

A Survey of Biological Entity Recognition Approaches

Gurinder Pal Singh Gosal
Department of Computer Science
Punjabi University
Patiala
e-mail: gosal.gps@gmail.com

Abstract—There has been growing interest in the task 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 (BER). The task of biological entity recognition (BER) has been proved to be a challenging task due to several reasons as identified by many researchers. The recognition of biological entities in text and the extraction of relationships between them have paved the way for doing more complex text-mining tasks and building further applications. This paper looks at the challenges perceived by the researchers in BER task and investigates the works done in the domain of BER by using the multiple approaches available for the task.

Keywords- Named Entity Recognition, NER, Biological named Entity Recognition, BER, Information Extraction, Text Mining, Bio-NLP.

I. INTRODUCTION

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 by identifying the facts inherent and inexplicit in the data [1]. Biomedical text mining, the term some practitioners use as synonymous with **Bio-NLP**, is applying the automated methods of text mining for exploiting the enormous amount of knowledge available in the biomedical literature [2]. Bio-NLP covers a wide range of applications covered under its ambit, such as, document classification, text mining, summarization, question-answering, ontology development, literature-based discovery etc. as shown in *Figure 1*.

The task with goal of extracting explicitly stated facts or structured facts from unstructured or semi-structured text is termed as **information extraction** in text mining systems. Information extraction often becomes the basis of other biomedical text mining applications and is considered as an initial processing step in this direction. For processing unstructured text, three major subtasks of information extraction are reported in literature:

- Named Entity Recognition (NER)
- Relation Extraction
- Event Extraction

Named entity recognition is a task that tries to find entities in the text and classifies these entities into some predefined classes. The task of **relation extraction** from the text, to detect binary relationships among named entities, is another subtask of information extraction relevant to the biomedical domain. The examples of relation extraction from biomedical text include gene-disease relationships, protein-protein interactions, drug-drug interactions etc. The task of

identifying highly complex relations among detected entities is known as **event extraction** and is the third sub-task in information extraction. The events such as gene expression and regulation and protein binding etc. are examples of event extraction task in the biomedical domain. However, the focus of this paper is to look in detail at the works done in the field of entity recognition in biological domain.

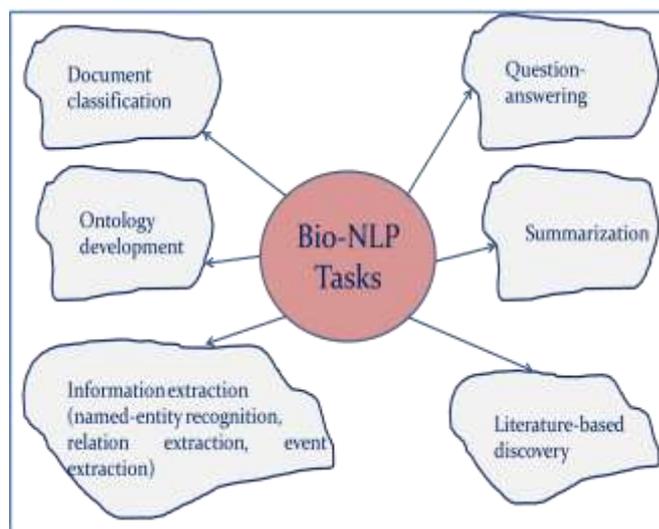


Figure 1. Bio-NLP Tasks

A. Biological Named Entity Recognition (BER)

Biological Named Entity Recognition (BER) is a subfield of NER when 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, or diseases. The idea is that recognizing biological entities in text allows for further extraction of relationships and other information by identifying the key concepts of interest and hence letting more complex text-mining tasks to be performed. *Figure 2* shows the sub-tasks performed under “Information Extraction” with BER being a sub-field under of NER.

B. Entity Recognition Approaches

Generally the NER, and therefore BER per se, can be grouped into the following categories as far as the approaches to perform the task are concerned:

B(i). Dictionary-Based Approach

One fundamental approach of performing NER is to utilize a descriptive list of terms (dictionary or lexicons), also termed as terminological resources, that 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 recognized as entity occurrence. This kind of approach is found to have a high degree of precision but has a poor recall. Many improvements have been suggested to increase the precision and recall of dictionary-based approaches and to overcome the stated difficulties, such as, generating spelling variants for the terms in a biomedical resource, appending additional terms to the underlying term lists etc. Despite these improvements, dictionary-based methods are most often used in conjunction with more advanced NER approaches.

