WI-ENRE in CLEF eHealth Evaluation Lab 2015: Clinical Named Entity Recognition Based on CRF

Jingchi Jiang[1], Yi Guan[1], Chao Zhao[1]

[1]School of Computer Science and Technology,
Harbin Institute of Technology, Harbin, China

jiangjingchi0118@163.com,
guanyi@hit.edu.cn, hitsa.zc@gmail.com

Abstract. Named entity recognition of biomedical text is the shared task 1b of the 2015 CLEF eHealth evaluation lab, which focuses on making biomedical text easier to understand for patients and clinical workers. In this paper, we propose a novel method to recognize clinical entities based on conditional random fields (CRF). The biomedical texts are split into sections and paragraphs. Then the NLP tools are used for POS tagging and parsing, and four groups of features are extracted to train the entity recognition model. In the subsequent phase for entity normalization, the MetaMap of Unified Medical Language System (UMLS) tool is used to search for concept unique identifiers (CUIs) category. In addition, CRF++ package is adopted to recognize clinical entities in another phase for entity recognition. The experiments show that our system named as WI-ENRE, is effective in the named entity recognition of biomedical texts. The $F_{measure}$ of EMEA and MEDLINE reach to 0.56 and 0.45 respectively in exact match.

Keywords: Named Entity Recognition, Conditional Random Fields, UMLS

1 Introduction

With the application of EMRs, hospitals and medical institutions generate masses of biomedical text. Based on biomedical text, the medical big data analytics and the building of heath knowledge network are the critical problem. As a precondition to solve the problem, named entity recognition can provide a solution to extract information and knowledge from biomedical text. Hence, the named entity recognition is becoming a research hotspot.

Biomedical text contains a wealth of information on patients covering their hospital stays, including health conditions, diagnoses, performed tests and treatments. Named entity recognition form biomedical text has a good research foundation[1,2]. In previous years, several NLP shared tasks have addressed information extraction tasks such as 2010 i2B2/VA Challenge[3] as well as identifying protected health information (PHI) at 2014 i2b2/UTHealth challenge. The 2013 ShARe/CLEF eHealth T2 task[4] was required to detect disorders spans and their concept unique identifiers (CUIs). On that basis, the 2014 ShARe/CLEF eHealth T2 shared task[5] focused on extracting information from biomedical text. In 2015, the CLEFeHealth addresses clinical named entity recognition on task 1b[6,7]. The aim is to automatically identify clinically relevant entities in medical text with French rather than English.

Methods for entity recognition can be roughly divided into three categories: rule-based, machine learning methods and a combination of both. The method of rule-based mainly relies on proper nouns dictionaries and rules which wrote by language experts or domain experts to identify the clinical entities. Compared to rule-based methods, many more researchers choose machine learning methods on entity recognition.

In this paper, we propose a novel method for task 1b of CLEFeHealth 2015. In order to testify this method, we design a named entity recognition system, WI-ENRE, which adopts machine learning method based on conditional random fields for the nine categories and lexicon-based approach for geographic areas.

The rest of this paper is arranged as follows. In Sec. 2, we discuss the materials and methods in detail, and also focus on feature optimizing selection. Moreover, we conduct the experiments to testify the effectiveness of WI-ENRE in Sec. 3. In Sec. 4, we conclude this paper and discuss the directions for further work.

2 Methods

In this study, the dataset which is called QUAERO French Medical Corpus[8] is provided by 2015 CLEFeHealth shared tasks. The training set consists of 11 text files with corresponding annotation files from EMEA and 833 text files with annotation files from MEDLINE. 80% of the text files from MEDLINE and EMEA folders are selected as the training data of model, respectively, while the remaining files are used for testing.

In the process of entity recognition and entity normalization, some related resources are used, which contain Stanford Parser based on French and UMLS tool. Then, the feature selection will be described as the significant part in this paper. Finally, the principle of conditional random field algorithm will be detailed in Sec. 2.4.

2.1 Data

The corpus is provided by the 2015 CLEFeHealth evaluation lab. The task 1b consists of clinical named entity recognition and entity normalization from the file of MEDLINE titles and EMEA documents.

Table 1. Description of the corpus.

	Training	Test
MEDLINE Documents	667	166
EMEA Documents	9	2
MEDLINE Words	8,406	2,149
EMEA Words	13,754	1,187
MEDLINE Entities	2,383	612
EMEA Entities	2,357	338
MEDLINE Entities(Deduplication)	1,879	541
EMEA Entities(Deduplication)	848	166

Table 2. Statistics of each category from the training corpus.

