Terminologies augmented recurrent neural network model for clinical named entity recognition

Ivan Lerner, Nicolas Paris, Xavier Tannier

Version of Record: https://www.sciencedirect.com/science/article/pii/S1532046419302734
Manuscript_0b08fe9007ac75a4be490614b8763e8c

Terminologies augmented recurrent neural network model for clinical named entity recognition

Ivan Lerner, MSc[1-2]; Nicolas Paris, MSc[2-3]; Xavier Tannier, PhD[4]

[1] Paris University, Paris, France

[2] AP-HP, DSI-WIND, Paris, France

[3] LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay, F-91405, Orsay, France

[4] Sorbonne Université, Inserm, Univ Paris 13, Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances pour la e-Santé, LIMICS, F-93017 Bobigny, France

Corresponding author: Xavier Tannier

Sorbonne Université, Inserm, Univ Paris 13, Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances pour la e-Santé, LIMICS, F-93017 Bobigny, France
E-mail: xavier.tannier@sorbonne-universite.fr

Abstract

Objective

We aimed to enhance the performance of a supervised model for clinical named-entity recognition (NER) using medical terminologies. In order to evaluate our system in French, we built a corpus for 5 types of clinical entities.

Methods

We used a terminology-based system as baseline, built upon UMLS and SNOMED. Then, we evaluated a biGRU-CRF, and a hybrid system using the prediction of the terminology-based system as feature for the biGRU-CRF. In French, we built APcNER, a corpus of 147 documents annotated for 5 entities (Drug names, Signs or symptoms, Diseases or disorders, Diagnostic procedures or lab tests and Therapeutic procedures). We evaluated each NER systems using exact and partial match definition of F-measure for NER. The APcNER contains 4,837 entities, which took 28 hours to annotate. The inter-annotator agreement as measured by Cohen's Kappa was substantial for non-exact match (K = 0.61) and moderate considering exact match (K = 0.42). In English, we evaluated the NER systems on the i2b2-2009 Medication Challenge for Drug name recognition, which contained 8,573 entities for 268 documents, and i2b2-small a version reduced to match APcNER number of entities.

Results

For drug name recognition on both i2b2-2009 and APcNER, the biGRU-CRF performed better that the terminology-based system, with an exact-match F-measure of 91.1% versus 73% and 81.9% versus 75% respectively. For i2b2-small and APcNER, the hybrid system outperformed the biGRU-CRF, with an exact-match F-measure of 85.6% versus 87.8% and 88.4% versus 81.9% respectively. On APcNER corpus, the micro-average F-measure of the hybrid system on the 5 entities was 69.5% in exact match and 84.1% in non-exact match.

Conclusion

APcNER is a French corpus for clinical-NER of five types of entities which covers a large variety of document types. The extension of the supervised model with terminology has allowed an easy increase in performance, especially for rare entities, and established near state of the art results on the i2b2-2009 corpus.

Keywords: "clinical natural language processing";"named entity recognition"; "information extraction"; "machine learning"; "APcNER"

Highlights

- For 28 hours of annotation time, we have built APcNER, a French corpus for clinical named-entity recognition that covers a large variety of document types
- APcNER allowed us to achieve on average 84% of non-exact F-measure over five types of clinical entities
- We provide consistent results on English and French corpora that give insight into the complementarity of terminology with a supervised model

Introduction

Within the range of data covered by electronic health records (EHRs), clinical documents (e.g., discharge summaries or physicians' letters) are rich sources of information for various applications such as patient enrollment in clinical research studies, epidemiological surveillance, medical coding and decision-making tools [1]. Information extraction tools must be tailored for applications in the medical field, where the language is both unstructured (e.g., free text) and semi-structured (e.g., drug lists), with a wide vocabulary.

Named-entity recognition (NER) is the process of mention detection and type classification of named entities, where named entities are concepts that can be referenced by various linguistic expressions. In recent decades, there has been a growing interest in clinical-NER, the task of NER for medical concepts such as drug names, diseases, or signs [2]. Supervised systems based on machine learning have proven to be more efficient than rule-based and terminology-based systems for NER [3]. Research efforts have then been made to unify these methods in hybrid systems, in a purely unsupervised [4,5] or semi-supervised fashion [6–8]. Such approaches are motivated by the necessity to reduce the need for manually annotated examples in the case of a supervised system, or the need for handwritten rules by experts in the field of rule-based and ontological systems. For instance, a supervised model can automatically learn context cues for the various expressions that are associated with the mentions of drug names, such as "*patient was given […]"* or "*[...] was stopped early*." A model relying on such context cues would increase its ability to detect new drug names or variants of drug name expressions that are not yet included in terminologies.

