching was 0.67 and for relaxed case (i.e. inexact matching), it was 0.73.

Keywords: Clinical Data, Named Entity Recognition, UMLS, Machine Learning, CRF

1 Introduction

There is a huge amount of raw medical data available in the form of textual information. The goal of Information Extraction, here, is to present the data in a way that enhances user experience and allows better comprehension. The major task involved in this process is the identification of named entities within the document. This allows the user to have a better understanding of the jargons. It also allows to identify important terms that may be helpful in summarising the medical document.

The Clinical Named Entity Recognition is different from other common sequence tagging problems, like POS (Part of Speech) tagging. The major point of difference is the existence of ambiguity in the medical document. The span of an entity may overlap with the span of another entity i.e. the same word can be a part of multiple entities. Another issue with it is the presence of non contiguous entities. That is, the span of the entity may be discontinuous over a sentence. Moreover, there are ample resources such as thesaurus, but almost all of them are English centric. The training data, being in French language, poses another challenge.

In order to annotate the clinical entities, UMLS (Unified Medical Language System) is used. It is a compendium of vocabularies in biomedical science. There are various semantic groups, under which an entity can lie.

In order to tackle the problem, an extensive list of features were used. CRFsuite software was used to train the tagger. The following sections explains in greater details the methods and tools used for tackling the problem.

2 Approach

2.1 Overview

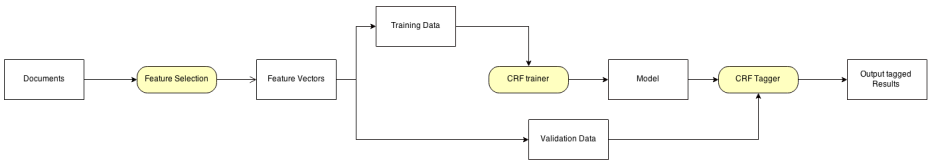

Fig. 1. Flow Diagram of the process.

The available data was first pre-processed by stemming all the words. We have considered the NER task as a sequence tagging problem. We, therefore have used CRF basic tagger to train the system. CRF is a state of the art method used for the purpose of sequence tagging. We have used CRFsuite software for this purpose.

2.2 Features

Following features were used for training the CRF:

1. **Lexical Features:** Uni-grams, bi-grams and tri-grams of words were used as features within a window of ± 3 around the current word. Then the POS tags of the words in the window were also chosen as features. Finally, we used capitalisation of letters and presence of digit as features too. In addition, prefixes and suffixes of 3 letter length were also used as features for each word.
2. **UMLS Features:** UMLS features were extracted using MetaMap. Since, MetaMap is English-centric, therefore, all the French words in the training data were translated into English, separately. Then, using MetaMap API, semantic group of these words were obtained and used as features for training.
3. **Global Features:** For every word, we calculated the position of the word in that sentence. To do this, we treated the word as *LEFT*, when it lied in the left quarter of the sentence. When it lied in the right quarter of the sentence, it was treated as being *RIGHT*. Otherwise, it was treated as being *CENTER*.

In order to account for the case when the span of entities was over multiple words in a contiguous manner, we used BIO format. For the starting of an entity, its name was prefixed with *-B*. If it was an intermediate word, the entity name was prefixed with *-I*. Otherwise, it was named as *O*. The system currently does not handle the discontinous terms.

3 Tools Used

1. Microsoft Bing translator was used for translating each french word (separately) into English.
2. Snowball stemmer was used for the stemming during pre-processing step
3. CRFSuite tagger was used to train the model based on the traing data and for tagging the test files.

4 Training Data

The training data was provided by the CLEF organisation itself. The data consisted of 833 MEDLINE documents, that contained single lines of medical data in French language. An additional 11 EMEA documents were also provided. Annotation files for each document were also given.

Some statistics of the training data is as follows:

Total Word Count	25,500
Number of Annotations	5,690
Non Contiguous Annotations	40
Overlapping Spans of entities	797

Table 1. Training data Stats

As can be seen, the non contiguous annotations account for just 0.7% of the total number of annotations and hence don't affect the validations, much.