Category	MEDLINE	EMEA
Anatomy(ANAT)	495	247
Chemical and Drugs(CHEM)	346	727
Devices(DEVI)	39	48
Disorders (DISO)	963	736
Geographic Areas (GEOG)	34	22
Living Beings (LIVB)	297	273
Objects (OBJC)	27	71
Phenomena (PHEN)	60	19
Physiology (PHYS)	160	119
Procedures (PROC)	574	433

In order to testify the method of entity recognition, the training set provided by CLEFeHealth is divided into two parts: the dataset for training which contains 676 documents and a total of 22,160 words, and the testing set contains 168 documents and a total of 3,336 words. Moreover, the number of entity and deduplicated entity are counted, respectively (as shown Tab. 1). In Tab. 2, we also give a few statistics for each category in the training corpus.

2.2 Resources

Stanford Parser. As an existing open source toolkit, Stanford Parser is utilized to split sentences of the biomedical text. Furthermore, Stanford Parser also provides the function of POS tagging for multi-languages, such as English, Chinese, French, German and so on.

UMLS. Unified Medical Language System (UMLS) is used for mapping clinical entity to the unique concept identifiers (CUIs). And MetaMap[9] is a highly configurable application to map biomedical text to the UMLS metathesaurus or equivalently to identify metathesaurus concepts. This is the case of task 1b which is required to recognize clinical entities and their CUIs.

2.3 Feature Selection

Before model training, a large number of features need to be extracted from biomedical texts. The features can be categorized into four groups: lexical features, orthographic features, context features and lexicon features, listed in Tab. 3.

Lexical features use the first and the last four characters of token to identify the categories of entities. The POS of a token is helpful in named entity recognition. The Stanford Parser tool is used to get POS tag of token, which is learnt on open domain corpus and supports multiple languages by loading template.

The tokens similar in shape can help the classifier “memorize” whether the token belong to one type of the entities. We replaced uppercase letters, lowercase letters, letters with diacritics and digits in a token by “A”, “a”, “b” and “0”, respectively. Length of a token is a significant feature to clinical entity recognition. Similarly, information of capital letters is also a strong feature to help us identify the entities which always consist of uppercase letters. For example, the tokens of "Bio-safety Cabinet", "CT" and other proper noun can be identified by capital feature.

The context features of the classifier contain the lowercase, first four characters, last four characters, POS tags of two tokens before and after the current token.

Table 3. Features used in the CRF classifier.

Category	Feature
Lexical features	lowercase of the current token
	first four characters of the current token
	last four characters of the current token
	POS of the current token
Orthographic features	shape of the current token
	length of the current token
	whether the current token contains a letter
	whether the current token begins with a capital letter
	whether all characters in the current token are capital letters
	whether the current token contains a digit
	whether all characters in the current token are digits
	whether the current token consists of letters and digits
Context features	first four characters of two previous tokens
	first four characters of two next tokens
	last four characters of two previous tokens
	last four characters of two next tokens
	POS of two previous tokens
	POS of two next tokens
Lexicon feature	whether the current token is in the "GEOG" dictionary

Finally, a dictionary of geography based on French is extracted from webpage[10] of city, state and country. All the words in the dictionaries are lowercased. Lexicon features are used to judge whether the lowercase of the current token is in the dictionary or not, rather than as a feature of CRF model. If the current token shows up in the "GEOG" dictionary, we can conclude this token belongs to the entity of geographical category

After the features of token are generated, extracting an optimal subset from all the features is the most important step for building an effective classification model. At present, search algorithms can be divided into complete-based search, heuristic-based search and random-based search. The sequential forward selection (SFS) and sequential backward selection (SBS) based on heuristic are the most commonly-used algorithms for selecting features. Beginning with an empty feature subset X, SFS add a feature x into X, and ensure the optimal performance of evaluation function J(X). After n-times iteration, the classification model is constructed based on local optimum. Instead of SFS, SBS starts a full feature set, and eliminate a feature from the feature set for each iteration.

