Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 160 (2019) 260–265

Procedia Computer Science

www.elsevier.com/locate/procedia

The 10th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN 2019)
November 4-7, 2019, Coimbra, Portugal

Study of Named Entity Recognition methods in biomedical field

Anna Śniegula[a,b], Aneta Poniszewska-Marańda[a,*], Łukasz Chomątek[a]

[a]*Institute of Information Technology, Lodz University of Technology, Łódź, Poland*
[b]*Department of Computer Science in Economics, Institute of Applied Economics and Informatics, Faculty of Economics and Sociology, University of Lodz, Łódź, Poland*

Abstract

Natural Language Processing (NLP) is very important in modern data processing taking into consideration different sources, forms and purpose of data as well as information in different areas our industry, administration, public and private life. Our studies concern Natural Language Processing techniques in biomedical field. The increasing volume of information stored in medical health record databases both in natural language and in structured forms is creating increasing challenges for information retrieval (IR) technologies. The paper presents the comparison study of chosen Named Entity Recognition techniques for biomedical field.

Keywords: Machine learning; Natural Language Processing; recurrent neural networks; Named Entity Recognition; Conditional Random Fields; UMLS; Long-Short Term Memory.

1. Introduction

Natural Language Processing (NLP) is very important in modern data processing taking into consideration different sources, forms and purpose of data as well as information in different areas our industry, administration, public and private life. Our studies concern Natural Language Processing techniques in biomedical field. The increasing volume of information stored in medical health record databases both in natural language and in structured forms is creating increasing challenges for information retrieval (IR) technologies.

Clinical data are often stored in natural language form. Narrative language is convenient for doctors to express events and medical concepts, unfortunately it makes the data difficult for searching, summarisation, decision support or statistical analysis.

* Corresponding author. Tel.: +48 42 632 97 57.
E-mail address: aneta.poniszewska-maranda@p.lodz.pl

10.1016/j.procs.2019.09.466

In order to perform above tasks the information has to be extracted with various Natural Language Processing (NLP) techniques [12]. Named Entity Recognition (NER) is a fundamental Natural Language Processing task to extract the entities of interest (e.g., disease names, medication names and lab tests) from clinical narratives, thus to support clinical and translational research.

The presented study aims to compare various Named Entity Recognition techniques. The paper is structured as follows: section 2 presents the current state of art in the field of Named Entity Recognition in biomedicine. Section 3 gives the theoretical description of natural language processing techniques while section 4 deals with the research methodology, conducted comparative tests and applied evaluation techniques of chosen Named Entity Recognition methods in biomedicine.

2. Related works of Named Entity Recognition in biomedicine

Named Entity Recognition (NER) is a sub-field of information extraction [12]. NER is a task of recognizing the expressions that should be categorised as expressions denoting entities. Example entity names in medical field are diseases, drugs, treatment, genes, cancer, protein and RNA [1, 6, 19]. Much of the research in biomedical informatics has centred on named entity recognition. According to [13] most of the methods are rule-based, although there are implemented some hybrid approach that combine machine learning with these rules.

The authors in [23] mentions Conditional Random Fields (CRF), Support Vector Machines (SVM) and Hidden Markov Model (HMM) as typical machine learning techniques that are currently applied for NER tasks in medical domain. The most recent papers concentrate on deep learning approaches applying recurrent neural networks (RNNs) such as Long-Short Term Memory (LSTM) [14], Gated Recurrent Units (GRU) [16].

Common trend is combining the RNN with statistical method on top of the recurrent layers. It ensures that the optimal sequence of tags over the entire sentence is obtained [15]. CRF is the most commonly used statistical method in this hybrid approach. The authors of [15] combined Residual Dilated Convolutional Neural Network with CRF. Due to the challenges listed below the Clinical NER attempts receive lower performance measures values (best F1 score obtained by [15] is amounts 91.32) in comparison of similar trials with corpuses in non-technical fields, where recently the authors of [3] obtained F1 score 93.5 on the on the CoNLL 2003 corpus.

