Research Article

A Survey on Biomedical Named Entity Extraction

Almas Tasneem,[1*] Archana B[1]

Abstract: There has been growing work in the field of Named Entity Recognition (NER) and a lot of research has been done in this direction in last two decades. Particularly, a lot of progress has been made in the biomedical domain with emphasis on identifying domain-specific entities and often the task being known as Biological Named Entity Recognition (BIO-NER).Biomedical named Entity Recognition is a crucial step towards efficient texts analysis in the medicine. The biomedical domain, due to its complicated language, has consistently lagged behind. Our aim is to extract named entities such as Gene, Protein or Cell Types. This paper looks at the challenges perceived by the researchers in BIO-NER task and investigates the works done in the field of BIO-NER by using the multiple approaches available for the task.

Asian Journal of Engineering and Technology Innovation
Volume 4, Issue 7
Published on: 7/05/2016

Cite this article as: Almas Tasneem, Archana B. A Survey on Biomedical Named Entity Extraction. Asian Journal of Engineering and Technology Innovation, Vol 4(7): 25-28, 2016.

INTRODUCTION

The term "Named Entity" was first coined in the sixth message understanding conference (MUC-6). The aim there was to extract named entities such as people, organization or location names from news articles. Over the past 12 years, the task of Named Entity Extraction has attracted considerable amount of research and a number of successful systems such as LBJ (Rizzolo and Roth, 2007) with accuracies of over 90% have been developed.

One of the fundamental tasks in Natural Language Processing (NLP) is text mining. Most of the text mining system depends upon the methods and tools of NLP. Text mining can be defined as a knowledge extracting method to extract useful and previously unknown information from a document or set of texts [1]. Biomedical text mining, is applying the automated methods of text mining for extracting the enormous amount of knowledge available in the biomedical literature [2]. It covers a wide range of applications, such as, document classification, text mining, question answering, ontology development, literature-based discovery etc.

Named entity recognition is a task that tries to find entities in the text and classifies these entities into some predefined classes. The examples of relation extraction from biomedical text include gene-disease relationships, protein-protein interactions, drug-drug interactions etc. However, the focus of this paper is to look in detail at the works done in the field of entity recognition in biological domain.

Biological Named Entity Recognition (BIO-NER)

Biological Named Entity Recognition (BIO-NER) is a subfield of NER where the text under consideration is a biological text and the predefined categories of entities are from biological domain, such as, the names of proteins, genes, diseases or cell types. The idea is recognizing biological entities present in the text and further extraction of relationships and other information by identifying the key values of interest and hence allowing more complex text-mining operations to be performed.

COMPARISON OF PROPOSED APPROACHES

The proposed methods can broadly be divided into three main categories: Rule Based, Dictionary Based and Machine Learning Based methods.

Dictionary-Based Approach

One fundamental approach of performing NER is to utilize a descriptive list of terms such as dictionary or lexicons, also termed as terminological resources, which can be the basis of identifying entity mentions in text. This type of approach is known as dictionary-based approach. If the word or group of words from the text matches with the term from the list, it is identified as entity occurrence. This method is found to have a high degree of precision but it has a poor recall. Many improvements have been suggested to increase the precision and recall of dictionary-based approaches and to overcome the difficulties, such as, generating spelling variants, appending additional terms to the underlying term lists etc.

[1]Reva Institute of Technology and Management, Rukmini Knowledge Park, Kattigenahalli, Yelahanka, Near Border Security Bustop, Bengaluru, Karnataka-560064, India.
E-mail: ashwin@revainstitution.org
*Corresponding author

1. Important Works on Dictionary-Based Approach

Dictionary-based approach is considered as the fundamental approach of identifying entity mentions in text by using a dictionary or lexicons. The works have discovered for this approach to have a high degree of precision but a poor recall. Tuason et al. reported that the low recall may be attributed to spelling mistake, character-level and word-level variations [3].

The major issue with the use of vocabularies of terms is that it is not possible to have limited list of terms and furthermore new terms are introduced by researchers and scientists around the world at very fast rate making most of these vocabularies out of date very soon.

The lower precision and recall and other reported issues in the dictionary-based methods led to adoption of many enhancements to these approaches. Generation of spelling variations to get the terms for a biomedical resource and then adding terms to the primary lists is one example of enhancement [4]. Then the expanded list can be used to do exact string matching. Although there are many of these enhancements have been attempted, yet dictionary-based methods are frequently used in combination with more advanced NER approaches.

Rule-Based Approach

Rule-based approach is another approach to NER. Here rules are defined in an attempt to recognize entities which describe the formation patterns and context of named entities. In this approach, the rules are developed manually using lexical-syntactic features or using existing information lists. Rule-based approaches are said to achieve better performance when compared to dictionary-based approaches. Lot of effort and time is invested to build the resources and rules. However, they are time-consuming and hard as rules are mainly handcrafted. Further, rules are very problem specific and domain specific to achieve high precision. These approaches have limited portability as far as transferring across other domains is concerned.

1. Important Works on Rule Based Approaches

The early years of NER task were predominately based on rule-based approaches [5]. Fukuda et al., (1998) proposed a method called PROPER (Protein Proper-noun phrase Extracting Rules), for identifying protein names from biomedical documents.