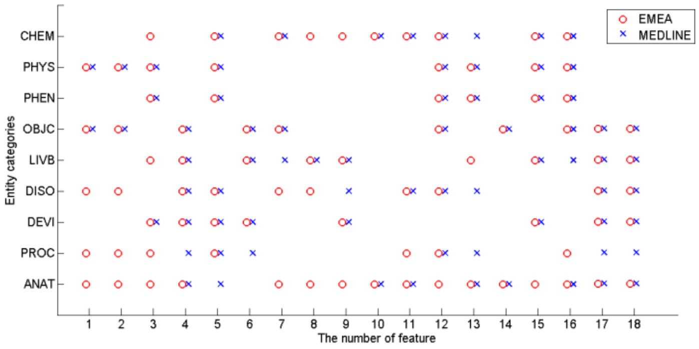

Fig. 1. The experiment is done to testify the effectiveness of BDS. The vertical and horizontal axes represent entity categories and feature categories, respectively. According to the different entity categories, WI-ENRE extracts the different feature set for building CRF model.

Compared with the above algorithms, we design and realize the bidirectional search (BDS) algorithm which combines the advantages of SFS and SBS, and improves the efficiency. The main idea of BDS is that SBS is used to search features, which is beginning with a full feature subset, while using SFS algorithm to search features beginning with an empty feature subset. Until a same feature subset is searched from both of SFS and SBS after n-iteration, BDS uses the same feature subset as the final results. After the selection step, the results for the different categories are shown in Fig. 1.

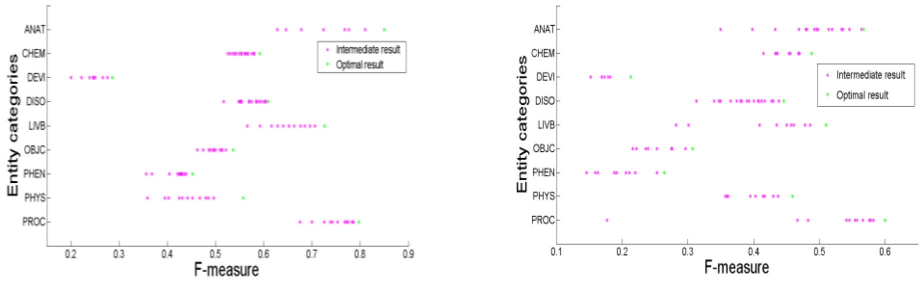

Fig. 2. The experiments of EMEA and MEDLINE demonstrate that the $F_{measure}$ of each categories change with the increase of iterations, and the most optimal combination of feature can be selected, respectively.

Furthermore, we list the $F_{measure}$ of the intermediate result, which is generated either SFS or SBS, in the process of n-iteration. For each category of entity, the most optimal combination of feature can be selected by BDS as shown in Fig. 2. Although the method of feature selection may make out the local optimum, it can give better results than full feature subset for the feature selection of different entity categories.

2.4 Conditional Random Field

The conditional random field algorithm is proposed by Lafferty in 2001. CRF is arbitrary undirected graphical model that bring together the best of generative models and Maximum Entropy Markov Models (MEMM). A potential function is defined as follow:

$$\phi_{y_c}(y_c) = \exp(\sum_k \lambda_k f_k(c, y \mid c, x)) \tag{1}$$

Where $\phi_{y_c}(y_c)$ is a potential function of the fully connected network of Y, which is built on undirected graph. $y \mid c$ represents random variables which correspond to the cth node in the fully connected network by boolean form. Given an observed sequence of tokens, $x = x_1 x_2 ... x_n$, CRF can predicts a corresponding sequence of labels, $y^* = y_1 y_2 ... y_n$. y^* , which maximizes the conditional probability $p(y \mid x)$, is defined as follow:

$$p(y \mid x) = \frac{1}{Z(x)} \exp(\sum_{c \in C} \sum_k \lambda_k f_k(c, y_c, x)) \tag{2}$$

The conditional random field algorithm is widely used in named entity recognition. The existing open source toolkit CRF++[11] is utilized to classify the tokens in a sequence into the BIO scheme. The "B" indicates a token is the beginning of the clinical entity. The "I" represents that a token is inside of the clinical entity. The "O" means that a token does not belong to any category of the clinical entity.

Firstly, the training and testing data are generated based on the features. A CRF model can be learnt after training on the training data which is described in Sec. 2.1. Then the tokens in the testing data can be classified into one of the entity categories or non-entity category using CRF model.

3 Experiments

3.1 System Design

The WI-ENRE system consists of two main modules, ten sub modules and one evaluation module. The purpose of this system is to automatically identify clinically relevant entities in medical text in French.

