Bootstrapping a Romanian Corpus for Medical Named Entity Recognition

Maria Mitrofan
Research Institute for Artificial Intelligence "Mihai Drăgănescu"
Romanian Academy
Calea 13 Septembrie, nr. 13,
Bucharest, România
maria@racai.ro

Abstract

Named Entity Recognition (NER) is an important component of natural language processing (NLP), with applicability in the biomedical domain, enabling knowledge discovery from medical texts. Due to the fact that for the Romanian language there are only a few linguistic resources specific to the biomedical domain, we have created a sub-corpus specific to this domain. In this paper we present a newly developed Romanian sub-corpus for medical domain NER, which is a valuable asset for the field of biomedical text processing. We provide a description of the sub-corpus, statistics about data-composition and we evaluate an automatic NER tool on the newly created resource.

1 Introduction

There is an increasing need for exploiting and managing the available biomedical texts due to the fact that each day huge amounts of medical data become available (Patel et al., 2009).

MEDLINE, the largest biomedical database resource, currently contains more than 26.9 million abstracts of the world's biomedical journal literature and each month 60,000 new abstracts are added, according to MEDLINE Database Summary Sheet (DBSS). The increasing rate of published biomedical literature has generated a pressing need for computation techniques to be used for information extraction from the available data (Coleman et al., 2009; Gabbay and Le May, 2010).

In general, most of the available data is noisy and/or unstructured as for instance in clinical reports. Consequently, NLP tools are required and used to turn this data into knowledge.

In the NLP domain NER is the task dedicated to the identification and classification of textual units, be they single words or multiple words (such as locations, names of persons, organizations, places).

NER systems are a prerequisite for many text processing applications such as relation extraction (Tasneem and Archana, 2016), question answering (Athenikos and Han, 2009), information extraction (Piskorski and Yangarber, 2012), etc. In fact, NER is a basic step in ordering and structuring all the existing domain information.

In particular, biomedical named entity recognition (BioNER) tools aim to detect biomedical terms such as human anatomical parts (Xu et al., 2014), drug names (Liu et al., 2015), gene and protein mentions (Tanabe and Wilbur, 2002), chemical compounds (Eltyeb and Salim, 2014), diseases (Jimeno et al., 2008) and to assign them the correct categories.

Although the NLP community has invested a lot of efforts in BioNER, the task is complex, because biomedical corpora contain specialized terminology, which is not easy to identify. Nevertheless, it is argued that the vocabulary in biomedical corpora is easier to deal with than the vocabulary in general corpora, due to the closure properties of sublanguages (Temnikova et al., 2013; Temnikova and Cohen, 2013).

Moreover NER systems trained and tested on news articles corpora achieve on average an accuracy of 90% (Passos et al., 2014), but similar techniques do not work well when applied to biomedical corpora, the accuracy obtained being about 10% less (Abacha and Zweigenbaum, 2011).

In this paper, we explore NLP techniques to identify biomedical named entities in text and also we present up-to-date statistics about a newly created Romanian medical sub-corpus.

Proceedings of Recent Advances in Natural Language Processing, pages 501–509,
Varna, Bulgaria, Sep 4–6 2017.
https://doi.org/10.26615/978-954-452-049-6_066

2 Challenges in BioNER

To minimize the gap mentioned before between performances of biomedical NER and other types of NER several techniques and algorithms have been proposed taking into consideration the peculiarities of biomedical texts.

Due to the fact that in biomedical literature there is not a unique naming convention, the spelling variations of the biomedical terms cause recognition ambiguity. For example, the same ”diabetes mellitus type 2” entity may be referred to in Romanian in different spelling forms: “T2DM” borrowed abbreviation from English, “DZ tip 2”(En. DM type 2) Romanian abbreviation, “diabet zaharat tip 2”(En. type 2 diabetes mellitus) the full Romanian form. Synonymy is a frequent linguistic feature of the biomedical subcorpus. For exmpample, the terms ”natriu” (En. sodium) and ”sodiu” (En. sodium) have the same meaning.

The phenomenon of polysemy is also present in Romanian biomedical text, for example for the Romanian abbreviation ”PA” there are two possible meanings: ”presiune arterială” (En. blood presure) and ”forfatază alcalină” (En. alkaline phosphatase).

