IHS-RD-BELARUS: Clinical Named Entities Identification in French Medical Texts

Maryna Chernyshevich, Vadim Stankevitch

IHS Inc. / IHS Global Belarus
131 Starovilenskaya St., 220123, Minsk, Belarus
{Marina.Chernyshevich, Vadim.Stankevitch}@ihs.com

Abstract. In this paper we present the results of our participation in the Task 1b of the 2015 CLEFeHealth challenge, whose goal was the identification of clinical entities of various types from medical texts in French and its normalization. We used the CRF-based system developed for disorder recognition in English and enhanced with French knowledge resources to recognize 10 types of clinic named entities from French medical texts: Anatomy, Chemical and Drugs, Devices, Disorders, Geographic Areas, Living Beings, Objects, Phenomena, Physiology and Procedures. Our system's performance in entity recognition task was evaluated at 0.70 and 0.52 F-measure in exact match mode and 0.80 and 0.70 F-measure in inexact match mode depending on test corpus. The obtained results are higher than the average of all submitted runs.

Keywords: Named-entity recognition, Biomedical Texts, Medical NLP

1 Introduction

Electronic medical records are of great value for both patients and health professionals as well as for multiple related domains and industries. Providing medical domain with automated tools for plain text processing, data extraction and classification is nowadays a challenge of major importance. To facilitate the development of effective approaches for the analysis of biomedical texts, corresponding shared tasks have been organized such as CLEF 2013 [5], SemEval 2014 [9], and SemEval 2015 [7]. The CLEF eHealth 2015 shared task initiates the research in languages other than English and our goal was to automatically identify clinically relevant entities in medical text in French and to normalize these entities to a specific UMLS Concept Unique Identifier (CUI) [10], [11]. The task was divided into three subtasks: plain entity recognition, normalized entity recognition and entity normalization.

Although our team participated in all subtasks our efforts were focused mainly on entities recognition task. For the normalization problem we have implemented a simple straightforward approach based on lookup in the UMLS terminology with quite poor performance. This approach needs further development.

In this paper we present a supervised CRF-based named entity recognition (NER) system that is capable of recognizing 10 types of clinical named entities from French medical texts: Anatomy, Chemical and Drugs, Devices, Disorders, Geographic Areas, Living Beings, Objects, Phenomena, Physiology and Procedures. The system was

originally developed for English disorder identification in order to participate in the SemEval-2015 shared task "Analysis of clinical text" [6]. Its performance was evaluated at F-measure of 0.898 for English dataset.

To adapt the system we added French lexicons and omitted some features not relevant for the proposed medical texts, for example, document section feature.

2 Materials and methods

2.1 Data

The dataset provided by the organizers is called QUAERO French Medical Corpus and comes from the European Medicines Agency (EMEA) and Medline [1]. This dataset has been developed as a resource for named entity recognition and normalization in 2013.

The training set contains 833 MEDLINE titles and 11 EMEA documents and the test set contains 832 MEDLINE titles and 12 EMEA documents.

The annotation of clinical entities was guided by concepts in the Unified Medical Language System (UMLS) [8] and covers 10 types of clinical entities. The training set contained 5,690 annotations while the test set contained 5,237 annotations. Table 1 represents the distribution of the annotated entities among the categories. The entities were annotated in a comprehensive fashion, so that nested entities were marked, and entities could be mapped to more than one UMLS concept. For instance, in the phrase "infarctus du myocarde" (myocardial infarction), the mention "myocarde" (myocardium) should be annotated with category "ANATOMY" (CUI C0027061) and the mention "infarctus du myocarde" should be annotated with category "DISORDER" (CUI C0027051) [11].

Table 1. Distribution of annotated entities

	Training set	Test set
Anatomy	742	649
Chemical and Drugs	1073	1237
Devices	87	76
Disorders	1699	1350
Geographic Areas	56	75
Living Beings	570	596
Objects	98	85
Phenomena	79	59
Physiology	279	291
Procedures	1007	819

2.2 Entities recognition system

Clinical entity recognition can be thought of as a sequence segmentation problem: each word is a token in a sequence to be assigned a label. The most popular and powerful sequential learning model is Conditional Random Fields (CRFs) – undirected statistical graphical models, a special case of which is a linear chain that corresponds to a conditionally trained finite-state machine [4].