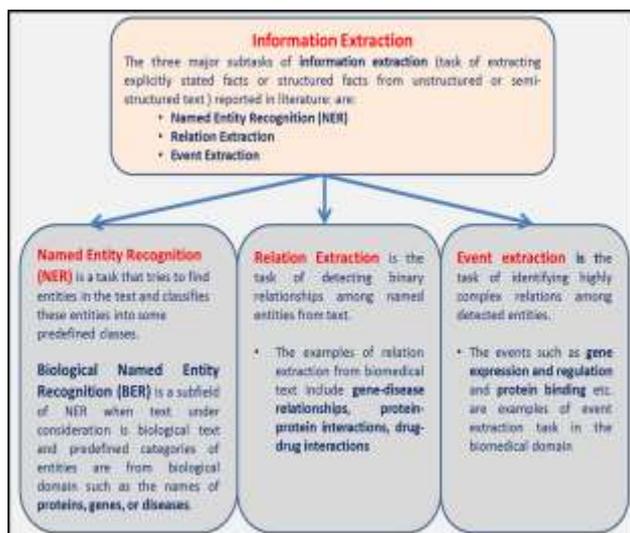


Figure 2. Information Extraction and its Sub-Tasks

B(ii). Rule-Based Approach

Rule-based approach is another approach to NER in which rules are defined in an attempt to recognize entities which describe the formation patterns of named entities and their context. In this approach, the rules are developed typically manually using lexical-syntactic features or using existing information lists, such as, dictionaries of typical term constituents like terminological heads, affixes, specific acronyms etc. Rule-based approaches are considered to typically achieve better performance than dictionary-based approaches. However, these kinds of knowledge engineering approaches are known to be time-consuming and hard as rules are mainly handcrafted. Further, rules tend to be very problem specific as well as domain specific in order to achieve high precision.

Even after putting so much effort and time to build the resources and rules, these approaches have limited

portability as far as transferring across other domains is concerned. Therefore, it is becoming common to shift towards more robust learning-based approaches instead of or in combination with dictionary-based and rule-based approaches [3].

B(iii). Learning Based Approaches

The task of extracting biomedical entities using statistical methods is usually accomplished by applying some kind of machine learning algorithm. The machine learning paradigm can be viewed as “programming by example.” It is a technique in which system learns automatically by using negative and positive **training examples** for the task with the help of distinguishing **features** linked with examples. The selected machine learning algorithms automatically differentiate negative examples from positive examples and this learning can be used further to identify similar information from the data which is still unseen [4].

Learning algorithms can be generally classified into three types:

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

B(iii)(a). Supervised Learning

Supervised learning technique is based on the idea of studying the features of positive and negative examples of NE over a large collection of annotated corpus and is the most frequently used and still the dominant approach in the NER community [5]. There are several supervised learning techniques which are used extensively for the task, such as, Support Vector Machines (SVM), Hidden Markov Models (HMM), Decision Trees, Maximum Entropy Models (ME), and Conditional Random Fields (CRF).

The main problem with the use of supervised learning methods in NER task is the requirement of large amount of training, usually manually annotated data which demands a lot of cost and time investment. Of late there have been efforts to automatically generate training data for the NER task by using bootstrapping and other semi-supervised statistical techniques.

B(iii)(b). 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. The main technique of semi-supervised learning in NER is “**bootstrapping**” that needs a lesser degree of supervision. The system is first trained on an initial small set of seeds (examples) to tag unlabeled data and the resulting annotations are then selected to increase the initial training set. The augmented training set is then used to re-train the system and this process iterates to progressively refine the learning model.

B(iii)(c). Unsupervised Learning (USL)

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

The various entity recognition approaches are depicted in the *Figure 3*.