And also there are no rules for the formation of biomedical terms and words may contain digits (T1DM, T2DM), Greek letters ”celula β” (En. β- cell), ”celule β-pancreatice” (En. pancreatic β- cells), hyphens ”19-nortestosteron” (En. ”19-nortestosterone”).

Another frequent problem is that biomedical literature is very rich in abbreviations. Many abbreviations are difficult to correctly classify because of their multiple forms. For example ”electrocardiogramă” (En. electrocardiogram) has two abbreviation forms ”ECG” and ”EKG” or ”fibrilație atrială” (En. atrial fibrillation) can be abbreviated as ”FA” or ”FiA”.

Chang et al. (2002) have shown that in every 5-10 MEDLINE abstracts there is one new abbreviation and Liu et al. (2002) showed that 81.2% of abbreviations found in MEDLINE abstracts are ambiguous. Moreover, new substances are discovered daily and this causes difficulties in recognizing them, especially for rule based systems.

Furthermore, another BioNER challenge is generated by the fact that one head noun may be shared by two or more biomedical named entities. For example, the following structure with coordination “micro- și macroangiopatiei” (En. micro- and macroangiopathy) consists of two entityes “microangiopatiei” (En. microangiophaty) and “macroangiopatiei” (En. macroangiopathy), the same case with “ateroscleroza aortei și a vaselor periferice” (En. atherosclerosis of the aorta and peripheral vessels), which should be read as ”ateroscleroza aortei și ateroscleroza vaselor periferice” (En. atherosclerosis of the aorta and atherosclerosis of the peripheral vessels).

As a particular type of coordination disjunctions also allow omission for the head noun in the second conjunct: ”celule beta pancreatice sau hepatice” (En. pancreatic beta or hepatic cells) should be interpreted as ”celule beta pancreatice sau celule hepatice” (En. pancreatic beta cells or hepatic cells).

Cascaded constructions represent another major challenge that can be encountered in BioNER, because one entity may be incorporated in another entity name. In GENIA V3.0 corpus almost 16.57% (Zhou and Su, 2004) of all biomedical entity names have cascaded construction (Sondhi, 2008). For example, for the Romanian language we may find cascaded constructions such as ”Anevrismele/B-DISO pot fi fusiforme/I-DISO (aspect cilindric al vasului/B-ANAT sangvin/I-ANAT) sau sacciforme/I-DISO.” (En. Aneurysms/B-DISO may be fusiforms/I-DISO (cylindrical appearance of the blood/B-ANAT vessel/I-ANAT) or sacciforms/I-DISO.) (see subsection 5.2).

Even though nowadays there are language independent BioNER systems, most of them rely on linguistic resources, which are not available for all languages and domains (Nadeau and Sekine, 2007), thus when language adaptation is needed the performance of BioNER systems is affected.

Consequently BioNER is much more complex than general named entity recognition applied in newswire domain (Sondhi, 2008).

3 Related Work

3.1 Biomedical Corpora

For English, there are multiple biomedical corpora that can be used for different NLP tasks. Since the release of the GENIA corpus (Kim et al., 2003) and thanks to the availability of annotated biomedical corpora (GENETAG corpus (Tanabe et al., 2005), SCAI IUPAC corpus (Kolarik et al., 2008), AnEM corpus (Ohta et al., 2012), and CellFinder corpus (Neves et al., 2012)), various systems have

been developed for information extraction from biomedical documents. Nowadays such systems can find diseases, drug names, clinical problems and gene names with performance (F score) better than 90% (Abacha and Zweigenbaum, 2011; Wang and Patrick, 2009; Boytcheva et al., 2010).

On the other hand, research on medical languages other than English is more scarce.

For the French language, the "Unified Medical Lexicon for French" (UMLF) (Zweigenbaum et al., 2005) has been created and aims at being a reference resource for NLP in the medical domain. Nevertheless, (Cartoni and Zweigenbaum, 2010) showed that even in large collections of terms there is a lack of specialized lexicons and they conducted an experiment to feed a French medical lexicon, in which the dimension of the specialized lexicon increased its coverage of the initial vocabulary from 14.1% to 25.7%.

