Biomedical Named Entity Recognition– A Theoretical Study

PUSHPALATHA M

Assistant Professor of Computer Science Maharani's Science College of Women, Mysuru-570005.

***Dr*. ANTONY *SELVADOSS* THANAMANI**

Associate Professor & HOD. M.Sc., M. Phil., Ph.D., PGDCA.

Research Department of Computer Science

NGM College, Pollachi-642001.

Abstract

The present paper deals with the review work of the present trend of data and its types. This paper deals with the named entity recognition in biomedical filed and its categories. Named entity is a task which involves the extraction of information in relation to biomedical data. It extracts the information regarding the clinical problems, solutions, practices and other details relating to problems which will help the clinicians to handle the other such cases in further. So in general the named entity recognition in the field of biomedical is a ready encyclopedia for the society and people.

Key Words: *Biomedical Named Entity Recognition*

1 Introduction

Named-entity recognition (NER) is a subtask of information extraction that seeks to locate and classify named entities in text into pre-defined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc. The present era is a developmental era and it is well acknowledged the easy expansion and distribution of the Internet has resulted in large quantity of information being produced and shared, which it exist in the form of textual data, images, videos and sounds. This shocking flow of data is also factual for specific area such as biomedical. The publications such as articles, books and technical reports, journals are available enormously. The valuable information are collected and accumulated in the form of structured data resources.

Named Entity Recognition (NER) is big task involved in Extracting Information in order to identify and classify the types of information .Named Entity serves as the basis for other important fields of information management like; Semantic Annotation, Question Answering, Ontology Population and Opinion Mining.

In general, named-entity recognition (NER) mainly based on identifying the names of persons, locations, and organizations in news articles, reports, and even tweets. The availability of annotated corpora, supervised learning methods have been widely accepted and prevail unsupervised. Such a state-of-the-art NER system has not only achieved high performance for human annotators but on another side Biomedical-Named Entity Recognition are getting better advantage with the annotated corpora to learn from. Recent advancement in the system could efficiently find clinical problems and gene names.

With the sudden increase of information in the biomedical domain there is a huge demand for automated biomedical information extraction techniques. The named entity (NE) recognition in the fields such as proteins, DNAs, RNAs, cells etc. has become an important tasks in the discovery of biomedical knowledge. While a number of algorithms have been planned for this task, biomedical named entity recognition (NER) is still remains a demanding task and an active area for research in the field of biomedical.

1.1 Why do we need Named Entity Recognition?

According to a market survey performed by IDC, between 2009 and 2020 the amount of digital information will grow along with the staffing and investment to manage it. Dealing with the disparity is very big challenge and is one of the proposals to improve the crisis by developing tools for the search and discovery of information which includes the ways to convert structure to unstructured data. Named Entity Recognition implies identifying the interest in unstructured texts which is exactly one of the major goal and serving as the basis for many other vital areas of information.

2 Evolution of Named Entity Recognition

The term "Named Entity" was first coined in the sixth message understanding conference by (MUC-6) (Grishman & Sundheim, 1996). In 2002 Petasis and co workers advocated NE definition as to "a proper noun, serving as a name for something or someone". They justified this restriction merely because of the significant percentage of proper nouns present in a corpus. On the other hand Alfonseca and Manandhar defined NER as "the task of classifying unknown objects in known hierarchies that are of interest for us for being very useful to solve a particular problem". This approach has been followed by the BBN hierarchy and the GENIA ontology for Information Retrieval (IR) and Question Answering (QA) tasks.

2.2 What is a Named Entity?

Experts in Named Entity Recognition have given several definitions for NE. After analysis we understand that NE can be classified into following four criterions: grammatical category, rigid designation, unique identification and domain of application.

There have been many attempts to develop techniques to identify NE in the biomedical literature. There are two main steps of named entity recognition: detecting boundaries of entity mentions and classifying the mentions into pre-defined semantic categories. The task of entity is to connect or normalize that is connecting a term to an exclusive concept identifier in a terminology.

3 Unsupervised Named Entity Recognition

NLP community has invested lot of efforts in unsupervised NER. Early work relies on heuristic rules and lexical resources such as WordNet. More recently, Alfonseca and Manandhar proposed named entity classification as a word sense disambiguation mission and cluster words based on the words with which they co-occur repeatedly in online search results. The context word frequency vector, which represents the semantics of words to be classified, is called "signature."