- One of the major components is the named entity recognition module, which can identify the clinical entity based on Conditional Random Field and generate the specific model for each category. In the pre-processing, the biomedical texts are divided into two parts: MEDLINE and EMEA. Then, using the CRF model to recognize the clinical entity, the results will be evaluated and determined whether the feature set should be optimized. Until the results meet the optimization condition, the CRF models will be stored in the model repositories.

- The second module integrated with UMLS can select the CUIs to map clinical entity, and generate the annotated biomedical texts automatically. Besides English, UMLS does not support the other languages, such as French, Chinese and so on. Therefore, the API of Google is used to translate the entities from French to English in the first step. Then the translated entities are put into UMLS and mapped to the CUIs which is selected with the first result.

In the part of named entity recognition, the first step is the preprocessing of the file, which contains the part-of-speech tagging by Stanford Parser and the generation of training files based on entity category. The next step includes the training of CRF model, the decoding of CRF by testing files and the evaluation of entity results. Then the module of feature optimization is performed until the optimum result is found. Finally, all of the optimum model for each category will be stored into model repositories.

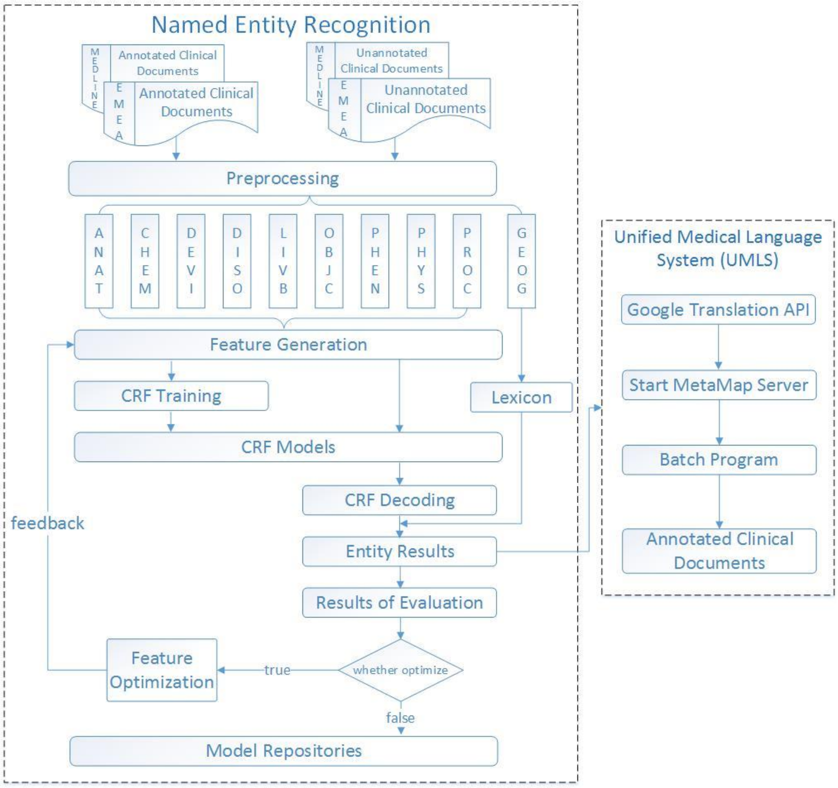

Fig. 3. The flow diagram of the WI-ENRE system is shown in this figure.

3.2 Evaluation Metrics

For task 1b, we determined the performance of WI-ENRE by comparing the system outputs against reference standard annotations. The system performance and performance for each category are evaluated rigorously. Precision, recall and $F_{measure}$[12] are calculated from true positive, false positive and false negative annotations, which are described as follows:

true positive (TP) = the annotation cue span from WI-ENRE overlapped with the annotation cue span from the reference standard

false positive (FP) = an annotation cue span from WI-ENRE did not exist in the reference standard annotations

false negative (FN) = an annotation cue span from the reference standard did not exist in WI-ENRE annotations

The formulas of the precision, recall, F_{measure} are shown in Eqs. (3) - (5).

$$Precision = TP/(TP+FP) \tag{3}$$

$$Recall = TP/(TP+FN) \tag{4}$$

$$F_{measure} = 2 * Recall \bullet Precision/(Recall + Precision) \tag{5}$$

3.3 Recognition Accuracy

Using the evaluation metrics described above, the results of the WI-ENRE system are shown in Tab. 4 and Tab. 5.