For Swedish an annotated gold standard corpus of medical records was developed (Velupillai, 2012) and also scientific medical corpus was created for linguistic exploration and terminology management. Mowery et al. (2012) proposed a clinical uncertainty and negation taxonomy and mapped an English annotation schema to a Swedish schema. Recently a corpus for BioNER recognition in Spanish have been created (Moreno et al., 2017).

For Bulgarian language important efforts have been made in collecting biomedical literature usable for NLP tasks. For example (Boytcheva et al., 2009) described a Bulgarian medical corpus formed by 6400 words, with 2000 of them belonging to Bulgarian medical terminology. Nikolova et al. (2016) used free textual data of diabetic patience to determine their smoking status.

3.2 BioNER Approaches

To tackle the challenges posed by BioNER, researchers use different NER approaches including: dictionary-based methods, rule-based methods and machine learning methods.

Terminology-driven BioNER methods such as dictionary and rule-based approaches, use regular expressions to match the information from terminological resources with text phrases. Fukuda et al. (1998) proposed a rule-based system for protein names identification and obtained a precision of 91.90% and a recall of 93.32%, when the system was evaluated on 30 annotated MEDLINE abstracts. Gaizauskas et al. (2000) used used terminology lexicons, standard biomedical suffixes and hand-designed grammar rules for terminology classes and achieved 86% precision and 68% recall. Nevertheless NER systems based on rules perform poorly for large scale tasks because of the spelling variations and different naming conventions of biomedical terms (Gaizauskas et al., 2000; Fukuda et al., 1998; Tuason et al., 2004).

Machine learning (ML) based systems are focused on the recognition of specific named entities using various statistical models. In machine learning area there are taken two main approaches. The former one is based on supervised learning techniques, where based on a learning algorithm a mapping from a known input to a desired output is performed.

The latter broad machine learning approach used for BioNER is unsupervised learning and the aim of this method is to find regularities in the data, based only on input data. The methods of unsupervised learning are mostly built upon clustering techniques, similarity based functions and statistics. Recently, there has been an increasing interest in using word embeddings from unlabeled biomedical corpora (Li et al., 2016).

4 Corpora Description and Annotation Tools

Even though at the international level the challenges of biomedical information processing have changed from where to collect resources to how to make use of them (Shaodian and Elhadad, 2013), at the national level linguistic resources specific to certain domains (biomedical area among them) are difficult to obtain. However, a relevant sub-corpus for biomedical domain has been collected in the context of the CoRoLa project (The reference corpus of the contemporary Romanian language created by the Romanian Academy Research Institute for Artificial Intelligence "Mihai Drăgănescu" and Institute for Computer Science in Iași) (Tufis et al., 2016).

The Romanian biomedical sub-corpus is composed of about 7 million tokens (including punctuation), about 300,000 sentences extracted from different biomedical sub-domains such as: diabetes, cardiology, endocrinology, neurology, oncology, etc. (Mitrofan and Tufiș, 2016) (Table 1).

# tokens	7,173,396
# words	6,287,246
# unique lemmas	136,330
# sentences	309,948
average tokens per sentence	23.14
average words per sentence	20.28
average punctuation per sentence	2.8

Table 1: Statistics over the Romanian medical sub-corpus.

4.1 Pre-processing Steps

NLP solutions are usually decomposed into subtasks that form processing pipelines that ensure specific functionalities such as: sentence splitting, tokenization, lemmatization and chunking, part-of-speech (POS) tagging, parsing.

In order to process the Romanian medical subcorpus we used the TTL platform (Ion, 2007), which is a language-independent text processing module (Todirașcu et al., 2011). Another processing tool for Romanian is the Modular Language Processing for Lightweight Applications (MLPLA) (Dumitrescu et al., 2017), which is a freely available[1] and language-independent processing tool that supports more than 50 languages.

The TTL tool is able to automatically perform specific functionalities (Tufiș et al., 2010) such as: sentence splitting (to identify the end of a sentence it uses regular expressions), tokenization, part-of-speech tagging (with an accuracy of more than 98%, when trained on newswire domain), lemmatization (it recovers for each word the corresponding lemma based on a human-validated Romanian word-form lexicon, the lemma guesser model has an accuracy of 83%), chunking (based on a set of regular expressions for each tagged and lemmatized lexical unit is assigned a syntactic phrase.