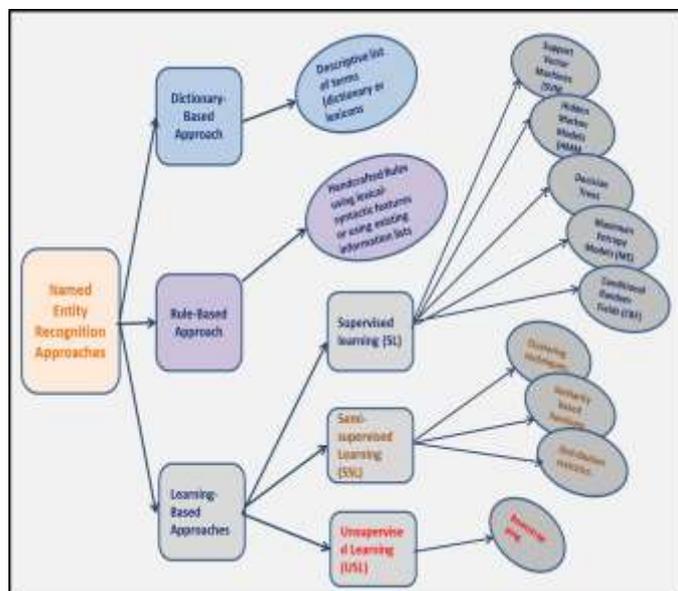


Figure 3. (Biological) Named Entity Recognition Approaches

II. SURVEY OF BER APPROACHES

A lot of progress has been made in the direction of Named Entity Recognition task since 1996 when the 6th Message Understanding Conference (MUC6) was organized [6]. The 6th Message Understanding Conference (MUC6) saw first time the use of term *Named Entity*, a named object of interest in the text [7]. However, five years prior to MUC6 Lisa F. Rau at the Seventh IEEE Conference on Artificial Intelligence Applications presented a paper describing a system to extract and recognize “Company” names which may be regarded as first in this field and was based on heuristics and rules [8]. Very few publications were there from 1991 leading to MUC6. At the MUC6 the need of the semantic identification of people, organizations and localizations, as well as numerical expressions, such as, time and quantities, was observed. In MUC6 names of entities are annotated using mark-up tags ‘ENAMEX’ and ‘NUMEX’ for entity name expression’ and numeric expression’ respectively.

Many dedicated tracks have been organized for NER tasks since the MUC6 conference and this upsurge has led to many quality publications in the area. After that subsequent 7th Message Understanding Conference (MUC7) [9] expanded the scope of named entity with inclusion of date and time entities. The task has progressed in time from simple identification of people, organizations and localizations to the technical domains to recognize domain-specific entities. Particularly Biological Named Entity Recognition (BER) has witnessed a lot of growth in the task of identifying entities from the biological text.

A. Challenges in BER

The task of biological entity recognition (BER) appearing to be straight forward at the first glance remains a challenging task for several reasons. Marrero M. et al. in their paper argue 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 [10]. The difficulty associated with the task of entity recognition in biomedical domain as compared to other domains has been attributed to several factors promulgated 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 or other lexicons to be all-inclusive and up-to-date [11].

The entity names in biomedical name are also usually longer than the names in other domains and detecting their boundaries is comparatively more difficult. There can be overlapping of the entity names making it hard to find which one is right (for example, in “left breast cancer”, what should be tagged “breast cancer”, “left breast cancer” or “left breast”?).

There is further problem of lack of consensus on the name to be used for a given entity and their conceptualization due to rapid progression of biomedical field and also the problem of similar or even identical names and acronyms being used for different concepts [12]. For example, “diabetes”, which is the name for both a disease and a mouse gene thus, having a different meaning in a closely related field [13].

If we look closely at the biomedical literature we observe that abbreviations are very frequently used in it [14]. The problem with use of abbreviations in the biomedical domain is that these can match common English words or have multiple homonyms [15]. The situation is aggravated by the fact that the naming conventions, although there are not many, are usually not followed by authors and they introduce their own abbreviations in their publications [11].

B. Important Works of BER

The three main categories of entity recognition approaches namely: dictionary based, rule-based and learning based approaches have witnessed a lot of research in the recent past.

B(i). Dictionary-Based Approach

Dictionary-based approach is considered as the fundamental approach of identifying entity mentions in text by using terminological resources, such as, 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, for example, they reported that if straightforward string matching was used, name variations could account for up to 79% of the missing genes [16]. They reported that the overall recall was only 36.2% when they experimented with mouse gene names. They observed that using different numerals (e.g. *syt4* and *syt iv*), “punctuation” variation (e.g. *bmp-4* and *bmp4*) or different transcriptions of Greek letters (e.g. *iga* and *ig alpha*) were the most frequent causes of the gene name recognition failures.

The researchers have witnessed lexical variants, synonymy when a concept is represented with several entities and homonymy when an entity has several meanings, and posing difficulties in recognizing entities using dictionary based approach [17]. The major issue with the use of such vocabularies of terms is that it is not possible to have exhaustive 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. Even with the use of exclusive “well-formed” and approved names by the scientists as an ideal scenario in the future, there still remains a challenge how to deal with a huge number of documents containing “legacy” and ad-hoc terms [17].

The lower precision and recall and other reported problems in the dictionary-based approaches led to adoption of many enhancements to these approaches. Generating spelling variations to get the terms for a biomedical resource and then augmenting terms to the primary lists is one example of such an enhancement [18]. Then the expanded list can be used to do exact string matching. There are other works which use algorithms, such as, probabilistic algorithm or BLAST algorithm to find variants of terms from raw text to augment the dictionary lists [19, 20].

Although there are many of these enhancements have been attempted, yet dictionary-based methods are frequently used in combination with more advanced NER approaches.