Table 4. Results for each category/Phase 1 (EMEA):

	TP	FN	FP	Precision	Recall	F_{measure}
GEOG	22	7	3	0.880	0.759	0.815
DISO	225	233	141	0.615	0.491	0.546
LIVB	141	135	2	0.986	0.511	0.673
CHEM	183	687	18	0.910	0.210	0.342
OBJC	15	35	2	0.882	0.300	0.448
PHEN	4	6	6	0.400	0.400	0.400
PHYS	29	111	11	0.725	0.207	0.322
DEVI	2	20	3	0.400	0.091	0.148
ANAT	123	32	46	0.728	0.794	0.759
PROC	160	90	13	0.925	0.640	0.757
Exact match (official)	971	1,289	234	0.429	0.805	0.56
Inexact match (official)	1,137	1,123	156	0.503	0.879	0.64

The evaluation results of EMEA and MEDLINE are presented respectively. The experiments show that results of EMEA are better than MEDLINE. In the 10 main categories, GEOG based on lexicon get the high Fmeasure above 80 and 70 percent in different corpus. Compared to GEOG, the categories which are based on CRF, such as ANAT, PROC and LIVB, have a low Fmeasure about 70 percent.

Table 5. Results for each category/Phase 1 (MEDLINE):

	TP	FN	FP	Precision	Recall	$F_{measure}$
GEOG	28	18	4	0.875	0.609	0.718
DISO	279	613	199	0.584	0.313	0.407
LIVB	142	178	28	0.835	0.444	0.580
CHEM	108	259	40	0.730	0.294	0.419
OBJC	8	27	10	0.444	0.229	0.302
PHEN	10	39	19	0.345	0.204	0.256
PHYS	31	120	53	0.369	0.205	0.264
DEVI	7	47	8	0.467	0.130	0.203
ANAT	232	262	78	0.748	0.470	0.577
PROC	267	302	188	0.587	0.469	0.521
Exact match (official)	1,068	1,909	671	0.358	0.614	0.452
Inexact match (official)	1,523	1,454	449	0.511	0.772	0.615

In addition, the rest categories are worse than ANAT, PROC and LIVB, with below 50 percent. Through the analysis, it is observed that the entity categories of low accuracy do not basically select the orthographic features which are inside the feature range of 6th and 11th (as shown in Fig.1). Moreover, we also found that the entity categories which select the feature of POS get higher percentage of accuracy than others.

3.4 Error Analysis

The errors in the WI-ENRE system are analyzed according to the error analysis method[13], which is roughly divided into three groups: type error (entity is correct but type is wrong), missing error (entity is in the gold standard but not in the system output) and spurious error (entity is in the system output but not in the gold standard). Based on the types of errors, Tab. 6 lists the error distribution of WI-ENRE system.

Table 6. Error distribution of the WI-ENRE system at the clinical entity recognition of CLEFeHealth 2015 task 1b:

	Error number	Percentage
Type error	101	1.65%
Missing error	3,221	52.72%
Spurious error	872	14.27%

According to the three groups of error, missing errors make up the highest proportion as 52.72%. Therefore, the recall of the WI-ENRE system is very low.

Table 7. Error details of the WI-ENRE system at the clinical entity recognition of CLEFeHealth 2015 task 1b:

	System output											
	ANAT	CHEM	DEVI	DISO	GEOG	LIVB	OBJC	PHEN	PHYS	PROC	missing	total
ANAT		2		1		1			2		294	6
CHEM	2						1		1	1	946	5
DEVI	1									1	67	2
DISO	3	3	1			3			1	10	846	21
GEOG											25	0
LIVB	2			2					2		313	6
OBJC			3								62	3
PHEN				1					1	2	45	4
PHYS	2			17						2	231	21
PROC		1		22				4	2		392	29
Spurious	124	58	11	340	7	30	12	25	64	201		872
total	10	6	4	43	0	4	1	8	9	16	3,221	

The experiment shows that the categories of CHEM and DISO have high missing error with the count of 946 and 846, respectively. Twenty-two PROC entities are identified as DISO while 10 DISO entities are marked as PROC. It is difficult to distinguish between PROC and DISO for WI-ENRE. In addition, ANAT, LIVB, PHYS have a missing count of above 200. All of these led to the low recall rate of WI-ENRE system. Compare to missing errors, the spurious errors of DISO are also much higher than others. It follows that the system cannot recognize the category of DISO well, which not only has the higher missing errors but also is the most serious error of spurious. For the type error, a normal level which can be remained within acceptance criteria is shown in Tab. 7.

