

A Supervised Named-Entity Extraction System for Medical Text

Andreea Bodnari^{1,2}, Louise Deléger², Thomas Lavergne²,
Aurélie Névéol², and Pierre Zweigenbaum²

¹ MIT, CSAIL, Cambridge, Massachusetts, USA

² LIMSI-CNRS, rue John von Neumann, F-91400 Orsay, France

Abstract. We present our participation in Task 1a of the 2013 CLEF-eHEALTH Challenge, whose goal was the identification of disorder named entities from electronic medical records. We developed a supervised CRF model that based on a rich set of features learns to predict disorder named entities. The CRF system uses external knowledge from specialized biomedical terminologies and Wikipedia. Our system performance was evaluated at 0.598 F-measure in the context of strict evaluation and 0.711 F-measure in the context of relaxed evaluation.

Keywords: Named-entity recognition, Natural Language Processing, Medical records, Machine learning

1 Introduction

Electronic medical records (EMRs) represent rich data repositories loaded with valuable patient information. Automated tools are required to process this patient information and make it available to medical professionals and specialized medical systems. These automated tools take as input the plain text of EMRs and output data of interest. For example, named-entity extraction tools process the plain text of EMRs and extract instances of named entities (i.e., noun phrases) that can be classified into a certain semantic category.

The 2013 CLEF-eHEALTH challenge aims to develop methods and resources that make EMRs more understandable by both patients and health professionals. The challenge spans over three tasks. We participate in the first task, which focuses on the identification of disorder named entities in electronic medical records, and develop a system that can perform NER on medical text. We propose a solution that combines a rich feature set with external knowledge gathered from both specialized and general domain knowledge repositories.

We present a named-entity recognition system specialized in the medical domain. Our system learns a CRF model from the training data, based on a rich feature set that combines external knowledge sources with information gathered from the EMR text. We discuss the system design, present the system results on the 2013 CLEF-eHEALTH [10] training and test data, and discuss the specific features that helped most with the system performance.

2 Related work

The natural language processing (NLP) community has organized specialized competitions to evaluate and help improve the state of the art in various NLP domains. Competitions in the general domain include the Text Retrieval Evaluation Conferences [2], the Semantic Evaluation (SemEval) [1], and the Conference on Natural Language Learning (CoNLL) shared-tasks [5]. In the medical domain, the Informatics for Integrating Biology and the Bedside (i2b2) center organized a series of NLP competitions focused on information extraction from unstructured clinical documents. The NLP competitions tried to support advancement in a series of NLP tasks like information extraction [13], information retrieval [11], semantic textual similarity, co-reference resolution [12]. Tasks like information retrieval, dependency parsing, and named entity recognition exhibit relatively well performing solutions that can be applied in real life settings. Yet, results of NLP competitions showed that the research community is still struggling in the specialized domain (i.e., medical, biomedical) compared to the general domain.

3 Materials and methods

3.1 Data

The corpus used for the 2013 CLEF-eHEALTH challenge consists of de-identified plain text EMRs from the MIMIC II database, version 2.5 [8]. The EMR documents were extracted from the intensive-care unit setting and included discharge summaries, electrocardiography reports, echo reports, and radiology reports. The training set contained 200 documents and a total of 94,243 words, while the test set contained 100 documents and a total of 87,799 words (see Table 2).

Annotation of disorder noun phrases (NPs) was carried out as part of the Shared Annotated Resources project [7]. The text of each EMR document was annotated by two professional coders trained for this task, followed by an open adjudication step. A disorder noun phrase is defined as any span of text which can be mapped to a concept in the SNOMED-CT terminology and which belongs to the Disorder semantic group. A concept is in the Disorder semantic group if it belongs to one of the following Unified Medical Language System (UMLS) [4] semantic types: congenital abnormality, acquired abnormality, injury or poisoning, pathological function, disease or syndrome, mental or behavioral dysfunction, experimental model of disease, anatomical abnormality, neoplastic process, signs and symptoms. The training set contained 5,874 annotations, while the test set contained 5,351 annotations (see Table 2). The non-contiguous entities accounted for approx. 10% of the training and test data.