Firstly, the data available for researchers in the biomedical field is limited, mostly due to the patient privacy and confidentiality requirements. The available annotated databases are usually insufficient for named entity recognition task to train the model [23]. Secondly, the medical texts are written in a specific manner, different from ordinary language. There are a lot of incomplete sentences, informal grammar and littered with misspellings and non-standard shorthand, abbreviations and acronyms. Moreover, the medicine is a rapidly expanding field with large number of researches conducted the contribute to constantly growing number of medical concepts. It makes extremely difficult to keeping the medical dictionaries up to date.

Moreover, concepts in medicine often carry amreferences bigous meaning, related to the concept. It implies the NER models to keep the word context information along the training process. Another typical feature is that medical language is characterised by long phrases containing special characters and dashes.

Most of the studies are performed on Corpuses in English. Second most popular language is Chinese. There are evidently lacking researches in other languages.

3. Natural Language Processing

The term *natural language* is used to describe any language used by human beings, to distinguish it from programming languages and data representation languages used by computers and described as artificial" [11, 12]. *Natural language processing (NLP)* term describes computional techniques that process spoken and written human language [10]. Natural language processing include data preprocessing techniques like data cleaning, tokenization, normalization (stemming, lemmazation or other forms of standardization).

Preparing the text requires choosing the optimal tools, however it helps to improve accuracy of proceeding NLP tasks. Other tasks of NLP concentrate on extracting the statistical features like term frequency (TF), inverse document frequency (IDF) or syntactical features including *Part Of Speach (POS)* tagging. NLP techniques are tools to achieve

the superior task. *Information Extraction (IE)* involving searching for relevant information in documents exist among the most applied tasks.

Named entity recognition (NER) is a stage of Information Extraction. It is one of key NLP tasks that helps to convert unstructured text into computer readable structured data [18, 20, 21, 22]. NER refers to the task of recognizing the expressions denoting entities (i.e. Named Entities), such as diseases, drugs or people's names, in free text documents [7]. NER can be solved with the use of many techniques that can be divided into several groups [2]: dictionary based approach, rule based approach, statistical approach, deep learning approach, hybrid approach.

4. Research methodology and implementation of selected methods

We performed NER task on GENIA corpus. Genia is commonly used corpus by researchers both as dictionary and as base corpus to perform NER task. It is available in different versions and different formats. We used version 1.0 that consists of 1001 abstract records from MEDLINE database and it is a taxonomy of 30 biologically relevant categories.

The corpus is annotated with various levels of linguistic and semantic information related to genes. Original files are in XML based mark-up format. They were transformed for NER task purpose to BIO format with words tagged with "O" which do not belong to any entity or with B_entity_name (B stands for beginning) and I_entity_name (I stands for inside) to indicate the first and subsequent words belonging to the entities.

The dataset was split into train and test subset. Training subset contains 35229 term occurrences. Table 1 presents the general distribution of the annotations among different categories. The most common category is protein molecule. Over 90% of any entity occurrences belong to one third entities.

Table 1. Genia Corpus 3.0.2 entities and their distribution among abstracts

Class name	No of genia abstrats	No of entity occurence	No of unique prhrases
protein_molecule	1774	20855	4005
other_name	1979	13132	6658
protein_family_or_group	1754	7665	2691
cell_type	1637	6844	1867
DNA_domain_or_region	1145	6368	3245
other_organic_compound	738	3938	1179
cell_line	1091	3472	1812
lipid	398	2345	378
protein_complex	674	2167	546
virus	398	2065	366
multi_cell	496	1660	452
DNA_family_or_group	709	1341	745
protein_domain_or_region	342	889	583
protein_subunit	251	817	295
amino_acid_monomer	227	765	179
tissue	290	656	366
cell_component	331	622	207
peptide	146	492	249
body_part	195	432	191
DNA_molecule	278	413	277
atom	115	331	65
inorganic	71	250	64
polynucleotide	150	242	175
RNA_molecule	280	241	134
nucleotide	86	236	59
RNA_family_or_group	185	236	88
mono_cell	76	221	89
other_artificial_source	85	167	94
protein_substructure	73	122	85
DNA_substructure	73	99	79
carbohydrate	21	92	44
protein_N/A	77	86	61
DNA_N/A	36	47	34
RNA_substructure	1	2	2