4 Conclusion

This paper described the clinical entity recognition by machine learning method for the task 1b of CLEFeHealth 2015. A suite of methods that included conditional random fields, feature selection with BDS algorithm and entity normalization using MetaMap performed the task well. Among these methods, the feature selection plays a crucial role to enhance the performance for each category. Using a suitable feature subset, we can obtain more accurate and reasonable classification than the full feature set. In order to testify this method, we design the system, WI-ENRE, to address the clinical entity based on CRF and achieve the normalization of clinical entity by UMLS.

The future study will be focused on the feature optimization and the improvement of recall rate. Moreover, the term vectors which are generated by word embedding can be taken as the characterizing attribute. The other useful features and more suitable methods will be researched to improve our system.

Acknowledgments. The MEDLINE title and EMEA documents used in this paper were provided by CLEFeHealth 2015 task 1b, and thanks to the organizing committee of CLEF and the annotators of the dataset.

References

1. Carol Friedman, Philip O. Alderson, John H. M. Austin, James J. Cimino, and Stephen B. Johnson. A general natural-language text processor for clinical radiology. J Am Med Inform Assoc, 1(2):161-174, 1994.
2. Pierre Zweigenbaum. Menelas: an access system for medical records using naturallanguage. Computer Methods and Programs in Biomedicine, 45:117-120, 1994.
3. Ozlem Uzuner, Brett R. South, Shuying Shen, and Scott L. DuVall. 2010 i2b2/VAchallenge on concepts, assertions, and relations in clinical text. J Am Med InformAssoc, 18(5):552-556, Sep-Oct 2011. Epub 2011 Jun 16.
4. Hanna Suominen, Sanna Salantera, Sumithra Velupillai, Wendy W. Chapman, Guergana K. Savova, No_emie Elhadad, Sameer Pradhan, Brett R. South, Danielle L. Mowery, Gareth J. F. Jones, Johannes Leveling, Liadh Kelly, Lorraine Goeuriot, David Martinez, and Guido Zuccon. Overview of the ShARe/CLEFeHealth evaluation lab 2013. In Proceedings of CLEF 2013, Lecture Notes in Computer Science, Berlin Heidelberg, 2013. Springer.
5. Liadh Kelly, Lorraine Goeuriot, Gondy Leroy, Hanna Suominen, Tobias Schreck, Danielle L. Mowery, Sumithra Velupillai, Wendy W. Chapman, Guido Zuccon and Joao Palotti. Overview of the ShARe/CLEFeHealth evaluation lab 2014. In Proceedings of the ShARe/CLEFeHealth Evaluation Lab. Springer-Verlag, 2014.
6. Goeuriot,L.,Kelly,L.,Suominen,H.,Hanlen,L.,Névéol,A.,Grouin,C.,Palotti,J.,Zuccon,G.: Overview of the clef ehealth evaluation lab 2015. In: CLEF 2015 - 6th Conference and Labs of the Evaluation Forum. Lecture Notes in Computer Science (LNCS), Springer (September 2015)
7. Névéol, A., Grouin, C., Tannier,X., Hamon, T.,Kelly,L., Goeuriot, L., Zweigenbaum, P.: CLEF eHealth evaluation lab 2015 task 1b: clinical named entity recognition. In: CLEF 2015 Online Working Notes. CEUR-WS (2015)
8. Névéol, A., Grouin, C., Leixa, J., Rosset, S., Zweigenbaum, P.: The QUAERO French medical corpus: A resource for medical entity recognition and normalization. In: Proc of Bio TextMining Work. pp. 24--30 (2014)
9. Alan A. Aronson: Effective mapping of biomedical text to the UMLS Metathesaurus: the Metamap program. In: AMIA, pp. p.17-21. (2001).
10. http://www.culturecommunication.gouv.fr/.
11. CRFsuite package: http://www.chokkan.org/software/crfsuite/.
12. Hripcsak, G., Rothschild, A.: Agreement, the F-measure, and reliability in information retrieval. J Am Med Inform Assoc 12(3) 296-8.
13. B. Wellner, M. Huyck, S. Mardis et al., “Rapidly retargetable approaches to de-identification in medical records,” Journal of the American Medical Informatics Association, vol. 14, no. 5, pp. 564-573, 2007.