B(ii). Rule-Based Approach

The entity recognition systems in the early years of NER task were predominately based on rule-based approaches [21]. The FASTUS system, which participated in the MUC-6 evaluation, uses manually designed regular expressions to extract names of entities [22]. LaSIE (Large Scale Information Extraction) [23] and LaSIE II [24] were two systems by The University of Sheffield NLP group that participated in 6th and 7th Message Understanding Conferences (MUC-6 and MUC-7) which used lookup lists of reference entity names and grammar rules for the identification of entities. The PASTA (Protein Active Site Template Acquisition) and EMPATHIE (Enzyme and Metabolic Pathways Information Extraction) systems were another rule based systems built for extraction of information from scientific abstracts and journal papers about enzymes and protein structure using context free grammars [25, 26]. Fukuda et al. proposed a method of identifying protein names from biological text by utilizing pattern-based rules [27]. In their work on biological named entity recognizer, Narayanaswamy et al. exploited the surface clues and simple linguistic and domain knowledge in identifying chemical names [28].

There are few efforts to consolidate upon the simpler ways of rule based approach by applying additional mechanisms. One of such efforts was the use of protein collocations extracted from a biological corpus to enhance the performance of protein name recognizer [29]. Protein name extraction system named Yapex improves recognition of protein names with result of syntactic parsing for determining entity boundaries [30].

Of late the emphasis has shifted towards using learning based approaches or using them in combination with, dictionary and rule-based approaches due to restricted

portability of dictionary and rule-based approaches in transferring across other domains [3].

B(ii). Learning-Based Approaches

Learning based approach of named entity extraction is primarily based on using statistical methods. In this approach some kind of machine learning algorithm is applied which makes use of automatically learning by using positive and negative training examples for the task with use of distinctive features associated with examples [31]. Nadeau et al. described three categories of features namely **word-level features**, **list look-up features**, and **document and corpus features**, as the features that are useful in the NER task using learning methods [5]. The feature are usually selected depending upon the task and what features do we select has a bearing on the performance of NER systems.

We find in the literature works reported on all three types of learning algorithms namely **supervised learning**, **semi-supervised learning** and **unsupervised learning**. Out of these three, supervised learning is considered to be the dominant approach for the named entity recognition task.

B(iii)(a). Supervised Learning

Support Vector Machines (SVM), Decision Trees, Hidden Markov Models (HMM), Maximum Entropy Models (ME), and Conditional Random Fields (CRF), all are frequently applied supervised learning techniques for NER task.

Support Vector Machines (SVM) has been used by many researchers in their NE works [32-38]. Various works have been reported using **Hidden Markov Models (HMM)** learning technique in the context of NER task [39-42]. The frameworks for biomedical NER system based on **CRF** are often reported to be better performing [43]. The CRF was used frequently among the highly ranked systems on the BioCreAtIve gene mention recognition tasks [44]. Kazama et al. used Wikipedia as external knowledge to improve CRF-based named entity recognition [45]. There are other machine learning methods that are also common for NER task, such as, Maximum Entropy model, Perceptron algorithms, Naïve Bayes, Decision Trees etc. and many works have been done using these.

We need large amount of training data which is mostly manually annotated in the supervised learning methods and this requires a lot of cost and time investments. In the recent past there have been attempts to automatic generate training data for the NER task by using bootstrapping and other semi-supervised statistical techniques.

B(iii)(b). Semi-supervised Learning (SSL)

The dependence of supervised learning on the training data is lessened by making use of both labeled and unlabeled data for the learning process in the semi-supervised (or “weakly supervised”) learning technique. **Bootstrapping** or self-training is the main technique of semi-supervised learning which requires a small degree of supervision. Bootstrapping method in SSL became quite popular and many NER methods are using bootstrapping approaches.

Brin et al. used a SSL technique which exploits the lexical features implemented by regular expressions in order to generate lists of book titles paired with book authors from the

World Wide Web [46]. They achieve this by using the duality between sets of patterns and relations to grow the target relation starting from a small sample.

Riloff and Jones [47] in their significant work presented mutual bootstrapping, which starts with a small set of seed entity names that are located in the corpus and their contexts are pruned to generalize extraction patterns. The ranking is done of these patterns using confidence score and new examples are found out using. This idea was further extended by Thelan et al. by using collective evidence from a large set of extraction patterns to give better results than earlier one [48].