4.1. Implementation of selected techniques

The purpose of the presenting research was to compare the most commonly used NER techniques – CRF and LSTM. it was checked how well these methods can find large number – 30 – of Named Entities with the uneven frequency distribution. Finally, it was also checked if the integration of CRF with the information from UMLS MetaThesaurus can increase the general performance of NER. The total number of 4 tests were performed.

During the performance of comparison it was also checked if the task can be done with the use of existing tool with minimal user effort. There are many open source libraries available in different languages. Most of them are given for general NER extraction purpose for extracting the traditional Entities like "Person", "Location", "Organisation". One of the most appreciated library is Stanford Named Entity Recognizer which is based on CRF algorithm. This library is written in JAVA but there are also plugins available in different programming languages as Python, .NET, PHP, Ruby [5].

Other well known NER tool is *spaCy* open-source library implemented in Python. The made researches [9] showed that spaCy's NER tool performs second best among four well-established open-source NER tools, regarding accuracy (Stanford NER performance was slightly better) and that it is the fastest in processing speed. However, there is no detailed information provided in its documentation which models are implemented in the background.

Something that is already allocated in medical domain was needed for our study. Therefore, the Python library "CliNER" was chosen – an open-source natural language processing system for named entity recognition in clinical text of electronic health records [4]. It was chose because it implemented all the methods needed for our study.

The library uses two NER extraction techniques: CRF and LSTM, moreover it implemented basic UMLS integration. The library is oriented to find medical terms like TEST, PROBLEM and TREATMENT. So, it was adapted to slightly recognize the entities from Genia corpus. First test was performed with the use of CRF-based classifier. CliNER CRF classifier implements linguistic features as follows:

- word unigram,
- part of speech tag (generated with *nltk pos_tagger*),
- last two characters,
- word shape,
- previous and next features,
- previous and next 3 unigrams,
- regex or units.

Second test was CRF-dictionary as the features set was extended with the domain knowledge from UMLS Metathesaurus. Features were based on semantic types of the phrases found in UMLS Metathesaurus database. Genia classes found their representatives in 119 UMLS different semantic types. Table 2 presents the mapping statistics. It shows how many UMLS semantic types were assigned to each Genia class representatives.

Finally, the bidirectional LSTM for the classiffcation task was employed. The library implements both character level and word level Bi-LSTM. Character sequence embeddings feed into word level LSTM. This approach is sensitive to misspellings [4].

4.2. Performance evaluation

To evaluate the performance of the methods used in this study the confusion matrix data of each test case was collected. A confusion matrix is a matrix used to designate the efficiency of a classification model on a group of test data for which the true values are known [17]. As the evaluation measure, the accuracy, precision, recall and F1 score were used. The precision is used to evaluate the correct degree of prediction power of the model defining the proportion of the correct positive identifications. Recall represents the proportion of positive cases that were correctly identified [17]. F-measure calculates the harmonic mean of precision and recall resulting in achieving the balance between precision and recall [8].