Rule-based systems initially seemed promising, but they failed to perform well on larger datasets. For example when Proux et al. evaluated their performance on a larger corpus of 25,000 MEDLINE abstracts by sampling, the precision fell to 70%. It is also very expensive to adapt these systems to verify new entity classes as the rules are to be developed manually. Moreover, these systems cannot identify new named entities since new entity names are frequently coined in the biomedical domain, this is a significant drawback. [14]

Machine Learning Based Approaches

We accomplish the task of extracting biomedical entities using statistical methods by applying some kind of machine learning algorithm. The machine learning paradigm can be viewed as - programming by example. In this technique a system learns automatically by using negative and positive training examples for the task with the help of features linked with examples. The selected machine learning algorithms automatically differentiate negative examples from positive examples and can be further used to identify similar information from the data which is still unseen [6]. Machine Learning algorithms are generally classified into three types:

- Supervised learning
- Semi-supervised Learning and
- Unsupervised Learning

1. Supervised Learning

Supervised learning technique is the most frequently used and still the dominant approach in the NER community. It is based on the idea of studying the features of positive and negative examples of NE over a large collection of annotated data [7]. There are several supervised learning techniques such as, Support Vector Machines (SVM), Hidden Markov Models (HMM), Maximum Entropy Models (ME), Decision Trees, and Conditional Random Fields (CRF). Supervised learning methods in NER task require a large amount of training, usually manually annotated data demands a lot of cost and time investment. Of late bootstrapping and other semi-supervised statistical techniques have been used to automatically generate training data.

a. Important works onSupervised learning

NER task uses Hidden Markov Models (HMM), Support Vector Machines (SVM), and Conditional Random Fields (CRF), for supervised learning.

SVMs have been highly successful in automated text classification. SVMs are primarily binary classifiers and are often trained using a one-vs-rest approach. The training time of an SVM is super linear to the size of the training set due to SVMs are quadratic optimization algorithms. Thus directly training with one-vs-rest approach on a data is not feasible. Much of the research has therefore been concentrated towards solving these problems.

Conditional Random Fields are said to be well suited for biomedical NER. Their use was first explored by Settles (2004) [15]. His system achieved an F-Score of close to 70%, the highest, on the JNLPBA 2004 task. Chan et al., (2008) dealt

with the issues of entity segmentation and classification separately in a cascaded manner. [12]

2. Semi-supervised Learning (SSL)

Semi-supervised learning uses both labeled data and unlabeled data for the learning process to reduce the dependence on training data. Bootstrapping is the main technique used for semi-supervised learning in NER with lesser degree of supervision. The system is first trained on an initial small set of examples and unlabeled data is tagged. The resulting annotations are then highlighted to increase the initial training set. The added training set is then used to re-train the system. This process iterates to progressively refine the learning model.

a. Important works on Semi-supervised Learning (SSL)

Labeled and unlabeled data lessens the dependence of supervised learning technique. The main technique of semi-supervised learning is Bootstrapping or self-training. They require a small degree of supervision. Bootstrapping method in SSL became quite popular and many NER methods are using bootstrapping approaches.

Cucchiarelli and Velardi used examples from existing NER systems for starting examples [8]. They relied on subject-object relations to find better contextual evidence about the entities.

M. Pasca et al. work was also motivated by method of mutual bootstrapping. The very evident limitation of the bootstrapping approach is propagating the error once it has been introduced. Another problem is that inadequate contextual information hinders the pattern generalization when low frequency classes of entities are present. [9]

3. Unsupervised Learning (USL)

In the unsupervised learning, decisions are made on unlabeled data. The methods of unsupervised learning are mostly built upon clustering techniques, similarity based functions and distribution statistics.

a. Important works on Unsupervised Learning (USL)

Unsupervised learning methods make decisions on a large unannotated data. The main approach used for the task of NER in unsupervised learning is clustering technique. There are other unsupervised approaches, such as, similarity based functions and distribution statistics.

Alfonseca et al. [10] presented a work of labelling an input word with an appropriate NE type taken from WordNet. In another work of USL, Evans [11] presented a system of Named Entity Recognition in the Open Domain (NERO) in which he worked on the problem of NER for identification of any types of entities useful in any scenario context.

There have many works reported in the literature which do use combination of different approaches to enhance the performance.

CHALLENGES IN BER

The task of biomedical entity recognition (BIO-NER) may appear to be straight forward at the first glance. But it is a challenging task for several reasons. Marrero M. et al. argues that NER is in fact not a solved problem, and acknowledged that the lack of agreement around the concept of Named Entity has important implications for NER research. The difficulty associated with the task of entity recognition in biomedical domain as compared to other domains has been attributed to several factors proposed by many researchers. The literature in biomedical domain makes use of millions of entity names with new ones being added to the list by every passing day, thus making it difficult for dictionaries and lexicons to be up-to-date [11].

Detecting their boundaries of biomedical name are also usually longer than the names in other domains and is comparatively more difficult. Entities names can be overlapping making it hard to find which one is right.