In addition, for languages other than English, annotated corpora and ontologies are scarcer. For instance, in French, there is only one annotated clinical corpus which covers a small subset of the medical domain [9], and international ontologies such as the Unified Medical Language System (UMLS®) are not fully translated [10]. The development cost of such annotated corpus is very high, as it has been reported that annotating 5 medical documents for 12 entities take on average 82 mins [9].

In this study, we aimed to evaluate a clinical-NER system for which development could scale up to the different uses of a large French data warehouse [11]. First, we built an annotated corpus for five clinical entity types, for which we present here the details of the annotation process. Second, we evaluated three different systems: 1) a terminology-based system built upon the Unstructured Information Management Architecture (apache-UIMA®) framework, 2) a supervised neural model based on a biGRU-CRF architecture, and 3) a hybrid system. This comparison focused on drug name recognition and was achieved on our French corpus as well as on a well-known, freely available corpus in English, for comparison purposes. On the French corpus exclusively, we evaluated the three systems on the task of jointly detecting the 5 types of annotated entities (Drug names, Signs or symptoms, Diseases or disorders, Diagnostic procedures or lab tests and Therapeutic procedures).

Methods

Corpus and annotation process

We used two corpora in this study, an English corpus from the i2b2-2009 Medication challenge (i2b2-2009) [12] and a French corpus APcNER, with clinical reports extracted from the Assistance Publique - Hôpitaux de Paris (AP-HP) data warehouse [11].

i2b2-2009 corpus

The original corpus included 1243 de-identified discharge summaries, 17 of which were annotated by the i2b2 team, and 251 collectively annotated by the challenge participants.[1] Overall, the 268 annotated documents of the corpus contained 337,745 tokens, 8,573 entities and 17,933 sentences, for a vocabulary of 23,214 tokens. The median sentence length was 13. The overall 8,573 entities comprised 6,488 (75%) unigrams, 1,053 (12%) bigrams and 1,029 (12%) longer mentions. These 268 discharge summaries were randomly assigned to a train (70%), development (15%) and test set (15%). We kept only the drug name annotations, and rejected the dose annotations and other drug-related information.
We also used a randomly sampled subset of the i2b2-2009 train set, to include the same number of drug name entities than the APcNER corpus, for quantitative comparison purposes. We call this smaller corpus “i2b2-small."

[1] Note that the original test set used during the i2b2 challenge is made of the 251 collectively annotated documents.

APcNER corpus

Document selection

We randomly sampled (stratified on document types) 147 documents from the dataset used for de-identification at AP-HP, excluding prescriptions and admission reports. The AP-HP de-identification dataset is a set of 3,223 French-language medical documents sampled (with upsampling of rare documents types) from over 50 million documents from the AP-HP data warehouse, which included EHR data from 39 hospitals. The APcNER included 4 main types of documents: discharge summaries, letters from physicians, operating reports and additional examination reports. Detailed document types can be found in the supplementary Table S1.

Annotation process

We based our annotation guideline on the UMLS® semantic types [13], Table 1 details the 5 medical entities that we annotated (Drug names, Signs or symptoms, Diseases or disorders, Diagnostic procedures or lab tests and Therapeutic procedures). We used BRAT Rapid Annotation Tool (BRAT) [14]. The general guidance for annotation was to annotate the most complete entities (e.g., “Non-ST segment elevation myocardial infarction” and not only “myocardial infarction”) with no possible overlap between entities (see Appendix Section 2 for the detailed annotation guideline).

Documents were annotated by groups of 10 and the annotation time monitored. Documents were pre-annotated using a terminology-based annotator (see below). IL, a medical resident, annotated all the dataset. Then, in order to estimate the quality of the guidelines, we randomly selected 10 documents that NG, a medical doctor, annotated blindly. We then assessed the agreement between the two annotators as F-measures for each class, considering IL as gold standard, and the overall Cohen’s Kappa, computed with exclusion of the outside class. Conflicting annotations have been discussed between IL and NG and are referenced in the Annotation Guideline (see Appendix Section 2). IL went through all the documents a second time in order to disambiguate the conflicting annotations.

For homogenization purposes, we fitted a simple conditional random field (CRF) model to the dataset with the default NER features, using Wapiti, an efficient open-source library for non-neural sequence labelling algorithms [15]. We used this model to detect annotation inconsistencies or case errors, and manually corrected them. Finally, we randomly divided all the documents into 6 folds by stratifying on the types of documents and the length of the documents. The corpus can be made available on condition that a research project is accepted by the scientific and ethics committee of the AP-HP health data warehouse (https://recherche.aphp.fr/eds/recherche/).

Results of the annotation process

The first round took on average 87 min per 10 documents and a total of 21 hours. The second round took on average 28 min per 10 documents and a total of 7 hours. Both rounds took on average 115 minutes per 10 documents and a total of 28 hours. Overall, the 147 documents of the corpus contain 80,421 tokens, 4,837 entities and 3,093 sentences, for a vocabulary of 12,523 tokens. The median sentence length was 14. The inter-annotator agreement after the first round and before the CRF harmonization is reported in Table 2. Cohen's Kappa was substantial for non-exact match (K = 0.61) and moderate considering exact match (K = 0.42).