Pre-processing

To facilitate feature generation for supervised CRF learning, sentences were pre-processed with French IHS Goldfire Linguistic Processor that performs the following operations: word splitting, part-of-speech tagging, parsing, noun phrase extraction, semantic role labelling within extended Subject-Action-Object (eSAO) relations [3].

Feature Set

Given a sentence S and a token under consideration W_k, we define features over W_k and window of 5 tokens: W_{k-2}, W_{k-1}, W_k, W_{k+1}, W_{k+2}.

1. **Lexical features:** Canonical form of the token W_k itself is used as feature. In order to model local context of the word this class of features also includes canonical forms of neighbouring words in the window [-2,+2].
2. **Orthographic features:** This set of features is used to represent case and characters of the token $\boldsymbol{W_k}$.

Letter case: token contains only upper case characters, token contains only lower case characters, first character is in upper case and the word is not the first in the sentence, token contains at least one upper case and one lower case characters.

Characters: token contains intra-word dash, token contains slash, token is a digit or contains a digit, token is a punctuation mark.

3. **Part of speech feature:** We include as features the part of speech information produced by IHS Goldfire Linguistic Processor.
4. **Word frequency in out-of-domain corpus**: We used social media texts as an out-of-domain corpus to calculate the word frequencies. The feature has four values: very rarely, rarely, frequently and vary frequently used with a empirically determined thresholds.
5. **Knowledge-based features:** In addition to orthography and syntactic structures, the model could also benefit from generalized semantic word groups. This sort of semantic domain knowledge can be provided in the form of lexicons. We created two types of lexicons: clinical lexicon and general lexicon.

The clinical lexicon was created using the 2014AA multilingual release of the UMLS Metathesaurus. It contains about 5 million entities for English and 2500 for French. We created dictionaries for each of 10 categories for both English and French. To comply with the annotation guidelines, each category combines many UMLS semantic types. For example, category ANATOMY encompasses following semantic types: Anatomical Structure; Body Location and Region; Body Part, Organ or Organ Component; Body Space or Junction; Body Substance; Body System; Cell; Cell Component; Embryonic Structure; Fully Formed Anatomical Structure; Tissue.

The general lexicon consists of lists of words from general domain translated automatically form English to French: materials ("*métal*"), units of measure (*"ml"*), person's professions (*"infirmière"*). These lexicons were originally created for English using the WordNet [2]. We have selected some top-level nodes, for example, physical property, human, process etc. and all subordinate terms were assumed to belong to the appropriate category [6].

Using the lexicons following features were assigned to each token:

Clinical lexicon features: 10 features representing presence of token or sequence of tokens in particular category. The value of the features indicates quantity of tokens

in a sequence that match lexicon exactly. For example, all tokens in the phrase **"**infarctus du myocarde" become value 3 for a disorder feature and token "myocarde" becomes value 1 for an anatomy feature. For lexicon entries that are multi-word, all words are required to match in the input sequence.

General lexicon features: This feature represents the semantic class to which the token belongs.

Classification

Based on the fact that there are nested and overlapping entities of different types we decided to model the problem as a supervised classification into two classes (target entity or not) instead of multi-class classification.

We created 10 training corpora with the same set of properties but with different entities labelled and then converted the sets into a BIO format, in which each word is assigned into one of three labels: B means the beginning of an entity, I means the inside of an entity, and O means the outside of an entity. In case of embedded entities of the same type we dismissed the entity of lower length. For example, in the training set the entity *"maladie de Parkinson"* is annotated as disorder and the word *"maladie"* is also annotated as separate disorder mention, we left only the entity *"maladie de Parkinson"*.

We trained 10 CRF models and then simply merged the classification results of all models. We didn't use any post-processing step to analyze cases of contradicting predictions of the classifiers, for example when the same token was recognized as entity of different type.