After running TTL on the biomedical subcorpus, about 7 million tokens were assigned a corresponding lemma and a POS tag. Table 6 shows the results after the POS-tagging step. We want to emphasize that most of the B-ANAT, B-DISO, B-PROC, B-CHEM named entity classes tagged as adjectives are in fact POS-tagging errors. This also happens for the category "Others" where nouns can be found tagged as verbs, adverbs, etc.

The TTL tagger marks the unknown words for which the tags and lemmas were predicted on the basis of the language model. This makes it easier to spot wrong predictions (tag, lemma or both) and correct them manually by a linguist. The bootstrapping method we adopted takes advantage of these corrections. It was shown that lexical features, especially part-of-speech tags, are important for BioNER as they may help to identify entity boundaries (Sondhi, 2008). Zhou and Su (2004) reported an increase in performance when part-of-speech features were integrated.

5 The Annotation Process

The first step, in order to apply NER techniques to the medical sub-corpus, was to manually annotate almost 40,000 tokens with BioNER tags and have all these labels checked by a medical expert, who was accustomed to the IOB standard.

Secondly to rapidly grow our sub-corpus used for BioNER we followed a typical bootstrapping procedure, in which, once a sub-portion of the subcorpus is available, a ML technique is used to learn how to automatically detect and label NEs in the unprocessed sections of the data. This way the manual annotation procedure is enhanced for the remainder corpora, because automatically inferred labels offer good guidelines and greatly speed-up the process. Therefore after the bootstrapping procedure other 60.000 tokens were automatically labeled with BioNER tags and then each one of them was corrected by hand.

5.1 Entity Classes

For the Romanian biomedical sub-corpus four top level entity classes were chosen Anatomy (anatomical structure, body part, organ, organ component,tissue, cell, cell component), Chemicals and Drugs (amino acid, peptide, protein, antibiotic, biologically active substance, chemical, clinical drug, enzyme, hormone, pharmacological substance, receptor), Disorders (anatomical abnormality, acquired abnormality, congenital abnormality, disease or syndrome, injury, mental dysfunction), Procedures (laboratory procedure, therapeutic or preventive procedure), defined by choosing the corresponding UMLS (Unified Medical Language System) semantic groups[2] :

- Anatomy (ANAT): "valvă aortică" (En. aortic valve), "stomac" (En. stomach), "țesut

[1] http://slp.racai.ro/index.php/mlpla-new/

[2] https://metamap.nlm.nih.gov/Docs/SemGroups_2013.txt - accessed 2017-05-04

epitelial" (En. epithelial tissue), "mitocondrie" (En. mitochondria);

- Chemicals and Drugs (CHEM): "penicilină" (En. penicillin), "acetilcolină" (En. acetylcholine), "lipază" (En. lipase);

- Disorders (DISO): "depresie" (En. depression), "delir" (En. delirium), "accident vascular cerebral" (En. stroke), "diabet zaharat" (En. mellitus diabetes);

- Procedures (PROC): "ecocardiografie transesofagiană" (En. transesophageal echocardiography), "radiografie" (En. radiography).

5.2 IOB Format Tagging

In order to apply language processing algorithms to BioNER, we converted the sub-corpus into IOB2 format (Sang and Veenstra, 1999), where "B" denotes the beginning chunk (a span of tokens) and "I" represents an inside chunk. "O" labels indicate tokens that do not belong to a chunk. Table 2 shows an example of a tagged sentence: "Examenul obiectiv al cordului identifică adesea tulburări de ritm, cele mai frecvente fiind fibrilația atrială și aritmia extrasistolică." (En. The objective examination of the heart often identifies rhythm perturbations, the most common being the atrial fibrillation and the extrasystolic arrhythmia.).