3.2 System design

We developed a supervised linear-chain Conditional Random Fields (CRF) model using 10-fold cross validation on the training set data. We first present the pre-processing steps we performed on the datasets. We then describe the model

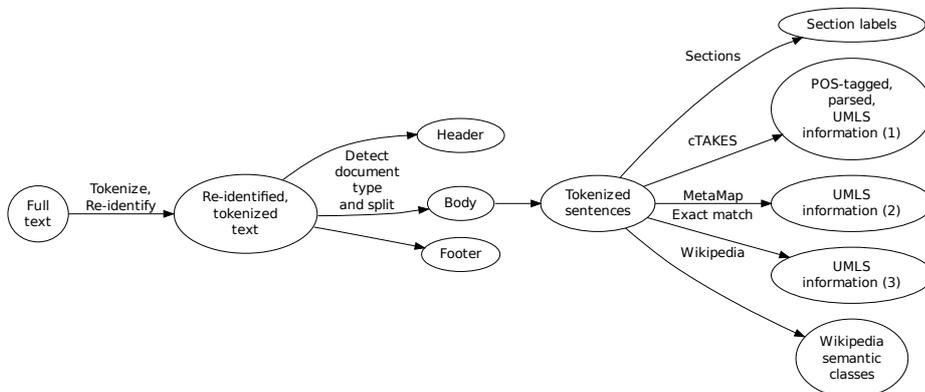
Table 1. Description of training and test data sets

	Training	Test
Word count	94,243	87,799
Annotation count	5,874	5,351
Non-Contiguous annotation count	660	439
Documents	200	100

feature set together with the CRF feature patterns. The feature production architecture is schematized in Figure 1.

Data pre-processing Before using the challenge corpora for training and testing, we performed several pre-processing steps:

- the training and test corpora provided by the challenge organizers were de-identified and thus contained special de-identification marks; to turn de-identification code into more normal phrases, we performed re-identification with pseudonyms on the input text.
- EMR documents present in general a well-structured form, with a header, document body, and a footer. The header and footer contain information relevant to clinical administration, but the disorder NPs are only encountered inside the document body. We thus removed the header and footer from the EMR documents and performed analysis on the document body only.

**Fig. 1.** Diagram of feature production.

System features Given a sentence $s = \dots w_{-2}w_{-1}w_0w_1w_2\dots$ and a token of interest w_k , we define features over w_k and n-grams centered at w_k .

1. **Lexical and morphological features:** we include information on the token and on the token lemma in the form of unigrams over w_{k-1} , w_k , w_{k+1} and bigrams over $w_{k-1}w_k$. We also include as unigram features the following token characteristics: token contains only upper case characters, token is a digit, is capitalized, or is a punctuation. Additionally we include as unigram features over w_k token suffixes ranging from 1 to 4 characters. Finally we add a 5-gram feature which detects patterns containing two slashes, such as “m/r/g”, which may reveal up to three disorders (such constructs are split into 5 tokens by our tokenizer), and apply it over w_{k-4} , w_{k-2} , w_k (the non-slash positions of the pattern).
2. **Syntactic features:** we tokenize and parse the EMR plain text using the cTakes [9] system. We include as features the part of speech information in the form of unigrams over w_{k-3} , w_{k-2} , \dots , w_{k+3} and bigrams over w_{k-3} , \dots , w_{k+2} . We further include the parse-tree dependency information for the current token w_k , the bigram $w_{k-1}w_k$, and trigram $w_{k-1}w_kw_{k+1}$.
3. **Document structure features:** the feature set contains as features the document type (e.g., radiology report, discharge summary) and the section type (e.g., Findings, Laboratory data, Social History). We extract the section type using a rule-based section extraction tool that identifies the occurrence of section names within the EMR. The section extraction tool uses a list of 58 manually defined section names. Both document type and section type are unigram features over w_k .
4. **UMLS features:** we include UMLS information from three sources. We first use the UMLS information provided by cTakes: the semantic type unique identifier (defined over unigrams w_{k-1} , w_k , w_{k+1} and bigram $w_{k-1}w_k$), and semantic group information (defined over unigrams w_{k-1} , w_k , w_{k+1} and bigrams $w_{k-1}w_k$ and w_kw_{k+1}). Secondly, we process the EMR plain text using MetaMap [3] and include the semantic group it identifies. We use an additional UMLS mapping where we directly search for UMLS noun phrases within the EMR text through exact match and include the semantic group and concept unique identifier (CUI) of the identified phrase. The MetaMap and the direct UMLS mapping features are unigram features over w_{k-1} , w_k , w_{k+1} . We also define a unigram binary feature for being a member of the Disorder semantic group and one for being a member of the Anatomy semantic group.
5. **Wikipedia features:** we make use of the Wikipedia Category information in order to classify the noun-phrases contained in EMRs. We group the Wikipedia categories into nine semantic groups: disorder, body part, living being, chemicals, phenomenon, object, geographical location, devices, and ‘other’. The ‘other’ category contains the Wikipedia categories not included in any of the defined categories. We use the article titles from the English Wikipedia and search for their occurrence within the EMR plain text. Once an article title is found we map its Wikipedia category to one of the categories

previously defined. We define the Wikipedia system feature as unigram over w_{k-1}, w_k, w_{k+1} .