Biomedical literature uses abbreviations that are very frequently used. The problem with use of abbreviations in the biomedical domain is that these can match common English words or have multiple homonyms [13]. The situation is aggravated by the fact that the naming conventions, although there are not many, are usually not followed.

CRITICAL ISSUES WITH MACHINE LEARNING TECHNIQUES

While machine learning techniques such as HMM, SVM and CRF have proven to be quite effective in building Bio-NER systems, their performance depends heavily on the quality and quantity of the selected features and the training set. Building a large training set requires considerable manual effort and any inconsistency in annotation may adversely affect the training and evaluation of these classifiers. Even in a standardized dataset such as GENIA for example, many entities have been doubly classified as “protein molecule or region” and “DNA molecule or region”. As was evident from the performances (max F-Score of 75%) in the JNLPBA 2004 and the BioCreative2004 task 1A, machine learning algorithms tend to get confused by these mistakes. Shen et al. (2004) presented an active learning approach in this regard, to minimize the human effort required to annotate the datasets. They considered three different criteria, namely information, representativeness and diversity and proposed measures to quantify them. Results showed that labelling costs could be reduced by at least 80% without degrading the performance [16].

CONCLUSION AND FUTURE WORK

In this paper we surveyed the area of BIO-NER and looked at the different approaches practiced for the task by many researchers. A lot of work has been done in BIO-NER by using the simple dictionary based method and lot many improvements have been tried to overcome the limitations found in this approach. However, we observe that despite these improvements, dictionary-based methods are most often used in combination with more advanced NER approaches. Rule based approaches have also been applied very frequently to the task of BIO-NER and we find many such works reported for it.

Although the supervised machine learning based approaches have made BIO-NER systems practical by far outperforming the rule or dictionary based methods, for them the problem remains in creating large enough training sets. It is thus the exploration of un-supervised or semi-supervised techniques is likely in future.

REFERENCES AND NOTES

1. Holzinger, A., Schantl, J., Schroettner, M., Seifert, C., &Verspoor, K. (2014). Biomedical text mining: State-of-theart, open problems and future challenges. In Interactive Knowledge Discovery and Data Mining in Biomedical Informatics (pp. 271-300). Springer Bio-NERlinHeidelBio-NERg.
2. Chapman, Wendy W. Cohen, K. Bretonnel et al. (2009). Current issues in biomedical text mining and natural language processing. Journal of Biomedical Informatics, Volume 42, Issue 5 , 757 – 759.
3. Tuason, O., Chen, L., Liu, H., Blake, J. A., & Friedman, C. (2004). Biological nomenclatures: a source of lexical knowledge and ambiguity. In Pac SympBiocomput (pp. 238-249).
4. Tsuruoka, Y., &Tsujii, J. I. (2003, July). Boosting precision and recall of dictionary-based protein name recognition. In Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine-Volume 13 (pp. 41-48). Association for Computational Linguistics.
5. Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. Lingvisticae Investigationes, 30(1), 3-26
6. Zhang, Z., Cohn, T., &Ciravegna, F. (2013). Topicoriented words as features for named entity recognition. In Computational Linguistics and Intelligent Text Processing (pp. 304-316). Springer Bio-NERlinHeidelBio-NERg.
7. Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. Lingvisticae Investigationes, 30(1), 3-26
8. Cucchiarelli, A., & Velardi, P. (2001). Unsupervised named entity recognition using syntactic and semantic contextual evidence. Computational Linguistics, 27(1), 123-131.
9. Pasca, M., Lin, D., Bigham, J., Lifchits, A., & Jain, A. (2006, July). Organizing and searching the world wide web of facts-step one: the one-million fact extraction challenge. In AAAI (Vol. 6, pp. 1400-1405).
10. Alfonseca, E., & Manandhar, S. (2002, January). An unsupervised method for general named entity recognition and automated concept discovery. In Proceedings of the 1st International Conference on General WordNet, Mysore, India (pp. 34-43).
11. Wilbur, J.; L. Smith; and T. Tanabe. (2007) BioCreative 2. Gene Mention Task. Proceedings of the Second BioCreative Challenge Workshop pp. 7-16.
12. Chen, L., H. Liu and C. Friedman (2005). "Gene name ambiguity of eukaryotic nomen-clatures." Bioinformatics 21(2): 248-256.
13. Liu, H., Aronson, A. R., & Friedman, C. (2002). A study of abbreviations in MEDLINE abstracts. Proceedings of the AMIA Symposium, 464–468.
14. Proux D, Rechenmann F, Julliard L, Pillet V, Jacq B.(1998) Detecting gene symbols and names in biological texts: a first step toward pertinent information extraction. In: Proceedings of genome inform ser workshop genome inform; pp. 72–80.
15. Settles B. (2004) Biomedical named entity recognition using conditional random fields and novel feature sets. In: Proceedings of the joint workshop on natural language processing in biomedicine and its applications, Geneva, Switzerland; pp. 104–7.
16. Shen D., Zhang J., Su J., Zhou G., Tan C. (2004) Multi-criteria-based active learning for named entity recognition, Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, pp.589-es, July 21-26, 2004, Barcelona, Spain.