2.3 Entities normalization

For the entities normalization subtask we implemented a simple algorithm that chooses all possible variants of normalized name for an entity. We generated putatively related strings, i.e. variants, synonyms and translations to English, and selected all CUIs that include all words from the particular entity variant. This approach generated large amount of normalized CUI variants for some ambiguous entities like *"traitement"* and caused thereby very low precision.

In the future we are going to implement a ranking algorithm to select the best of all CUI variants.

3 Results

The system's ability to correctly identify the clinical entities was evaluated using precision, recall, and F-measure.

Evaluation was carried out under two settings:

- exact match: a predicted mention is considered a true positive if the predicted span is exactly the same as for the gold-standard mention;
- inexact match: a predicted mention is a true positive if there is any word overlapping between the predicted mention span and the gold standard span.

A total of seven teams participated in the task, submitted 10 system runs. The results of our best submitted run compared to average and median results are summarized in the table 2.

Our best run produced 0 in exact match for EMEA due to a technical issue: we submitted predictions with shifted spans of start and end positions. The results with corrected spans are included in the table 3 under "Later submitted run" title. Our system obtained very competitive results: 0.80 and 0.70 F-measure on test set under inexact match setting and 0.70 and 0.52 F-measure under exact match setting. These results are almost two times better than the Average/Median results. The official rank of submissions is not published at the time of publication.

Table 2. System results for entity recognition subtask under exact and inexact evaluation settings (F-measure)

	EMEA		MEDLINE	
	exact match	inexact match	exact match	inexact match
Best submitted run	0	0.80	0.52	0.70
Later submitted run	**0.70**	**0.80**	**0.52**	**0.70**
Average	0.31	0.40	0.39	0.57
Median	0.22	0.55	0.45	0.66

Our system performs well in recognizing chemical and drugs, living beings and disorders, but it fails in recognizing phenomena and devices. Table 3 provides more precise named entity classification results by categories.

Table 3. System results for particular entity type in entity recognition subtask (F-measure)

	EMEA		MEDLINE	
	exact match	inexact match	exact match	inexact match
Anatomy	0.71	0.80	0.59	0.68
Chemical and Drugs	0.77	0.86	0.58	0.70
Devices	0.38	0.47	0.10	0.17
Disorders	0.63	0.79	0.56	0.80
Geographic Areas	0.45	0.45	0.46	0.49
Living Beings	0.82	0.85	0.54	0.60
Objects	0.61	0.56	0.14	0.15
Phenomena	0.12	0.11	0.03	0.07
Physiology	0.40	0.43	0.22	0.26
Procedures	0.74	0.84	0.48	0.58

These results demonstrate the dependence of CRF classification performance on the training set volume. The types that were rarely encountered in the training set (Devices, Phenomena, Objects) have the lowest F-measure.

In order to determine the importance of individual features, ablation experiments were carried out. Table 4 shows the resulting changes in the F-measure in inexact match mode. Rows are ordered by features set impact on the full gold standard. Positive values indicate that a feature group has a negative impact on classification quality: results are improved by omitting the features.

Table 4. Results of feature ablation experiments (F-measure)

	EMEA inexact match	MEDLINE inexact match
All features:	0.80	0.70
– knowledge-based features	0.62 (-0.18)	0.52 (-0.18)
– lexical features	0.67 (-0.13)	0.58 (-0.12)
– orthographic features	0.79 (-0.01)	0.70 (0.0)
– out-of-domain frequency	0.79 (-0.01)	0.70 (0.0)
– part-of-speech feature	0.84 (+0.04)	0.68 (-0.02)

The most important features are knowledge-based features, closely followed by lexical features. Other features contributed relatively small individual effects, but were necessary to achieve the overall performance in combination.

As for the entities normalization subtask, our system performs poorly and needs further development. The submitted results of the system are illustrated in table 5 and are compared against the average and median results.