Token	Tag
Examenul (The examination)	O
obiectiv (objective)	O
al (of)	O
cordului (heart)	B-ANAT
identifică (identifies)	O
adesea (often)	O
tulburări (perturbations)	B-DISO
de (of)	I-DISO
ritm (rhythm)	I-DISO
,	O
cele (the)	O
mai (most)	O
frecvente (frequent)	O
fiind (being)	O
fibrilația (the fibrillation)	B-DISO
atrială (atrial)	I-DISO
și (and)	O
aritmia (the arrhythmia)	B-DISO
extrasistolică (extrasistolic)	I-DISO
.	O

Table 2: Example of a tagged sentence

6 Corpus and Automatic Biomedical NER Evaluation

At the time we are writing this paper, the annotation of the sub-corpus is on-going in parallel with enlarging its size. However, we consider that the available data has reached maturity, in the sense that it can already find its use in the field of research. In what follows we provide relevant statistical information about our corpora composition such as: (a) the distributions of the named entities based on their type; (b) the average length and standard deviation of named entities (also based on their types); (c) distribution of underlying part-of-speech type for each NE type and (d) the results obtained by our pretrained NE models.

For clarity, all information regarding the corpus is rendered in subsection 6.1, while subsection 6.2 deals with the process of training and testing our automatic NE technique based on the newly created sub-corpus.

This section is oriented toward providing preliminary information about the sub-corpus and before we proceed with, we will motivate the statistics we extracted.

6.1 Corpus Statistics

- **NE type distribution**: this information is very helpful for establishing if the sub-corpus is well-balanced and what the expected results will be if one trains an automatic NE identification tool on the available data (Table 3).

Tag	Number of tags
B-DISO	3992
I-DISO	2942
B-ANAT	1387
I-ANAT	996
B-PROC	947
I-PROC	714
B-CHEM	2525
I-CHEM	816

Table 3: NE type distribution.

- **Average size (in tokens) of NEs:** knowing what is the average span of a NE is impor-

tant in the feature-selection process. As such, compact NEs (short and without interleaved non-NE tokens) make it possible to use small context windows in the feature extraction process, while long-range NEs (with interleaved non-NE tokens) require other approaches (in practice modified SHIFT-REDUCE schemes can achieve good results) (Table 5). Table 4 shows that most of the medical NEs are compound of more than one token, as can be seen also in table 5. "CHEM" is the entity class that contains the shortest NEs, 75% of NEs are compound of only one token, and the NEs with length greater than three tokens appear seldom, as can be seen from both tables 5 and 4.

NE	NE length				
	1	2	3	4	5
B-DISO	48%	35%	12%	3%	2%
B-ANAT	43%	42%	12%	2%	1%
B-PROC	40%	47%	10%	2%	1%
B-CHEM	75%	20%	4%	1%	0%

Table 4: NE type length.

Tag	Average	Stdev.
DISO	1.747	0.951
ANAT	1.723	0.743
PROC	1.762	0.177
CHEM	1.329	0.656
Overall	1.626	0.846

Table 5: Average size of NEs.

- **POS statistics:** provide good clues whether one should or should not use the POS information as features for training a automatic NE tool. In our case,it would be expected that most tokens would be nouns, adjectives and abbreviations (Table 6).

6.2 Automatic NE for Biomedical Sub-corpus

As can easily be seen our NEs are mostly compact with a POS distribution that motivates using the grammatical category as a feature. This, combined with the average length of our NEs has driven us to go for a straight-forward NE identification procedure: we trained a classifier to label each token inside a sentence with a IOB tag, based on features extracted from the context windows.

In our approach, the context-window size is 3 (centered on the current token) and the features are composed of the word-form and POS information for each context-word. A particularity is that, instead of using standard approaches (CRF, SVM, Decision Tree etc.) we employed a Partitioned Convolutional Neural Network for classification and we used automatically extracted word-embeddings (Mikolov et al., 2013), computed using Word2Vec[3] from a corpus composed of the Romanian section of Wikipedia, concatenated with our own medical sub-corpus. The architecture of the network is composed of two partitions followed by two fully connected layers and a softmax output layer. Each partition is trained independently on its own feature category:

- The wordform partition works directly over the word embeddings inside the receptive field (window size of 3) and is based on 128 convolutional filters (size 1x64 - a word embeddings size of 64);

- The POS partition is trained on automatically inferred feature embeddings, that feed into 16 convolutional filters. The automatic feature-embeddings process is inspired by (Danqi and Christopher, 2014) and is implemented as a set of deconvolutional filters (one filter for each possible POS label).