Clinical-NER systems

For all experiments, we used the inside, outside, beginning (IOB) tagging scheme [16]. For an entity of type DRUG, the first token of such entity is coded B-DRUG, if the entity is constituted of multiples tokens, the following tokens are coded I-DRUG, and all tokens outside entities are coded O. For instance, "*placed on heparin sodium*" is encoded "O O B-DRUG I-DRUG."

Terminology based system

In English, we extracted drug names using regular expressions from UMLS® (including SNOMED 3.5 CT®) to create a large dictionary of drug names. In French, we used 10 terminologies, 8 of which were previously referenced in [10] (ATC, BPDM, CCAM, CIM-10, DRC, SNOMED, UMLS), and 2 terminologies held by AP-HP (GLIMS, QDOC). Table 1 details the extracted UMLS semantic types by entity type. Terms were extracted using minimal regular expression rules, and then tokenized using Stanford CoreNLP [17]. We discarded common terms based on Wikipedia word count. The matching rules were based on the apache-UIMA framework, CoreNLP and dkPRO and allowed multiple words matching, stop words, accent normalization and case insensitive matching. The source code is available with the GLP-3 license (https://github.com/EDS-APHP/uima-aphp/tree/master/uima-dict). Resolution of conflicting (overlapping) entities was done by randomly picking one of the conflicting entities.

Supervised system

Sentence segmentation and tokenization were performed using Stanford CoreNLP [17]. Numbers were normalized to a unique token. We learned a biGRU-CRF (Bidirectional - Gated Recurrent Unit - Conditional Random Field) [3], based on the NCRF++ implementation [18]. The model takes 2 types of inputs. First, English or French word

embeddings trained with the Skip-Gram model [19], from 2 million AP-HP documents for the French version (dimension 200), and 2 million MIMIC [20] clinical notes for the English version (dimension 100). Second, character embeddings processed by 1-dimensional convolution (kernel size 3) with max-pooling. The global token representation is the concatenation of the word and character embeddings (see Figure 1). The sequence of token representation is then processed forward and backward by the biGRU, which outputs an emission probability score for each entity class. Finally, the CRF decodes the sequence of labels by associating the emission probability score with a transition probability score (see Figure 2). We used a dropout rate of 0.5, an L2-norm on the model weights and early stopping to prevent overfitting. We used Bayesian optimization [21] to perform hyperparameter tuning of the architecture (number of layers, number of neurons, character embedding dimension), learning rate and L2-norm. Note that with 1 entity type, NER is a 3 class classification problem, and with *k* entity types it is a *k* x 2 + 1 classification problem, e.g., with the 5 entities types of APcNER there are 11 classes to predict.

Hybrid system

We proposed a hybrid system in which a supervised model is associated with a terminology-based model. For each token, we added a feature representing the class predicted by the terminology-based system described above, which is then encoded as a categorical embedding of dimension 5. This embedding is then concatenated to the word embedding from the supervised system. This terminology based feature can take two values per entity types (e.g., B-Drug Name; I-Drug name), as well as one value for the “Outside” class. We also added a context feature based on the terminology of section titles that we have developed internally. The French section headings terminology was created based on documents of the same distribution as the APcNER corpus, the English section heading terminology was created for the 2018 AP-HP Datathon based on MIMIC notes. For each token, the context feature was the class of the last section heading, following the order of the document. Then, the context feature is encoded as a categorical embedding of dimension 5. This embedding is concatenated to the word embedding from the supervised system and the other feature embedding. These feature embeddings are initialized randomly and learned during the optimization procedure.

Evaluation methodology and metrics

First, we evaluated the systems on the i2b2-2009 corpus, its reduced version i2b2-small, and on the APcNER corpus with labels limited to drug names. Then we evaluated the systems on the entire APcNER corpus, as one multi-class task. We compared the models

based on F-measure, precision and recall using the CONLL definition: "*precision is the percentage of named entities found by the learning system that are correct. Recall is the percentage of named entities present in the corpus that are found by the system. A named entity is correct only if it is an exact match of the corresponding entity in the data file.*" We also compared the model based on partial match, which allowed the boundaries of the entity to mismatch.

Neural network models' training is highly non-deterministic and is subject to the random seed choice. Because of this variability during the training phase, we performed five experiments for each model presented in this work, and reported the mean, and 95% confidence interval assuming a t-distribution of each metric. Note that such confidence interval only accounts for the variability originated from the optimization procedure, and not the variability originated when sampling the documents from the data warehouse.