In another variant of mutual bootstrapping, A. Cucchiarelli and Velardi used seeds from existing NER systems for starting examples [49]. They relied on syntactic relations (e.g., subject-object) to find better contextual evidence about the entities. M. Pasca et al. work was also motivated by method of mutual bootstrapping [50]. Their approach was innovative in the sense that they used pattern generalization by semantic class similarity. Another contribution of theirs was that they were able to show that beginning with a small set of (10 examples) it is possible to apply this SSL technique on large corpora (100 million web documents) and still have high precision (88%) in generating millions of facts.

Nadeau *et al.* discussed in his paper that experiments in semi-supervised NERC have performances comparable with baseline supervised approaches [5]. Vlachos and Gasperin et al. presented a bootstrapping method in which they used new test set for the task constructed on an annotation scheme which separates gene names from gene mentions, allowing a more reliable annotation [51].

Another recent work using bootstrapping is reported by Olsson [52] in which he uses a method that is focused on the creation of annotated data, as opposed to the creation of classifiers, and the eventual result of the method is a corpus of marked up textual documents. Knopp [53] presented a bootstrapping approach based on Wikipedia's category system to classify the NERs contained in HeiNER, a multi-lingual NE corpus prepared in earlier work by Wentland in 2008 and he was able to classify more than two million named entities to improve the resource's quality.

The very evident limitation of the bootstrapping approach is propagating the error once it has been introduced. Clearly once the noisy patterns or entities are surfaced in the bootstrapping process there is a sharp decline in the performance due to the very nature of the process. Another problem is that when low frequency classes of entities are there, inadequate contextual information hinders the pattern generalization [47].

B(iii)(c). Unsupervised Learning (USL)

Unsupervised learning methods make decisions on a large unannotated corpus or data. For the task of NER the main approach in unsupervised learning is clustering technique. There are other unsupervised approaches, such as, similarity based functions and distribution statistics. Fundamentally, the techniques are based on lexical resources, on lexical patterns and on statistics on a large unlabeled data.

The following are some prominent works using USL for NER task:

In an unsupervised learning method in the domain of NER, Alfonseca et al. [54] presented a work of labeling an input word with an appropriate NE type taken from WordNet. In another work of USL, Evans [55] 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. The unsupervised approach he used was extracting sequences of capitalized words appearing in a document that are likely to be entity names and then submitting queries to a search engine to search possible hypernyms of the capitalized sequences together with Hearst patterns..

Shinyama et al. [56] in their work of unsupervised learning discovered a strong connection between being a named entity and appearing in multiple news sources at the same time and presented a technique for detecting rare named entities in an unsupervised manner. Although the accuracy was reported high for this approach but recall was not that good.

There were two significant works on using unsupervised learning for named entities considering them multi-word units (MWU). Silva et al. used mutual information measures and the frequency of words to identify n-grams from corpus as the possible entities on the assumption that named entities as multi word occur more often together than disjointedly [57]. The other work which built upon their idea and extended further by applying similar method but with Web data was reported by Downey et al. [58].

There have many works reported in the literature which do use combination of different approaches to enhance the performance. Many efforts on for performing biomedical NER are often quite successful at using dictionary- or rule-based methods with the statistical methods. In a work on medical entity recognition, Abacha et al. [59] observed that hybrid approaches using domain knowledge and statistical methods of machine learning did best. Sasaki et al. [60] combined a dictionary-based approach with part-of-speech tagging to extract known protein names in parallel. There are number of other successful works on hybrid NER systems in the biomedical domain, therefore underlining the importance of combining the different approaches for the specific problem in hand for better performance.

C. Evaluation of ER Systems

Thorough evaluation of any system is crucial to its growth. The NER task is generally evaluated by comparing the outputs of an NER system with that of gold standard i.e. manual annotations on the same dataset. Many methods are in use to judge systems based on their capability to mark a text like an expert. Due to upsurge in the task of biomedical named entity recognition in recent past, many efforts have taken place to evaluate the NER systems based on community-wide evaluation.

MUC conferences were a series of Message Understanding Conference Evaluations. In MUC events [61], the scoring for the system is done on two parameters: ability of the system to detect the correct type (TYPE); and ability of the system to detect the exact text (TEXT). For both of these parameters of TYPE and TEXT, there are three measures: (COR) the number of correct answers; (ACT) the number of

guesses of actual system; and (POS) the number of possible entities in the solution. The micro-averaged f-measure (MAF) is the final MUC score, which is the harmonic mean of precision and recall calculated over all entity spaces on both parameters.

IREX [62] was an evaluation-based project for information retrieval and information extraction in Japanese and it used a simple scoring mechanism. An exact match with the corresponding entity in the gold standard makes the named entity to be correct in this task and hence this evaluation is classified in the “exact-match evaluation” category along with similar evaluations. Another such evaluation task was CONLL shared task [63] in 2003, which also compared systems on the micro-averaged f-measure (MAF). In this system as in IREX, precision is the percentage of named entities detected by the system that are correct and recall is the percentage of named entities present in the corpus that are found by the system.