5 Experiments

Precision, Recall and F1 measure were used as evaluation metrics to evaluate the system. These are defined as follows:

$$Precision = \frac{TruePositives}{TruePositives+Falsepositives}$$

$$Recall = \frac{TruePositives}{TruePositives+FalseNegatives}$$

$$F1Score = \frac{2*Precision*Recall}{Precision+Recall}$$

We used ten fold cross validation techniques. The available data was partitioned into training data and validation data. The partition was done randomly, i.e. random samples were taken from the data and used for validation. The partition was done four times in the ratio of 60:40, 70:30, 80:20 and 90:10. Ten different and random partitions for each ratio were created. Then, the average precision and recall was calculated.

When MetaMap was not used as an external source of information, following results were obtained:

Training %	**Validation %**	**Avg. Precision**	**Avg. Recall**	**Avg. F1 Score**
60%	40%	0.928	0.453	0.609
70%	30%	0.933	0.466	0.622
80%	20%	0.932	0.485	0.638
90%	10%	0.936	0.488	0.642

Table 2. Run1: Validation Results

When MetaMap outputs were used by the tagger, following results were obtained:

Training %	**Validation %**	**Avg. Precision**	**Avg. Recall**	**Avg. F1 Score**
60%	40%	0.928	0.515	0.662
70%	30%	0.925	0.521	0.667
80%	20%	0.926	0.531	0.675
90%	10%	0.919	0.541	0.681

Table 3. Run2: Validation Results

As can be seen, some improvement in recall was observed when MetaMap thesaurus was used in the system.

6 Official Results

6.1 Runs

We submitted two runs, which are described as follows:

1. **Run 1:** Predictions were completely made by CRFsuite based on the model generated using training data.

2. **Run 2:** Predictions made use of CRFsuite software as well as UMLS Metathesaurus information obtained by MetaMap. Whenever the model tagged a token as a non-entity, information from MetaMap was used to specify the tag for it. It was done at the level of single word only.

The runs were submitted for entity identification only. We didn't participate in entity normalisation.

6.2 Results

For the EMEA documents, following results were obtained:

Run	**Exact Match**			**Inexact Match**		
	Precision	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**
Run 1	0.8591	0.5478	0.669	0.9091	0.6085	0.7290
Run 2	0.3567	0.5792	0.4415	0.3926	0.6653	0.4938

Table 4. EMEA Results

For the MEDLINE titles, following results were obtained:

Run	**Exact Match**			**Inexact Match**		
	Precision	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**
Run 1	0.7126	0.4081	0.519	0.8188	0.5084	0.6273
Run 2	0.3973	0.4582	0.4256	0.4634	0.5845	0.5170

Table 5. MEDLINE Results

7 Conclusion

We present a supervised Clinical named Entity Recognition system that can detect the named entities from a French medical data, using an extensive list of features, with an F1 Score of 0.68. It also uses UMLS Metathesaurus information obtained by MetaMap, using the English translated version of French words.

The surprising thing to observe was that although, we used extra UMLS Metathesaurus information obtained by MetaMap in run2, although it improved the precision, but reduced the recall drastically. This remains to be investigated.

The system does not handle the non contiguous entities properly. We can use mutual information as a technique to identify the relationship between different entities, in order to detect that.

References

1. Clef e health: Task 1b 2015. https://sites.google.com/site/clefehealth2015/task-1/task-1b.
2. Crfsuite: a fast implementation of conditional random fields (crfs). http://www.chokkan.org/software/crfsuite/.
3. F. S. F. I. K. C. Czajkowski, K. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*.
4. A. A. et. al. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*.
5. L. Goeuriot, L. Kelly, H. Souminen, L. Hanlen, A. Névéol, C. Grouin, J. Palotti, and G. Zuccon. Overview of theclef ehealth evaluation lab 2015. In *CLEF 2015 - 6th Conference and Labs of the Evaluation Forum*. Lecture Notes in Computer Science (LNCS), Springer, September 2015.
6. A. Névéol, C. Grouin, X. Tannier, T. Hamon, L. Kelly, L. Goeuriot, and P. Zweigenbaum. CLEF eHealth evaluation lab 2015 task 1b: clinical named entity recognition. In *CLEF 2015 Online Working Notes*. CEUR-WS, 2015.