In order to evaluate our approach we used 80% of the data for training, 10% for development, and 10% for testing. Table 7 summarizes the results obtained on the test-set: column 2 (ident.) refers to the number of correctly identified instances of the corresponding label and column 3 (act.) represents the actual number of instances in the test-set.

7 The Availability of the Data

The biomedical subcorpus will be available in the context of the CoRoLa project copyright agreement signed with the editorial offices representatives and with the publishing houses. All the data from the CoRoLa will be available for the public through KorAP platform (Bingel et al., 2013). This platform allows various linguistic types of searches in the data, but the corpus will not be downloadable. However, all the results of the interrogation of the corpus outside the scope of the copyright restrictions will be downloadable.

[3]https://github.com/dav/word2vec - accessed 2017-05-03

Tag	Nouns	Adjectives	Abbreviations	Others
B-DISO	3634	19	285	54
I-DISO	418	2263	10	251
B-ANAT	1352	19	16	0
I-ANAT	150	788	19	39
B-PROC	907	10	20	10
I-PROC	160	491	15	48
B-CHEM	2195	125	179	26
I-CHEM	248	410	49	109

Table 6: POS-statistics.

Tag	Ident.	Act.	Precision	Recall	F-score
Dev-set					
B-ANAT	63	178	0.67	0.35	0.46
I-ANAT	56	171	0.74	0.32	0.45
B-DISO	208	409	0.64	0.50	0.56
I-DISO	150	341	0.68	0.43	0.53
B-PROC	4	166	0.80	0.02	0.04
I-PROC	13	156	0.81	0.08	0.15
B-CHEM	46	107	0.29	0.42	0.34
I-CHEM	5	26	0.18	0.19	0.18
Test-set					
B-ANAT	52	136	0.75	0.38	0.50
I-ANAT	31	104	0.77	0.29	0.43
B-DISO	162	387	0.61	0.41	0.49
I-DISO	137	297	0.68	0.46	0.55
B-PROC	23	53	0.51	0.43	0.46
I-PROC	17	34	0.47	0.50	0.48
B-CHEM	81	189	0.42	0.42	0.42
I-CHEM	13	74	0.48	0.17	0.25

Table 7: Evaluation results on the development and test sets.

8 Conclusions and Future Work

In this paper we introduced a newly created text sub-corpus aimed at proving support for NLP on biomedical text. We provided relevant information about the sub-corpus itself (at token/NE level), we described our annotation process (both automatic: tokenization, lemmatization and part-of-speech tagging – and manual: the NE labeling procedure).

Additionally, we assessed the validity and maturity of our data by introducing a custom-designed ML method for identifying NEs in the biomedical domain.

Currently our corpus is still under development, but we consider that the available data and the pre-trained tool can already be used on Romanian biomedical text.

The annotated section of the corpus is freely available for download[4] and non-commercial use. Special use-cases require license permissions from the author.

Acknowledgements

We want to thank to my colleague Tiberiu Boroș for helping us with technical issues whenever they have arisen and for the comments that greatly improved the quality of the paper. Moreover we wish to acknowledge the help provided by the PhD student Grigorina Mitrofan. Also Verginica Barbu Mititelu and Elena Irimia, thank you for giving us valuable comments on the draft version of this paper.

[4] http://slp.racai.ro/index.php/resources/

References

A. Abacha and P. Zweigenbaum. 2011. Medical entity recognition: a comparison of semantic and statistical methods. In *Proceedings of BioNLP 2011 workshop*. pages 56–64.

Sofia J. Athenikos and Hyoil Han. 2009. Biomedical question answering: A survey.

P. Banskiand J. Bingel, N.Diewald, E. Frick, M. Hanl, M. Kupietz, P. Pezik, C. Schnober, and A. Witt. 2013. The new corpus analysis platform at ids mannheim.

S. Boytcheva, I. Nikolova, E. Paskaleva, G. Angelova, D. Tcharaktchiev, and N. Dimitrova. 2009. Extraction and exploration of correlations in patient status data. in proceedings of the workshop on biomedical information extraction. In *Proceedings of the Workshop on Biomedical Information Extraction*. pages 1–7.