Table 5. System results for entity normalization subtask under exact and inexact evaluation settings (F-measure)

	EMEA		MEDLINE	
	exact match	inexact match	exact match	inexact match
Best submitted run	0.10	0.69	0.07	0.57
Average	0.61	0.82	0.47	0.65
Median	0.87	0.89	0.67	0.68

4 Conclusion

In this paper we presented a supervised statistical system originally developed for English disorder recognition and adapted for French language to participate in shared task 1b of the CLEF eHealth 2015 lab on Clinical Named Entity Recognition.

Our system makes use of CRF for identifying clinical entities of 10 types: Anatomy, Chemical and Drugs, Devices, Disorders, Geographic Areas, Living Beings, Objects, Phenomena, Physiology, Procedures. The results achieved for the entity recognition subtask are quite promising: 0.80/0.70 F-measure in inexact match and 0.70/0.52 F-measure in exact match mode depending on test corpus. These results are close to the ability of our system to recognize disorder mentions in English texts: F-measure of 0.89. This fact proves promising adaptability of proposed approach to different languages. Although these results are positive, there is still room to improve the systems. In future, we would like to explore semi-supervised learning approaches to take advantage of large amount of unannotated clinical text. It would be also interesting to adapt the proposed system to other languages.

In the future we will focus especially on the entity normalization subtask to improve our result.

References

1. Névéol, A., Grouin, C., Leixa, J., Rosset, S., Zweigenbaum, P.: The QUAERO French Medical Corpus: A Ressource for Medical Entity Recognition and Normalization. Fourth Workshop on Building and Evaluating Ressources for Health and Biomedical Text Processing – BioTxtM2014, 24-30 (2014)
2. Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K.: Introduction to WordNet: An on-line lexical database. Technical report, Princeton. CSL Report 43, revised March 1993
3. Todhunter, J., Sovpel, I., and Pastanohau, D.: System and method for automatic semantic labeling of natural language texts. U.S. Patent 8 583 422, November 12, 2013
4. Lafferty, J., McCallum, A., and Pereira, F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: Proceedings of the Eighteenth International Conference on Machine Learning (2001)
5. Suominen, H., Salanterä, S., Velupillai, S., Chapman, W. W., Savova, G., Elhadad, N., Pradhan, S., South, B.R., Mowery D. L., Jones G. J. F., Leveling J., Kelly, L., Goeuriot, L., Martinez, D., and Zuccon, G.: Overview of the shARe/CLEF eHealth evaluation lab 2013. In: Proceedings of ShARe/CLEF eHealth Evaluation Labs (2013)
6. Chernyshevich, M. and Stankevitch, V.: IHS-RD-Belarus: Identification and Normalization of Disorder Concepts in Clinical Notes. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), 380–384, Denver, Colorado, June 4-5, 2015. c 2015 Association for Computational Linguistics (2015)
7. Elhadad, N., Pradhan, S., Lipsky Gorman, S., Manandhar, S., Chapman, W., Savova, G.: SemEval-2015 Task 14: Analysis of Clinical Text. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), 303–310. Denver, Colorado, June 4-5, 2015. c 2015 Association for Computational Linguistics (2105)
8. Bodenreider, O.: The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. Nucleic Acids Research, 267–270 (2004)
9. Pradhan, S., Elhadad, N., Chapman, W., Manandhar, S. and Savova, G.: SemEval-2014 Task 7: Analysis of Clinical Text. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pp.54- 62 (2014)
10. Goeuriot, L., Kelly, L., Suominen, H., Hanlen, L., Névéol, A., Grouin, C., Palotti, J., and Zuccon, G.: Overview of the CLEF eHealth Evaluation Lab 2015. In: Proceedings of CLEF 2015 - 6th Conference and Labs of the Evaluation Forum. Lecture Notes in Computer Science (LNCS), Springer (2015)
11. Névéol, A., Grouin, C., Tannier, X., Hamon, T., Kelly, L., Goeuriot, L. and Zweigenbaum, P.: CLEFeHealth Evaluation Lab 2015 Task 1b: clinical named entity recognition. In: Proceedings of CLEF 2015. CLEF 2015 Online Working Notes, CEUR-WS (2015)