In the Automatic Content Extraction (ACE) entity detection and tracking (EDT) task [], the evaluation process is a complex one. The task definition in the ACE evaluation is more elaborate in the sense that all mentions of an entity, whether a name, a description, or a pronoun, are to be identified and collected into equivalence classes based on reference to the same entity. ACE includes practical coreference resolution and also deals with various evaluation issues such as, partial match, wrong type etc.

Although evaluation tasks, such as ACE, with more elaborate mechanisms and wide coverage of the problem may be considered more powerful, however, these complex methods make error analysis and evaluation hard.

III. CONCLUSION

In this paper we surveyed the area of BER and looked at the different works done for the task by many researchers using multiple approaches. A lot of work has been done in BER by using the simple dictionary based approach and further to overcome the limitations found in this approach lot many improvements have been tried. 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 BER and we find many such works reported for the same.

However, we notice that due to restricted portability of dictionary and rule-based approaches in transferring across other domains, the focus in recent times has shifted towards using learning based approaches or using them in combination with, dictionary and rule-based approaches. We find in the literature works reported on different types of learning algorithms though supervised learning is considered to be the dominant learning method for the entity recognition task. There have been also many useful hybrid ER systems reported in the biomedical domain, therefore highlighting the importance of combining the different approaches for better results. We also looked at the different evaluation tasks carried out in BER, while observing that more elaborate mechanisms and wide coverage evaluation systems adds to the complexity making error analysis and evaluation harder.

REFERENCES

- [1]. Holzinger, A., Schantl, J., Schroettner, M., Seifert, C., & Verspoor, K. (2014). Biomedical text mining: State-of-the-art, open problems and future challenges. In *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics* (pp. 271-300). Springer Berlin Heidelberg.
- [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]. Simpson, M. S., & Demner-Fushman, D. (2012). Biomedical text mining: A survey of recent progress. In *Mining text data* (pp. 465-517). Springer US.
- [4]. Zhang, Z., Cohn, T., & Ciravegna, F. (2013). Topic-oriented words as features for named entity recognition. In *Computational Linguistics and Intelligent Text Processing* (pp. 304-316). Springer Berlin Heidelberg.
- [5]. Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1), 3-26.
- [6]. Grishman, R., & Sundheim, B. (1996, August). Message Understanding Conference-6: A Brief History. In *COLING* (Vol. 96, pp. 466-471).2q
- [7]. Grishman, R., & Sundheim, B. (1996, August). Message Understanding Conference-6: A Brief History. In *COLING* (Vol. 96, pp. 466-471).2q
- [8]. Rau, L. F. (1991, February). Extracting company names from text. In *Artificial Intelligence Applications, 1991. Proceedings., Seventh IEEE Conference on* (Vol. 1, pp. 29-32). IEEE.
- [9]. Chinchor, N., & Marsh, E. (1998, July). Muc-7 information extraction task definition. In *Proceeding of the seventh message understanding conference (MUC-7), Appendices* (pp. 359-367).
- [10]. Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J., & Gómez-Berbís, J. M. (2013). Named Entity Recognition: Fallacies, challenges and opportunities. *Computer Standards & Interfaces*, 35(5), 482-489.
- [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]. Leser, U.; and J. Hakenberg. (2005) What makes a gene name? Named entity recognition in the biomedical literature. *Briefings in Bioinformatics* 6, pp. 357-369.
- [13]. Chen, L., H. Liu and C. Friedman (2005). "Gene name ambiguity of eukaryotic nomen-clatures." *Bioinformatics* 21(2): 248-256.
- [14]. Okazaki, N., Ananiadou, S., & Tsujii, J. I. (2010). Building a high-quality sense inventory for improved abbreviation disambiguation. *Bioinformatics*, 26(9), 1246-1253.
- [15]. Liu, H., Aronson, A. R., & Friedman, C. (2002). A study of abbreviations in MEDLINE abstracts. *Proceedings of the AMIA Symposium*, 464–468.
- [16]. Tuason, O., Chen, L., Liu, H., Blake, J. A., & Friedman, C. (2004). Biological nomenclatures: a source of lexical knowledge and ambiguity. In *Pac Symp Biocomput* (pp. 238-249).
- [17]. Krauthammer, M., & Nenadic, G. (2004). Term identification in the biomedical literature. *Journal of biomedical informatics*, 37(6), 512-526.
- [18]. 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.
- [19]. Tsuruoka, Y., & Tsujii, J. I. (2003, July). Probabilistic term variant generator for biomedical terms. In *Proceedings of*