S. Boytcheva, I. Nikolova, E. Paskaleva, G.Angelova, D. Tcharaktchiev, and N. Dimitrova. 2010. Obtaining status descriptions via automatic analysis of hospital patient records.

B. Cartoni and P. Zweigenbaum. 2010. Semi-automated extension of a specialized medical lexicon for french. In *Proceedings of LREC*.

J. Chang, H. Schutze, and R. Altman. 2002. Creating an online dictionary of abbreviations from medline.

K. Coleman, BT. Austin, and C. Brach andEH. Wagner. 2009. *Evidence on the chronic care model in new millenium.*. Millrood.

Chen Danqi and D. Manning Christopher. 2014. A fast and accurate dependency parser using neural networks. In *EMNLP*. pages 740–750.

Stefan Daniel Dumitrescu, Tiberiu Boroş, and Dan Tufiş. 2017. Racai's natural language processing pipeline for universal dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Association for Computational Linguistics, Vancouver, Canada, pages 174–181. http://www.aclweb.org/anthology/K/K17/K17-3018.pdf.

Safaa Eltyeb and Naomie Salim. 2014. Chemical named entities recognition: a review on approaches and applications.

K. Fukuda, T. Tsunoda, A. Tamura, and T. Takagi. 1998. Toward information extraction: identifying protein names from biological papers. In *Proceedings of the Pacific Symposium on Biocomputing.*. pages 707–718.

John Gabbay and Andrée Le May. 2010. *Practice-based evidence for healthcare: clinical mindlines*. Routledge.

R. Gaizauskas, G. Demetriou, and K. Humphreys. 2000. Term recognition and classification in biological science journal articles. In *Proceedings of Workshop on Computational Terminology for Medical and Biological Applications.*. pages 37–44.

Radu Ion. 2007. *Word Sense Disambiguation Methods Applied to English and Romanian (in Romanian)*.

Antonio Jimeno, Ernesto Jimenez-Ruiz, Vivian Lee, Sylvain Gaudan, Rafael Berlanga, and Dietrich Rebholz-Schuhmann. 2008. Assessment of disease named entity recognition on a corpus of annotated sentences.

J.D. Kim, T. Ohta, Y. Tateisi, and J. I. Tsujii. 2003. Genia corpus—a semantically annotated corpus for bio-textmining.

Corinna Kolarik, Roman Klinger, Christoph M Friedrich, Martin Hofmann-Apitius, and Juliane Fluck. 2008. Chemical names: terminological resources and corpora annotation. In *Workshop on Building and evaluating resources for biomedical text mining (6th edition of the Language Resources and Evaluation Conference)*.

L. Li, L. Jin, Y. Jiang, and D.Huang. 2016. Recognizing biomedical named entities based on the sentence vector/twin word embeddings conditioned bidirectional lstm.

H. Liu, A. Aronson, and C. Friedman. 2002. A study of abbreviations in medline abstracts. In *Proceedings of the American Medical Informatics Association Symposium 2002*. page 327–332.

Shengyu Liu, Buzhou Tang, Qingcai Chen, Xiaolong Wang, and Xiaoming Fan. 2015. Feature engineering for drug name recognition in biomedical texts: Feature conjunction and feature selection.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. pages 3111–3119.

Maria Mitrofan and Dan Tufiș. 2016. Building and evaluating the romanian medical corpus. In *Proceedings ofthe 12 th International Conference "Linguistic Resources and tools for processing the Romanian language"*. pages 29–36.

I. Moreno, E. Boldrini, P. Moreda, and M. T. Roma-Ferri. 2017. Drugsemantics: a corpus for named entity recognition in spanish summaries of product characteristics.

D. L. Mowery, S. Velupillai, and W. W. Chapman. 2012. Medical diagnosis lost in translation: analysis of uncertainty and negation expressions in english and swedish clinical texts. In *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing. Association for Computational Linguistics*.