Nadeau et al. give a system of retrieving entity lists by web page wrapper, followed by disambiguation through heuristic rules. Sekine and Nobata give definitions and rule-based taggers for 200 categories of entities, as well as a standard taxonomy of general entities.

3.1 Biomedical Named Entity Recognition

There are two major research directions in BM-NER: finding gene, protein, and related biological and genetic terms, and also finding disease, drug names, and other medical terms. We use biological NER and medical NER to represent these two research sub-domains respectively.

The early NER systems in both the fields are typically rule-based or lexicon-based, several of which are widely accepted. MedLEE is a general natural language processor for clinical texts, encoding and mapping terms to a controlled vocabulary; GENIES is a system extracting molecular pathways from journal articles, which is modified from MedLEE; EDGAR is a natural language processing system that add information about drugs and genes relevant to cancer from the biomedical literature; AbGene is one of the most and best successful NER systems for gene and protein; MetaMap, developed by National Library of Medicine(NLM), is a tool discovering UMLS Metathesaurus concepts referred to in text. Many of these systems highly resort to lexical knowledge resources such as GO and UMLS.

Very recently cTAKES provides concept identification and normalization to UMLS in clinical texts. In the medical domain, the first publicly available corpus for NER evaluation was created in the i2b2 challenge 2010. In this event, 22 supervised and semi-supervised systems were developed for entity extraction, and most of the leading systems used CRF, except for the best performed system. Before the availability of i2b2 corpus, recent research very much focused on evaluation on, extension to, and comparison with MetaMap and its predecessor MMTx. Meystre and Haug evaluate MMTx with a automatically list of clinical problems.

4 Challenges and Opportunities in NER

NER is not a solved task, but it can be solved. At least, to the extent any other domain-dependent task can be considered as solved. The difficulty is that present assessment, practices and resources in NER do not permit us to decide. NER has been considered a solved problem when the system achieves a minimum performance with a good result of NE types. We are not sure about the present techniques which perform with other types of NE with different kinds of documents. There are no traditionally accepted ways to assess the new types of NE tools and its recognition nowadays. The new evaluation methods have to solve some of the limitations and they are not enough to assess the development of NER because they asses systems with unusual goals which are not valid for most NER applications.

Conclusions

Named Entity Recognition play an important role in Information Extraction tasks such as Identification of Relationships and Scenario Template Production, as well as other areas such as Semantic Annotation, Ontology Population or Opinion Mining, just to name a few. However, the definitions given for Named Entity have been very diverse, ambiguous and incongruent so far. It is necessary to take NER back to the research community and develop adequate evaluation forums, with a clear idea about the task and user models, and the use of appropriate measures and standard methodologies. Only by doing so may we really contemplate the possibility of NER being a solved problem.

REFERENCES:

1. Abacha, A.; Zweigenbaum, P. Medical entity recognition: a comparison of semantic and statistical methods. Proceedings of BioNLP 2011 Workshop; Association for Computational Linguistics; 2011. p. 56-64.
2. Alfonseca, E.; Manandhar, S. An unsupervised method for general named entity recognition and automated concept discovery. Proceedings of the 1st International Conference on General WordNet; Mysore, India. 2002. p. 34-43.
3. Aronson, A. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. Proceedings of the AMIA Symposium; American Medical Informatics Association; 2001. p. 17
4. Ashburner M, Ball C, Blake J, Botstein D, Butler H, Cherry J, Davis A, Dolinski K, Dwight S, Eppig J, et al. Gene ontology: tool for the unification of biology. Nature genetics. 2000; 25(1):25. [PubMed: 10802651]
5. Bodenreider O. The unified medical language system (umls): integrating biomedical terminology. Nucleic acids research. 2004; 32(suppl 1):D267–D270. [PubMed: 14681409]
6. Brunstein, “Annotation guidelines for answer types,” 2002.
7. D. Nadeau and S. Sekine, “A survey of named entity recognition and classification,” Linguisticae Investigationes, vol. 30, no. 7, 2007.
8. D. Nadeau, “Semi-Supervised Named Entity Recognition: Learning to Recognize 100 Entity Types with Little Supervision,” Department of Information Technology and Engineering, University of Ottawa, 2007.
9. E. Alfonseca and S. Manandhar, “An unsupervised method for general named entity recognition and automated concept discovery,” in 1st International Conference on General WordNet, 2002.
10. Finkel, J.; Grenager, T.; Manning, C. Incorporating non-local information into information extraction systems by gibbs sampling. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics; Association for Computational Linguistics; 2005. p. 363-370.
11. Friedman C, Kra P, Yu H, Krauthammer M, Rzhetsky A. Genies: a natural-language processing system for the extraction of molecular pathways
12. Fukuda K, Tsunoda T, Tamura A, Takagi T, et al. Toward information extraction: identifying protein names from biological papers. Pac Symp Biocomput. 1998; 707:707–718. [PubMed: 9697224]
13. G. Petasis, A. Cucchiarelli, P. Velardi, G. Paliouras, V. Karkaletsis, and C. D. Spyropoulos, “Automatic adaptation of Proper Noun Dictionaries through cooperation of machine learning and probabilistic methods,” in 23rd annual international ACM SIGIR conference on Research and development in information retrieval, 2000, pp. 128-135.
14. Grisman R. and Sundheim B. (1996). Message Understanding Conference – 6: A Brief History. Inproceedings of International Conference on Computational Linguistics.
15. J. Gantz and D. Reinsel, “The Digital Universe Decade, Are You Ready?,” 2010.
16. Jd. Kim, T. Ohta, Y. Teteisi, and J. Tsujii, “GENIA corpus: a semantically annotated corpus for bio-textmining,” Bioinformatics (Oxford, England), vol. 19, no. 1, p. 180, 2003.
17. Jurafsky, D.; Martin, J.; Kehler, A. Speech and language processing: an introduction to natural language processing, computational linguistics and speech recognition. Vol. 2. MIT Press; 2002.
18. McCallum, A.; Li, W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. Proceedings of the seventh conference on Natural language learning at HLT-NAACL; 2003; Association for Computational Linguistics; 2003. p. 188-191.
19. Meystre, S.; Haug, P. Comparing natural language processing tools to extract medical problems from narrative text. AMIA Annual Symposium Proceedings; American Medical Informatics Association; 2005. p. 525
20. Morgan A, Lu Z, Wang X, Cohen A, Fluck J, Ruch P, Divoli A, Fundel K, Leaman R, Hakenberg J, et al. Overview of biocreative ii gene normalization. Genome biology. 2008; 9(Suppl 2):S3. [PubMed: 18834494]
21. Nadeau D, Sekine S. A survey of named entity recognition and classification. Lingvisticae Investigationes. 2007; 30(1):3–26.
22. Nadeau D, Turney P, Matwin S. Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity. Advances in Artificial Intelligence. 2006:266–277.
23. Proux D, Rechenmann F, Julliard L, Pillet V, Jacq B, et al. Detecting gene symbols and names in biological texts: a first step toward pertinent information extraction. GENOME INFORMATICS SERIES. 1998:72–80. [PubMed: 11072323]
24. Ratinov, L.; Roth, D. Design challenges and misconceptions in named entity recognition. Proceedings of the Thirteenth Conference on Computational Natural Language Learning; Association for Computational Linguistics; 2009. p. 147-155.
25. Rindflesch, T.; Tanabe, L.; Weinstein, J.; Hunter, L. Edgar: extraction of drugs, genes and relations from the biomedical literature. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, NIH Public Access; 2000. p. 517
26. Sea, K.; deBruijn, B.; Cherry, C. Nrc at i2b2: one challenge, three practical tasks, nine statistical systems, hundreds of clinical records, millions of useful features. Proceedings of the 2010 i2b2/ VA Workshop on Challenges in Natural Language Processing for Clinical Data; 2010.
27. SGK, MJJ, OPV ZJ, SS, K-SKC, CCG. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. 2010; 17(5):507–513. [PubMed: 20819853]
28. Wang, Y.; Patrick, J. Cascading classifiers for named entity recognition in clinical notes. Proceedings of the Workshop on Biomedical Information Extraction; Association for Computational Linguistics; 2009. p. 42-49.
29. Zhou, G.; Su, J. Named entity recognition using an hmm-based chunk tagger. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics; Association for Computational Linguistics; 2002. p. 473-480.