- the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval* (pp. 167-173). ACM.
- [20]. Krauthammer, M., Rzhetsky, A., Morozov, P., & Friedman, C. (2000). Using BLAST for identifying gene and protein names in journal articles. *Gene*, 259(1), 245-252.
- [21]. Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1), 3-26.
- [22]. Appelt, D. E., Hobbs, J. R., Bear, J., Israel, D., Kameyama, M., Martin, D., ... & Tyson, M. (1995, November). SRI International FASTUS system: MUC-6 test results and analysis. In *Proceedings of the 6th conference on Message understanding* (pp. 237-248). Association for Computational Linguistics.
- [23]. Gaizauskas, R., Humphreys, K., Cunningham, H., & Wilks, Y. (1995, November). University of Sheffield: description of the LaSIE system as used for MUC-6. In *Proceedings of the 6th conference on Message understanding* (pp. 207-220). Association for Computational Linguistics.
- [24]. Humphreys, K., Gaizauskas, R., Azzam, S., Huyck, C., Mitchell, B., Cunningham, H., & Wilks, Y. (1998, April). University of Sheffield: Description of the LaSIE-II system as used for MUC-7. In *Proceedings of the Seventh Message Understanding Conferences (MUC-7)*.
- [25]. Humphreys, K., Demetriou, G., & Gaizauskas, R. (2000, January). Two applications of information extraction to biological science journal articles: Enzyme interactions and protein structures. In *Pac Symp Biocomput* (Vol. 5, pp. 505-516).
- [26]. Gaizauskas, R., Demetriou, G., Artymiuk, P. J., & Willett, P. (2003). Protein structures and information extraction from biological texts: the PASTA system. *Bioinformatics*, 19(1), 135-143.
- [27]. Fukuda K, Tsunoda T, Tamura A, Takagi T (1998) Toward information extraction: identifying protein names from biological papers. In: Proceedings of the Pacific Symposium on Biocomputing'98 (PSB'98). pp. 707–718.
- [28]. Narayanaswamy, M., Ravikumar, K. E., Vijay-Shanker, K., & Ay-shanker, K. V. (2003). A biological named entity recognizer. In *Pac Symp Biocomput* (p. 427).
- [29]. Hou, W. J., & Chen, H. H. (2003, July). Enhancing performance of protein name recognizers using collocation. In *Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine-Volume 13* (pp. 25-32). Association for Computational Linguistics.
- [30]. Franzén, K., Eriksson, G., Olsson, F., Asker, L., Lidén, P., & Cöster, J. (2002). Protein names and how to find them. *International journal of medical informatics*, 67(1), 49-61.
- [31]. Zhang, Z. (2013). Named entity recognition: challenges in document annotation, gazetteer construction and disambiguation.
- [32]. Ekbal, A., & Bandyopadhyay, S. (2010). Named entity recognition using support vector machine: A language independent approach. *International Journal of Electrical and Electronics Engineering*, 4(2), 155-170.
- [33]. Isozaki, H., & Kazawa, H. (2002, August). Efficient support vector classifiers for named entity recognition. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1* (pp. 1-7). Association for Computational Linguistics.
- [34]. Kazama, J. I., Makino, T., Ohta, Y., & Tsujii, J. I. (2002, July). Tuning support vector machines for biomedical named entity recognition. In *Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain-Volume 3* (pp. 1-8). Association for Computational Linguistics.
- [35]. Mitsumori, T., Fation, S., Murata, M., Doi, K., & Doi, H. (2005). Gene/protein name recognition based on support vector machine using dictionary as features. *BMC bioinformatics*, 6(Suppl 1), S8.
- [36]. Takeuchi, K., & Collier, N. (2005). Bio-medical entity extraction using support vector machines. *Artificial Intelligence in Medicine*, 33(2), 125-137.
- [37]. Yamamoto, K., Kudo, T., Konagaya, A., & Matsumoto, Y. (2003, July). Protein name tagging for biomedical annotation in text. In *Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine-Volume 13* (pp. 65-72). Association for Computational Linguistics.
- [38]. Asahara, M., & Matsumoto, Y. (2003, May). Japanese named entity extraction with redundant morphological analysis. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1* (pp. 8-15). Association for Computational Linguistics.
- [39]. Morwal, S., Jahan, N., & Chopra, D. (2012). Named entity recognition using hidden Markov model (HMM). *Int. J. Nat. Lang. Comput. (IJNLC)*, 1(4), 15-23.
- [40]. GuoDong, Z., & Jian, S. (2004, August). Exploring deep knowledge resources in biomedical name recognition. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications* (pp. 96-99). Association for Computational Linguistics.
- [41]. Ponomareva, N., Pla, F., Molina, A., & Rosso, P. (2007). Biomedical named entity recognition: a poor knowledge HMM-based approach. In *Natural Language Processing and Information Systems* (pp. 382-387). Springer Berlin Heidelberg.
- [42]. N. Collier, C. Nobata, and J.-i. Tsujii (2000). Extracting the names of genes and gene products with a hidden Markov model. In *Proceedings of the 18th Conference on Computational Linguistics -Volume 1*, pages 201–207.
- [43]. Simpson, M. S., & Demner-Fushman, D. (2012). Biomedical text mining: A survey of recent progress. In *Mining text data* (pp. 465-517). Springer US.
- [44]. L. Smith, L. Tanabe, R. Johnson nee Ando, C.-J. Kuo, I.-F. Chung, C.-N. Hsu, Y.-S. Lin, R. Klinger, C. Friedrich, K. Ganchev, M. Torii, H. Liu, B. Haddow, C. Struble, R. Povinelli, A. Vlachos, W. Baumgartner, L. Hunter, B. Carpenter, R. Tzong-Han Tsai, H.-J. Dai, F. Liu, Y. Chen, C. Sun, S. Katrenko, P. Adriaans, C. Blaschke, R. Torres, M. Neves, P. Nakov, A. Divoli, M. Mana Lopez, J. Mata, and W. Wilbur (2008). Overview of BioCreAtIve II: Gene mention recognition. *Genome Biology*, 9(Suppl 2):S2.
- [45]. Kazama, J. and Torisawa, K. (2007). Exploiting Wikipedia as External Knowledge for Named Entity Recognition. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Prague, Czech, pp.698-707.
- [46]. Brin, S. (1999). Extracting patterns and relations from the world wide web. In *The World Wide Web and Databases* (pp. 172-183). Springer Berlin Heidelberg.
- [47]. Riloff, E., & Jones, R. (1999, July). Learning dictionaries for information extraction by multi-level bootstrapping. In *AAAI/IAAI* (pp. 474-479).
- [48]. Thelen, M., & Riloff, E. (2002, July). A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10* (pp. 214-221). Association for Computational Linguistics.