First, using default parameters of NCRF++, we tested the main features of the architecture, ablation of the CRF, GRU versus LSTM, ablation of character embeddings, and different set of fast-text and skip-grams embeddings. This first step led to the selection of our baseline architecture, the biGRU-CRF with skip-gram and character embeddings. Second, we performed fined tuning of this architecture. For the i2b2-2009 corpus, we selected the optimal set of hyperparameters for the supervised model based on a development set, including the optimal epoch stop, and evaluated on the test set the models trained on the train+dev set. For APcNER, we selected the optimal set of hyperparameters for the supervised model by 6 fold cross-validation, we defined the optimal epoch stop as the mean of the optimal epoch stop for each fold. We evaluate the model on each fold, after training it on the remaining 5 folds, using the same set of optimal hyperparameters. We then report the evaluation metrics computed over the 6 folds.

Results

Table 3 summarizes the results of Drug name recognition in the i2b2-2009 corpus, i2b2-small (reduced version of the former dataset) and the APcNER corpus. For both i2b2-2009 and APcNER, the biGRU-CRF outperforms the terminology-based system, with an exact-match F-measure of 91.1% [95% CI, 90.3-91.9] versus 73% and 81.9% [81.2-82.6] versus 75% respectively. For both i2b2-small and APcNER, the hybrid system outperforms the biGRU-CRF, with an exact-match F-measure of 87.8% [86.4-89.2] versus 85.6% [84.8-86.3] and 88.4% [86.1-86.7] versus 81.9% [81.2-82.6] respectively. The performance on i2b2-small is very close to the performance on APcNER for the hybrid system (with an exact-match F-measure of 87.8% and 86.4%).

Table S4 summarises the results of clinical-NER on all the entity types of the APcNER corpus. For all three systems, the exact-match performance for Sign or symptom, Disease or disorder, Diagnostic procedure or lab test, and Therapeutic procedure are much lower compared to Drug name. The difference between the biGRU-CRF and the terminologies are also more important than for Drug names, with exact-match F-measures of 55.2% versus 15.7% (Sign or symptom), 59.5% versus 30.9% (Disease or disorder), 75.9% versus 30.4% (Diagnostic procedure or lab test) and 61.3% versus 16.6% (Therapeutic procedure). The difference between exact-match and partial match metrics is also greater than for Drug names. As illustrated in Figure 3, the hybrid system outperforms the other systems for all entity types except Therapeutic procedure. Table S2 summarizes the hyperparameters of the models.

Discussion

In this study, we built APcNER, a corpus for clinical-NER of 5 types of entities (Drug names, Signs or symptoms, Diseases or disorders, Diagnostic procedures or lab tests and Therapeutic procedures). We then systematically evaluated a supervised model (biGRU-CRF) against a terminology-based system. Finally, we proposed to extend the supervised system by encoding the predictions of the terminology-based system as categorical embeddings. This hybrid model learns from both inputs: word embeddings carrying semantic and syntactic information from words, and terminology-based embeddings carrying information from external resources. Thus, the powerful modelling capability of the biGRU allows the system to learn from the noisy terminology-based predictions. On the APcNER and on the i2b2-2009 corpora, the biGRU-CRF outperformed the terminology-based system. The hybrid system was more efficient in low regime of entities, for APcNER (except for the Therapeutic procedure class) and i2b2-small.

Both the biGRU-CRF and its extended version outperform previous results from the i2b2 2009 Medication Challenge (90% F-measure for the best team) [21]. These results (mean 92.2 [91.4-92.9]) are very close to FABLE [22] which used *bootstrapping*, a semi-supervised approach, leading to 93% F-measure. As the number of examples increases, the information brought by the terminology should become redundant with the one brought by the annotations, which could explain the relatively larger performance gain of the hybrid system in low regime of trained examples (see Table 3). Note that, as mentioned above, our test set is a sub-sample of the i2b2 test set used during the challenge.

As illustrated in Figure 3, the performance gain of the hybrid system from the purely supervised system appears to be proportional to the performance of the terminology-based system, and inversely proportional to the performance of the supervised system. However, the actual relation is more complex, and the performance of the hybrid system for Signs and Symptoms does not fit with this hypothesis. The number of co-occurrence terms between terminologies could shed light on this outlier, as the Signs and Symptoms category has much more co-occurrence term (15 ‰) than the other entities. Hence, the hybrid system gain for this entity could be boosted by the features from other categories.

The difference in performance for the biGRU-CRF between the i2b2-2009 and the APcNER corpora is partially explained by their difference in number of annotated entities. Indeed, our results (see Table 3) on the reduced version of i2b2 show close performance with the hybrid system when the training set is reduced to the same number of entities than APcNER. Differences remain with the system using no external resources, but this may be due to the fact that the scope covered by the APcNER corpus is much broader in terms of document types and medical specialties. Another noteworthy difference is that a drug name followed by its commercial name between brackets is annotated as a single entity in i2b2-2009, but as several separate entities in APcNER, which explains the difference in the distribution of long n-grams ($n \geq 3$) between the two corpus (see Table 2).