Table 2. Genia to UMLS semantic type mapping

Genia Class	Umls most frequent semantic type	No. Of different Umls types	UMLS influence on F1
RNA_molecule	Intellectual Product	16	10.15
body_part	Body Part, Organ, or Organ Component	32	9.05
mono_cell	Disease or Syndrome	14	6.86
other_organic_compound	Organic Chemical	60	6.64
atom	Element, Ion, or Isotope	14	4.71
protein_substructure	Spatial Concept	20	4.57
amino_acid_monomer	Biologically Active Substance	14	4.33
nucleotide	Nucleic Acid, Nucleoside, or Nucleotide	9	3.5
peptide	Biologically Active Substance	31	2.95
lipid	Organic Chemical	27	2.06
multi_cell	Finding	43	1.72
protein_domain_or_region	Spatial Concept	38	1.63
protein_N/A	Functional Concept	18	1.59
RNA_family_or_group	Nucleic Acid, Nucleoside, or Nucleotide	21	1.14
protein_subunit	Intellectual Product	24	1.02
cell_type	Cell	54	1
DNA_family_or_group	Qualitative Concept	46	0.64
protein_molecule	Enzyme	73	0.39
cell_component	Cell Component	29	0,17
protein_family_or_group	Enzyme	71	0,13
carbohydrate	Organic Chemical	10	0
DNA_N/A	Intellectual Product	15	0
DNA_substructure	Qualitative Concept	15	0
cell_line	Cell	54	-0,17
inorganic	Element, Ion, or Isotope	19	-0,33
virus	Virus	30	-0.52
protein_complex	Intellectual Product	30	-0.58
polynucleotide	Nucleic Acid, Nucleoside, or Nucleotide	12	-1.01
DNA_domain_or_region	Intellectual Product	73	-1.17
tissue	Body Part, Organ, or Organ Component	39	-1.65
other_artificial_source	Functional Concept	20	-2.68
DNA_molecule	Qualitative Concept	32	-3.55

5. Conclusions

Firstly, the data available for researchers in the biomedical field is limited, mostly due to the patient privacy and confidentiality requirements. The available annotated databases are usually insufficient for named entity recognition task to train the model [23]. Secondly, the medical texts are written in a specific manner, different from ordinary language. There are a lot of incomplete sentences, informal grammar and littered with misspellings and non-standard shorthand, abbreviations and acronyms. Moreover, the medicine is a rapidly expanding field with large number of researches conducted the contribute to constantly growing number of medical concepts. It makes extremely difficult to keeping the medical dictionaries up to date.

Moreover, concepts in medicine often carry amreferences bigous meaning, related to the concept. It implies the NER models to keep the word context information along the training process. Another typical feature is that medical language is characterised by long phrases containing special characters and dashes. Most of the studies are performed on Corpuses in English. Second most popular language is Chinese. There are evidently lacking researches in other languages.

The experiments presented in the papers showed that the use of chosen NER methods gives good results, better then the other methods and state the point for the next research and experiments.

References

[1] A. B. Abacha and P. Zweigenbaum, "Medical Entity Recognition: A Comparison of Semantic and Statistical Methods", Proceedings of BioNLP 2011 Workshop, BioNLP'11, pp. 56–64, 2011.

[2] M. Allahyari and S. Pouriyeh and M. Assefi and S. Safaei and E. Trippe and J. B. Gutierrez and K. Kochut, "A Brief Survey of Text Mining: Classifiation, Clustering and Extraction Techniques, 2017.

[3] A. Baevski and S. Edunov and Y. Liu and L. Zettlemoyer and M. Auli, "Cloze-driven Pretraining of Self-attention Networks", http://arxiv.org/abs/1903.07785.

[4] W. Boag and E. Sergeeva and S. Kulshreshtha and P. Szolovits and A. Rumshisky and T. Naumann, "CliNER 2.0: Accessible and Accurate Clinical Concept Extraction, http://arxiv.org/abs/1803.02245.

[5] J. R. Finkel and T. Grenager and C. Manning, "Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling", Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL'05, pp. 363–370, 2005.

[6] V. Hatzivassiloglou and P. A. DubouÃl' and A. Rzhetsky, "Disambiguating proteins, genes, and RNA in text: A machine learning approach" Suppl 1:S97–106. ISSN 1367-4803.

[7] M. A. Hearst, "Untangling Text Data Mining", Proceedings of 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics, ACL'99, pp. 3–10, 1999.

[8] B. Huang and M. T. Kechadi and B. Buckley, "Customer churn prediction in telecommunications", Expert Systems with Applications, Vol. 39(1), pp. 1414–1425, 2012.