D. Nadeau and S. Sekine. 2007. A survey of named entity recognition and classification.

Mariana Neves, Alexander Damaschun, Andreas Kurtz, and Ulf Leser. 2012. Annotating and evaluating text for stem cell research. In *Proceedings of the Third Workshop on Building and Evaluation Resources for Biomedical Text Mining (BioTxtM 2012)*.

I. Nikolova, S. Boytcheva, G. Angelova, and Z. Angelov. 2016. Combining structured and free textual data of diabetic patients' smoking status. in international conference on artificial intelligence: Methodology, systems, and applications. In *Proceedings of the International Conference on Artificial Intelligence: Methodology, Systems, and Applications*. Springer, pages 57–67.

Tomoko Ohta, Sampo Pyysalo, Jun'ichi Tsujii, and Sophia Ananiadou. 2012. Open-domain anatomical entity mention detection. In *Proceedings of ACL 2012 Workshop on Detecting Structure in Scholarly Discourse (DSSD)*. pages 27–36.

Alexandre Passos, Vineet Kumar, and Andrew McCallum. 2014. Lexicon infused phrase embeddings for named entity resolution. In *Proceedings of CoNLL.*.

V.L. Patel, E.H. Shortliffe, M. Stefanelli, P. Szolovits, M.R. Berthold, and R. Bellazzi. 2009. The coming of age of artificial intelligence in medicine.

Jakub Piskorski and Roman Yangarber. 2012. Information extraction: Past, present and future.

E. F. Sang and Jorn Veenstra. 1999. Representing text chunks. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics.*.

Zhang Shaodian and Noémie Elhadad. 2013. Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts.

Parikshit Sondhi. 2008. A survey on named entity extraction in the biomedical domain.

Lorraine Tanabe and John Wilbur. 2002. Tagging gene and protein names in biomedical text.

Lorraine Tanabe, N. Xie, LH. Thom, W. Matten, and WJ. Wilbur. 2005. A tagged corpus for gene/protein named entity recognition.

Almas Tasneem and B. Archana. 2016. A survey on biomedical named entity extraction.

I. P. Temnikova and K. B. Cohen. 2013. Recognizing sublanguages in scientific journal articles through closure properties. In *Proceedings of BioNLP*. pages 72–79.

I. P. Temnikova, I. Nikolova, W.A. Baumgartner Jr, G. Angelova, and K. B. Cohen. 2013. Closure properties of bulgarian clinical text. In *Proceedings of RANLP 2013*.

Amalia Todirașcu, Radu Ion, Mirabela Navlea, and Laurence Longo. 2011. French text preprocessing with ttl. In *PROCEEDINGS OF THE ROMANIAN ACADEMY*. page 151–158.

O. Tuason, L. Chen, H. Liu, J.A. Blake, and C. Friedman. 2004. Biological nomenclature: A source of lexical knowledge and ambiguity. In *Proceedings of Pac Symp Biocomput.*. pages 238–249.

D. Tufis, V. B. Mititelu, E. Irimia, S. D. Dumitrescu, and T. Boros. 2016. The ipr-cleared corpus of contemporary written and spoken romanian language. In *Proceedings of the International Conference on Language Resources and Evaluation - LREC*.

Dan Tufiș, Radu Ion, Alexandru Ceaușu, and Dan Ştefănescu. 2010. Reifying the alignments. accurat-project.eu.

S. Velupillai. 2012. *Shades of certainty: annotation and classification of swedish medical records.*.

Y. Wang and J. Patrick. 2009. Cascading classifiers for named entity recognition in clinical notes. In *Proceedings of the workshop on biomedical information extraction.*. pages 42–49.

Yan Xu, Ji Hua, Zhaoheng Ni, Qinlang Chen, Yubo Fan, Sophia Ananiadou, Eric I-Chao Chang, and Junichi Tsujii. 2014. Anatomical entity recognition with a hierarchical framework augmented by external resources.

G. Zhou and J. Su. 2004. Exploring deep knowledge resources in biomedical name recognition. In *Proceedings of the joint workshop on natural language processing in biomedicine and its applications*. pages 96–99.

P. Zweigenbaum, R. Baud, A. Burgun, F. Namer, É Jarrousse, N. Grabar, and S. Darmoni. 2005. Umlf: a unified medical lexicon for french.