Another notable result is that the performance of the biGRU-CRF is much lower for other types of entities than Drug names, and the difference between exact match and partial match is larger (see Table S4). Along with the results of the APcNER annotation process (Table 2), it suggests that it is partly due to the longer size of the entities. Indeed, compared to Drug names, other categories have between 3 and 6 times more entities composed of at least 2 tokens. The inter-annotator agreement is also lower for these types of entities, and the results of the biGRU-CRF are consistent with those of the inter-annotator agreement. Moreover, it is also consistent with feedback from the annotation process that boundaries are more difficult to define for longer entities. In addition, the conflict between overlapping entities rarely concerns Drug names, whereas they are more likely to occur between Diagnostic and Therapeutic procedure (e.g., angiography), or between Disorder and Sign

(e.g., hemiparesis). Following this analysis, we argue that for the APcNER corpus, the metric of reference should be the non-exact F-measure for entities other than Drug names.

In comparison with MERLOT [9], which include 44,740 entities of 12 types, for 500 documents from Hepato-gastro-enterology and Nutrition ward, APcNER is both smaller (147 documents) and covering a broader scope (no restriction of medical specialty). The inter-annotator agreement for class common to both corpora are comparable, with Signs and Symptoms exact F-measure 59% versus 55%, Drug names 90% versus 85%, and Diseases and disorders 77% versus 65% for MERLOT and APcNER, respectively.

To our knowledge, our study is the first to provide an estimation of the annotation efficiency for clinical-NER for a distribution of medical documents that is representative of that of an EHR (with the exception of imagery reports and drug prescriptions). The annotation efficiency estimated is achieving on average 84% of non-exact match F-measure for the cost of 28 hours of annotations. Using active learning is likely to diminish this cost by 40 to 80% [23,24], hence reaching performance greater than 95% on this task seems possible. In addition, we found consistent results in English and French, which provide an insight into the complementarity of a terminology with a supervised model.

The main limitation of our study is the small size of our corpus compared to the broad scope it covers. However, using cross-validation allowed us to maintain comparable regime of entities with the test set of other corpora. If cross-validation is known to present a risk of overfitting [25], we did not tune the hyperparameters for the hybrid systems, hence the performance gain relative to the biGRU-CRF is a lower bound. Finally, in regards of the average low performance of the supervised model on APcNER (average F-measure of 67.1%), one could think the corpus unfit to allow for supervised learning. However, it still constitutes an important tool to evaluate semi-supervised or unsupervised systems. Indeed, large pre-trained language models such as XLNET [26] could help overcome this low annotation regime by providing better contextualized word representation, which is therefore a direction that we will explore in future research. In addition, combined with a more focused dataset (such as MERLOT [9]), it could allow interesting transfer learning approaches that are to be tested.

Conclusion

APcNER is a French corpus for clinical named-entity recognition of five types of entities which covers a large variety of document types. The extension of the supervised model with terminology has allowed an easy increase in performance, especially for rare entities.

Acknowledgements

We thank Nicolas Griffon for taking part in the annotation process. We thank Christel Daniel for resources and review of the manuscript. We thank Raphaël Veil for proofreading the manuscript.

Authors' contributions

IL contributed to conceptualization; data curation; formal analysis; methodology; software; writing - original draft. NP contributed to conceptualization; data curation; formal analysis; software ; resources; writing - review & editing. XT contributed to conceptualization; data curation; formal analysis; methodology; project administration; resources; supervision; validation; writing - original draft; writing - review & editing.

List of abbreviations

NER: named entity recognition
biGRU-CRF: Bidirectional - Gated Recurrent Unit - Conditional Random Field
AP-HP:Assistance Publique - Hôpitaux de Paris
IOB: inside outside beginning
ATC: Anatomical Therapeutic Chemical Classification System
BPDM: Base publique du médicament
CCAM: Classification commune des actes médicaux
CIM-10: Classification internationale des maladies
DRC: Dictionnaire des Resultats de Consultation
SNOMED: SYSTEMATIZED NOMENCLATURE OF MEDICINE CLINICAL TERMS

Declarations

Availability of data and material: The code made for simulations is available at https://gitlab.com/lerner.ivan/apcner. This project was accepted by the scientific and ethics committee of the AP-HP health data warehouse as CSE-18-0035. The corpus can be made available on condition that a research project is accepted by the scientific and ethics committee of the AP-HP health data warehouse (https://recherche.aphp.fr/eds/recherche/).

Competing interests: The authors declare that they have no competing interests.

Tables

Table 1. UMLS semantic types extracted for each entity

Table 2. APcNER, a French corpus for 5 clinical entities

The F-measure is the harmonic mean between precision and recall computed as in Conll 2003. Inter-annotator agreement is evaluated on a random subset of APcNER of 10 documents.