[9] R. Jiang and R. E. Banchs and H. Li, "Evaluating and Combining Name Entity Recognition System", pp. 21–27, https://aclweb.org/anthology/papers/W/W16/W16-2703/.

[10] D. Jurafsky and J. H. Martin, "Speech and Language Processing", 2nd Edition. Prentice Hall, ISBN 978-0-13-187321-6.

[11] G. Lample and M. Ballesteros and S. Subramanian and K. Kawakami and C. Dyer, "Neural Architectures for Named Entity Recognition, http://arxiv.org/abs/1603.01360.

[12] S. M. Meystre and G. K. Savova and K. C. Kipper-Schuler and J. F. Hurdle, "Extracting information from textual documents in the electronic health record: A review of recent research", pp. 128–144, ISSN 0943-4747.

[13] S. Pradhan and N. Elhadad and B. R. South and D. Martinez and L. Christensen and A. Vogel and H. Suominen and W. W. Chapman and G. Savova, Evaluating the state of the art in disorder recognition and normalization of the clinical narrative", Vol. 22(1), pp. 143–154, ISSN 1527-974X.

[14] Y. Qin and Y. Zeng, "Research of Clinical Named Entity Recognition Based on Bi-LSTM-CRF", Journal of Shanghai Jiaotong University (Science), Vol. 23(3), pp. 392–397, 2018.

[15] J. Qiu and Q. Wang and Y. Zhou and T. Ruan and J. Gao, "Fast and Accurate Recognition of Chinese Clinical Named Entities with Residual Dilated Convolutions", Proceedings of IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 935–942, 2018.

[16] A. P. Quimbaya and A. S. MÃžnera and R. A. GonzÃąlez Rivera and J. C. Daza RodrÃŋguez and O. M. MuÃśoz Velandia and A. A. Garcia PeÃśa and C. LabbÃl', "Named Entity Recognition Over Electronic Health Records Through a Combined Dictionary-based Approach", Procedia Computer Science, Vol. 100, pp. 55–61, 2016.

[17] S. R. Gottumukkala and P. J. Lum and R. Slack and S. Thirumurthi and P. M. Lynch and E. Miller and B. R. Weston and M. L. Davila and M. S. Bhutani and M. A. Shafi and R. S. Bresalier and A. A. Dekovich and J. H. Lee and S. Guha and M. Pande and B. Blechacz and A. Rashid and M. Routbort and G. Shuttlesworth and L. Mishra and J. R. Stroehlein and W. A. Ross, "Natural Language Processing As an Alternative to Manual Reporting of Colonoscopy Quality Metrics", Gastrointestinal endoscopy, Vol. 82(3), pp. 512–519, 2015.

[18] Y. Sasaki and Y. Tsuruoka and J. McNaught and S. Ananiadou, "How to make the most of NE dictionaries in statistical NER", Vol. 9(11):S5, ISSN 1471-2105.

[19] Y-J. Song and B.-C. Jo and C.-Y. Park and J.-D. Kim and Y.-S. Kim, "Comparison of named entity recognition methodologies in biomedical documents", BioMedical Engineering OnLine, Vol. 17(2):158, 2018.

[20] W. Sun and Z. Cai and Y. Li and F. Liu and S. Fang and G. Wang, "Data Processing and Text Mining Technologies on Electronic Medical Records: A Review", Journal of Healthcare Engineering, 2018:4302425, 2018.

[21] C. Sutton and A. McCallum, An Introduction to Conditional Random Fields", arXiv:1011.4088 [stat], November 2010.

[22] Y. Wu and M. Jiang and J. Xu and D. Zhi and H. Xu, "Clinical Named Entity Recognition Using Deep Learning Models", Proceedings of AMIA Annual Symposium, pp. 1812–1819, 2018.

[23] J. Zhang and J. Li and S. Wang and Y. Zhang and Y. Cao and L. Hou and X. Li, "Category Multi-representation: A Unified Solution for Named Entity Recognition in Clinical Texts", Advances in Knowledge Discovery and Data Mining, Springer, pp. 275–287, 2018.