All features pertaining to multi-token expressions instead of only single tokens (for instance, being a UMLS term with a given semantic group, or being an abbreviation) are encoded with the begin inside outside (B-I-O) scheme: given a label L , the first token is labeled B- L , the next tokens are labeled I- L , and tokens not having this feature are labeled O. All features can use both unigrams and bigrams of *classes*: this leverages the specific capabilities of linear-chain CRFs to label sequences instead of isolated tokens.

Problem formulation We model the problem as a supervised classification task with three labels: B-Disorder, I-Disorder, and O (outside). We include in the gold standard of the training set the contiguous entities and only partially took into account the non-contiguous entities. Observing that non-contiguous entities generally follow the general SNOMED v3.5 model, with a morphology / dysfunction part, which is akin to a disorder, and another (generally, anatomy) part, we only include the morphology / dysfunction part of non-contiguous entities, based on their UMLS semantic types, labelling them with B-Disorder and I-Disorder classes. This allows us to handle them as though they were contiguous entities without polluting the gold standard labels with disorder-labeled anatomy segments that would perturb training and classification.

We used the Wapiti [6]¹ implementation of CRFs because it is fast and offers convenient patterns (e.g., patterns using regular expressions on field values).

4 Results and discussion

4.1 Evaluation metrics

We evaluate the system’s ability to correctly identify the spans of disorder NPs. The evaluation measures are precision, recall, and F-measure, defined as

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

$$F\text{-measure} = \frac{2 * Precision * Recall}{Precision + Recall} \quad (3)$$

where

TP = count of system NPs presenting same span as gold standard NPs;

FP = count of system NPs presenting divergent span from gold standard NPs;

FN = count of gold standard NPs not present in the system NPs.

¹ <http://wapiti.limsi.fr/>

We compute the system NP span overlap to the gold standard NP under two settings: the strict evaluation setting, where the system NP span is identical to the gold standard NP span, and relaxed evaluation setting, where the system NP span overlaps the gold standard NP span.

4.2 Results

Our best run evaluated at 0.730 F-measure on the training data and 0.598 F-measure on the test data under the strict evaluation setting; under the relaxed evaluation setting, it obtained an 0.887 F-measure on the training data and 0.711 F-measure on the test data. In general, the system precision was higher than the system recall (0.814 precision on the test data vs. 0.473 recall under strict evaluation, and 0.964 precision vs. 0.563 recall on the test data under relaxed evaluation). The system recall is lower as we did not handle the non-contiguous NPs that accounted for approx. 10% of the training and test data.

Table 2. System results on training and test data sets under strict and relaxed evaluation settings.

	Training			Test		
	Precision	Recall	F measure	Precision	Recall	F measure
Strict	0.791	0.677	0.730	0.814	0.473	0.598
Relaxed	0.961	0.823	0.887	0.964	0.563	0.711

4.3 Discussion

The 2013 CLEF-eHEALTH Task 1a is the second challenge after the 2010 i2b2/VA Shared-Task aiming to identify disorder named entities in clinical text. Even though the final goals of the two challenges were similar, they differed in several structural points. First, the 2010 i2b2/VA corpus had token-based annotations and was already segmented into tokens, thus requiring no further pre-processing. In contrast, the 2013 CLEF-eHEALTH corpus marked annotations at the character level and pre-processing was desirable, which motivated the first steps of our pipeline. Secondly, entity boundaries were defined differently in the two challenges. A first example is by the determiners which were included as part of the annotation in the 2010 i2b2/VA Shared-Task but were excluded in the 2013 CLEF-eHEALTH challenge (e.g., *a brain tumor* vs *brain tumor*). A second example is the non-contiguous entities included only in the 2013 CLEF-eHEALTH challenge (e.g., given the EMR sentence “The pain reported by the patient is occurring in lower back”, the annotated entity is “pain...in lower back”). A final difference between the two challenges is the corpus size: approx. 18,550 problem entities in the i2b2 training corpus, compared to 5,874 disorder entities in CLEF-eHEALTH training corpus.

Error analysis on the training data revealed some systematic mistakes performed by our NER model. The majority of the incorrectly predicted entities were NPs with morphological structure resembling the one of the disorder NPs (e.g., *anastomosis* contains the Greek suffix ‘-osis’ meaning abnormal condition and resembles the name of several disorders like *necrosis*, *osteoporosis*, but can be both a disorder and a procedure), disorder names used as findings (e.g., the noun phrase *esophageal varices* is a finding based on the context *which showed non-bleeding grade III esophageal varices*), and disorder entities used in a negated context (e.g., NP *fasciculations* inside the context *tongue midline without fasciculations*). The system also predicted parts of NPs it was trained on as being standalone entities; for example, the phrase *left ventricular* was predicted as a disorder entity after the system encountered the phrase *left ventricular aneurysm* during training.