Table 3. Drug name recognition

Comparison between a terminology-based system, a supervised model (biGRU-CRF) and a hybrid system on: the i2b2-2009 corpus, i2b2-small (a reduced version of the former corpus), and APcNER. We report the mean over 5 runs for exact match and partial match, along with its 95% confidence interval assuming a t-distribution.

Figures

Figure 1. Word representation

The word representation is the concatenation of a word embedding, a character embedding and a terminology-based feature embedding.

Figure 2. biGRU-CRF architecture

Figure 3. The hybrid system gain from biGRU-CRF

We plotted the hybrid system gain on the y-axis, defined as the difference of F-measure between the hybrid system and the biGRU-CRF, as a function as the ratio of the terminology-based system F-measure and the biGRU-CRF F-measure. The color represents the different entity types, and the marker sizes are arbitrary.

References

1. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: A literature review. J Biomed Inform. 2018;77: 34–49.

2. Liu F, Chen J, Jagannatha A, Yu H. Learning for Biomedical Information Extraction: Methodological Review of Recent Advances [Internet]. arXiv [cs.CL]. 2016. Available: http://arxiv.org/abs/1606.07993

3. Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural Architectures for Named Entity Recognition. arXiv [cs.CL]. 2016. Available: http://arxiv.org/abs/1603.01360

4. Zhang S, Elhadad N. Unsupervised biomedical named entity recognition: experiments with clinical and biological texts. J Biomed Inform. 2013;46: 1088–1098.

5. Alicante A, Corazza A, Isgrò F, Silvestri S. Unsupervised entity and relation extraction from clinical records in Italian. Comput Biol Med. 2016;72: 263–275.

6. Del Vigna F, Petrocchi M, Tommasi A, Zavattari C, Tesconi M. Semi-supervised Knowledge Extraction for Detection of Drugs and Their Effects. Social Informatics. Springer International Publishing; 2016. pp. 494–509.

7. Gupta S, Pawar S, Ramrakhiyani N, Palshikar GK, Varma V. Semi-Supervised Recurrent Neural Network for Adverse Drug Reaction mention extraction. BMC Bioinformatics. 2018;19: 212.

8. Fries J, Wu S, Ratner A, Christopher R. SwellShark: A Generative Model for Biomedical Named Entity Recognition without Labeled Data [Internet]. arXiv [cs.CL]. 2017. Available: http://arxiv.org/abs/1704.06360

9. Campillos L, Deléger L, Grouin C, Hamon T, Ligozat A-L, Névéol A. A French clinical corpus with comprehensive semantic annotations: development of the Medical Entity and Relation LIMSI annOtated Text corpus (MERLOT). Language Resources and Evaluation. 2018;52: 571–601.

10. Névéol A, Grosjean J, Darmoni SJ, Zweigenbaum P, Others. Language Resources for French in the Biomedical Domain. LREC. 2014. pp. 2146–2151.

11. Daniel C, Serre P, Orlova N, Bréant S, Paris N, Griffon N. Initializing a hospital-wide data quality program. The AP-HP experience. Comput Methods Programs Biomed. 2018; doi:10.1016/j.cmpb.2018.10.016

12. Uzuner O, Solti I, Xia F, Cadag E. Community annotation experiment for ground truth generation for the i2b2 medication challenge. J Am Med Inform Assoc. 2010;17: 519–523.

13. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology [Internet]. Nucleic Acids Research. 2004. p. 267D–270. doi:10.1093/nar/gkh061

14. Stenetorp P, Pyysalo S, Topić G, Ohta T, Ananiadou S, Tsujii J 'ichi. BRAT: A Web-based Tool for NLP-assisted Text Annotation. Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational

Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics; 2012. pp. 102–107.

15. Lavergne T, Cappé O, Yvon F. Practical Very Large Scale CRFs. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics; 2010. pp. 504–513.

16. Ramshaw LA, Marcus MP. Text Chunking Using Transformation-Based Learning. In: Armstrong S, Church K, Isabelle P, Manzi S, Tzoukermann E, Yarowsky D, editors. Natural Language Processing Using Very Large Corpora. Dordrecht: Springer Netherlands; 1999. pp. 157–176.

17. Manning C, Surdeanu M, Bauer J, Finkel J, Bethard S, McClosky D. The Stanford CoreNLP natural language processing toolkit. Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations. 2014. pp. 55–60.

18. Yang J, Zhang Y. NCRF++: An Open-source Neural Sequence Labeling Toolkit [Internet]. arXiv [cs.CL]. 2018. Available: http://arxiv.org/abs/1806.05626

19. Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space [Internet]. arXiv [cs.CL]. 2013. Available: http://arxiv.org/abs/1301.3781

20. Johnson AEW, Pollard TJ, Shen L, Lehman L-WH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016;3: 160035.

21. Uzuner O, Solti I, Cadag E. Extracting medication information from clinical text. J Am Med Inform Assoc. 2010;17: 514–518.

22. Tao C, Filannino M, Uzuner Ö. FABLE: A Semi-Supervised Prescription Information Extraction System. AMIA Annu Symp Proc. 2018;2018: 1534–1543.

23. Shen D, Zhang J, Su J, Zhou G, Tan C-L. Multi-criteria-based Active Learning for Named Entity Recognition. Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics; 2004. doi:10.3115/1218955.1219030

24. Chen Y, Lasko TA, Mei Q, Denny JC, Xu H. A study of active learning methods for named entity recognition in clinical text. J Biomed Inform. 2015;58: 11–18.

25. Ng AY, Others. Preventing“ overfitting” of cross-validation data. ICML. 1997. pp. 245–253.

26. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le QV. XLNet: Generalized Autoregressive Pretraining for Language Understanding. arXiv [cs.CL]. 2019. Available: http://arxiv.org/abs/1906.08237

terminology or section features

word representation (word2vec)

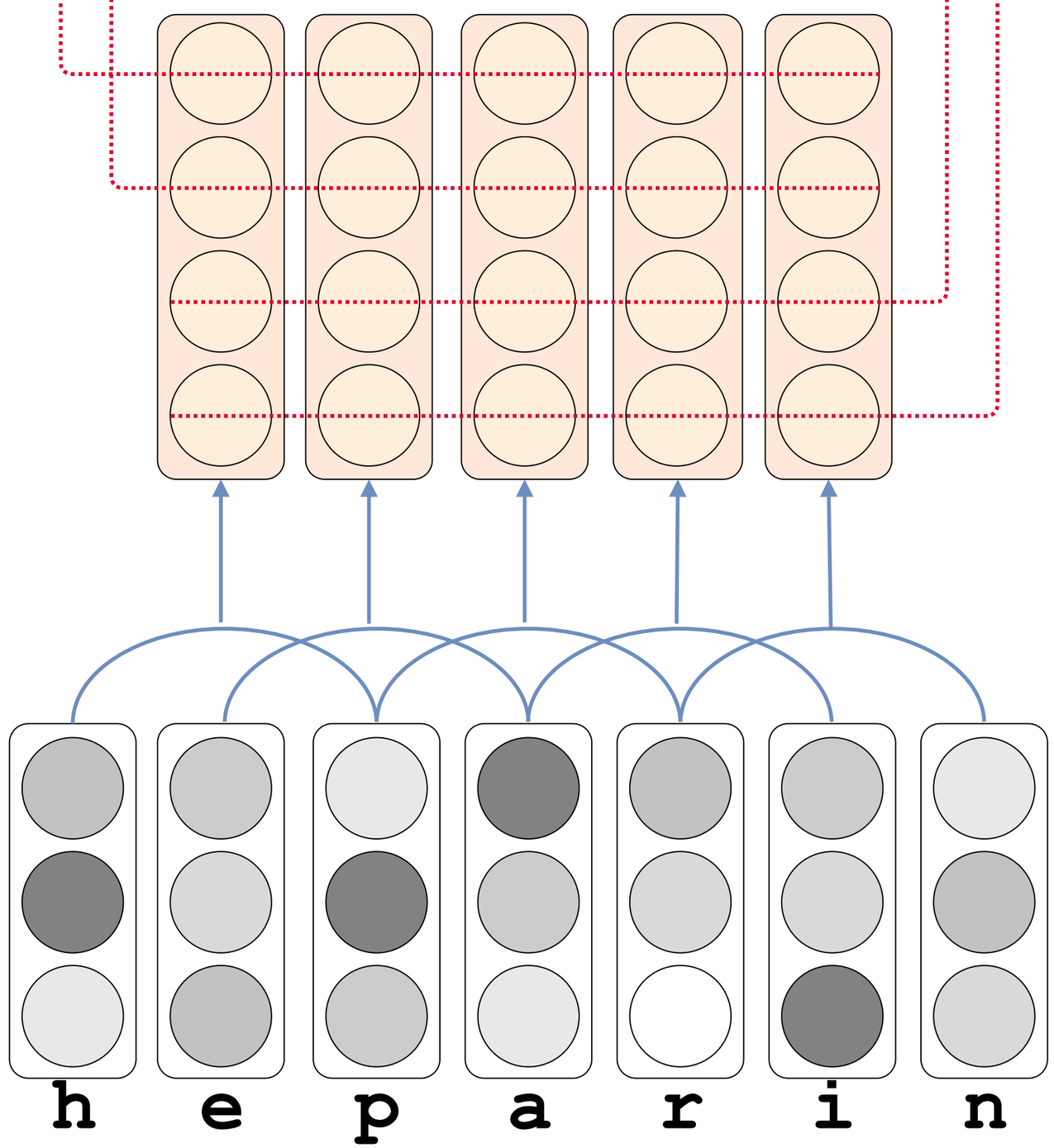
character representation

concatenation

max pooling

convolution

character embeddings



prediction

concatenation
+
projection

backward GRU

forward GRU

word
representation

CRF

bi-GRU

... O O B-DRUG I-DRUG O O O O END

... placed on intraveinous heparin until the following morning

Entity types	Semantic Type	Number of terms (All sources)	Co-occurences of term across type (‰)
Drug name	Antibiotic Clinical Drug Pharmacologic Substance Vitamin	French: 24,932 English: 96,547	French: 4
Sign or symptom	Sign or Symptom	5,125	15