In general, our NER system failed to identify the non-contiguous named entities (it could at best identify some of their parts), abbreviations or NPs rarely encountered in the training set (e.g., *aaa*, *3vd*, *inability to walk*), misspelled disorder entities (e.g., *hematochezeia*) and the full span of several NPs longer than 2 tokens (e.g., *chronic subdural hematoma*). Out-of-vocabulary tokens, i.e., terms not found in the UMLS because of lack of coverage, variants, or misspellings, were an important source of lack of recall.

5 Conclusion and perspectives

We present a clinical NER system designed for participation in Task 1a of the 2013 CLEF-eHEALTH challenge. We design our system as a CRF with a rich feature set and external knowledge gathered from specialized terminologies and general domain knowledge repositories. Our system evaluates at 0.598 F-measure in the strict evaluation context and 0.711 F-measure in the relaxed evaluation context, obtaining a mid-range position.

Our entity detection system presents good precision but performs worse in terms of recall. In order to improve system recall, additional textual data can be integrated into the CRF model. We expect that including the Brown word clustering information as part of the feature vector would provide a fallback for some out-of-vocabulary tokens and help increase accuracy based on its unsupervised word classes. The non-contiguous entities are only partially handled by our system, thus better handling of all parts of non-contiguous entities, for instance through syntactic dependencies, should result in improved recall.

6 Acknowledgments

This work was partly funded through project Accordys² funded by ANR under grant number ANR-12-CORD-0007-03. The first author was funded by the

² Accordys: *Agrégation de Contenus et de CO*nnaissances *pour Raisonner à partir de cas de DY*Smorphologie *foetale*, Content and Knowledge Aggregation for Case-based Reasoning in the field of Fetal Dysmorphology (ANR 2012-2015).

Châteaubriand Science Fellowship 2012–2013. We acknowledge the Shared Annotated Resources (ShARe) project funded by the United States National Institutes of Health with grant number R01GM090187.

References

1. ACM. SemEval Portal. http://aclweb.org/aclwiki/index.php?title=SemEval_Portal. Accessed July 19, 2012.
2. Timothy G. Armstrong, Alistair Moffat, William Webber, and Justin Zobel. Improvements that don't add up: ad-hoc retrieval results since 1998. In *Proceedings of the 18th ACM conference on Information and knowledge management, CIKM '09*, pages 601–610, New York, NY, USA, 2009. ACM.
3. Alan A. Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: the Metamap program. In *Proceedings of the AMIA Annual Symposium*, pages 17–21. American Medical Informatics Association, 2001.
4. Olivier Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acid Res*, 32:D267D270, 2004.
5. CoNLL. CoNLL: the conference of SIGNLL. <http://ifarm.nl/signll/conll/>. Accessed July 19, 2012.
6. Thomas Lavergne, Olivier Cappé, and François Yvon. Practical very large scale CRFs. In *ACL Proc*, pages 504–513, 2010.
7. Shared Annotated Resources. Shared Annotated Resources. https://www.clinicalnlpannotation.org/index.php/Main_Page. Accessed May 24, 2013.
8. M. Saeed, M. Villarroel, A.T. Reisner, G. Clifford, L. Lehman, G.B. Moody, T. Heldt, T.H. Kyaw, B.E. Moody, and R.G. Mark. Multiparameter intelligent monitoring in intensive care II (MIMIC-II): A public-access ICU database. *Clinical Care Medicine*, 39:952–960, 2011.
9. Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of American Medical Informatics Association*, 17:507–513, 2010.
10. Hanna Suominen, Sanna Salanterä, Wendy W. Sumitra Velupillai Chapman, Guergana Savova, Noemie 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/CLEF eHealth evaluation lab 2013. In *Proceedings of CLEF 2013*, Lecture Notes in Computer Science, Berlin Heidelberg, 2013. Springer. To appear.
11. Özlem Uzuner. Recognizing obesity and comorbidities in sparse data. *Journal of American Medical Informatics Association*, 16(4):561–570, Aug 2009.
12. Özlem Uzuner, Andreea Bodnari, Shuying Shen, Tyler Forbush, John Pestian, and Brett South. Evaluating the state of the art in coreference resolution for electronic medical records. *Journal of American Medical Informatics Association*, 17:514–518, February 2010.
13. Özlem Uzuner, Brett R. South, Shuying Shen, and Scott L. DuVall. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556, 2011.