Disease or disorder	Mental or Behavioral Dysfunction Cell or Molecular Dysfunction Anatomical Abnormality Congenital Abnormality Acquired Abnormality Injury or Poisoning Pathologic Function Neoplastic Process Disease or Syndrome	104,104	2
Diagnostic procedure or lab test	Laboratory or Test Result Laboratory Procedure Diagnostic Procedure	16,974	8
Therapeutic procedure	Therapeutic or Preventive Procedure	20,926	1

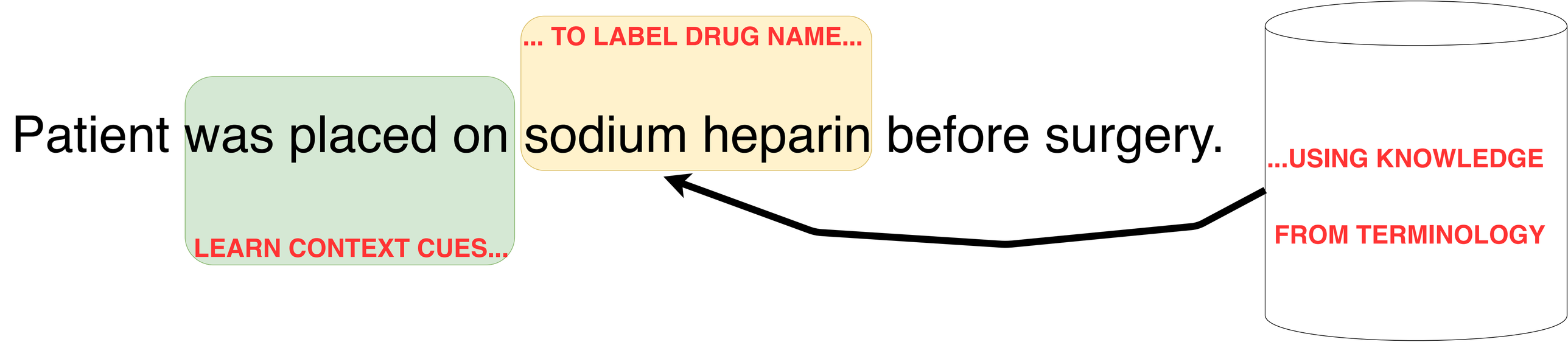
Table 1. UMLS semantic types extracted for each entity

Entity types	Non exact F-measure	Exact F-measure	# entities	n-grams (%)		
				n = 1	n = 2	n ≥ 3
Drug name	.92	.85	1076	1014 (94)	54 (.5)	8 (.1)
Sign or symptom	.71	.55	432	356 (82)	65 (15)	11 (.3)
Disease or disorder	.77	.65	1672	1238 (74)	330 (20)	104 (.6)
Diagnostic procedure or lab test	.87	.70	1156	808 (70)	297 (27)	51 (.4)
Therapeutic procedure	.71	.51	501	414 (83)	73 (15)	14 (.3)

Table 2. APcNER, a french corpus for five clinical entities.

		F-measure [95% CI]	
Corpus	**System**	**Exact-match**	**Partial-match**
i2b2-2009	**Terminologies**	73	84.6
	biGRU-CRF	91.1 [90.3-91.9]	93.5 [92.7-94.3]
	Hybrid system	**92.2** [91.4-92.9]	94.7 [94.1-95.2]
i2b2-small	**biGRU-CRF**	85.6 [84.8-86.3]	90.4 [89.8-91.1]
	Hybrid system	**87.8** [86.4-89.2]	90.6 [86.7-94.6]
APcNER	**Terminologies**	75	77.7
	biGRU-CRF	81.9 [81.2-82.6]	86.4 [85.3-87.6]
	Hybrid system	**86.4** [86.1-86.7]	90.4 [89.8-90.9]

Table 3. Drug name recognition.


... TO LABEL DRUG NAME...
Patient was placed on sodium heparin before surgery.
LEARN CONTEXT CUES...
...USING KNOWLEDGE
FROM TERMINOLOGY