- [49].Cucchiarelli, A., & Velardi, P. (2001). Unsupervised named entity recognition using syntactic and semantic contextual evidence. *Computational Linguistics*, 27(1), 123-131.
- [50].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).
- [51].Vlachos, A., & Gasperin, C. (2006, June). Bootstrapping and evaluating named entity recognition in the biomedical domain. In *Proceedings of the HLT-NAACL BioNLP Workshop on Linking Natural Language and Biology* (pp. 138-145). Association for Computational Linguistics.
- [52].Olsson, F. (2008). *Bootstrapping named entity annotation by means of active machine learning: a method for creating corpora* (Doctoral dissertation, University of Gothenburg).
- [53].Knopp, J. 2011. Extending a Multilingual Lexical Resource by Bootstrapping Named Entity Classification using Wikipedia's Category System. In *Proceedings of the 5th International Workshop On Cross Lingual Information Access*, at IJCNLP2011. Chiang Mai, Thailand, pp.35-43.
- [54].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).
- [55].Evans, R., & Street, S. (2003). A framework for named entity recognition in the open domain. *Recent Advances in Natural Language Processing III: Selected Papers from RANLP, 260*, 267-274.
- [56].Shinyama, Y., & Sekine, S. (2004, August). Named entity discovery using comparable news articles. In *Proceedings of the 20th international conference on Computational Linguistics* (p. 848). Association for Computational Linguistics.
- [57].Silva, J. F., Kozareva, Z., Noncheva, V., & Lopes, G. P. (2004). Extracting named entities. a statistical approach. *Proceedings of the XIth Conference sur le Traitement des Langues Naturelles—TALN, 19–22 Avril, Fez, Marroco*, 347-351.
- [58].Downey, D., Broadhead, M., & Etzioni, O. (2007, January). Locating Complex Named Entities in Web Text. In *IJCAI* (Vol. 7, pp. 2733-2739).
- [59].Abacha, A. B., & Zweigenbaum, P. (2011, June). Medical entity recognition: A comparison of semantic and statistical methods. In *Proceedings of BioNLP 2011 Workshop* (pp. 56-64). Association for Computational Linguistics.
- [60].Sasaki, Y., Tsuruoka, Y., McNaught, J., & Ananiadou, S. (2008). How to make the most of NE dictionaries in statistical NER. *BMC bioinformatics*, 9(Suppl 11), S5.
- [61].Chinchor, Nancy (1999) Overview of MUC-7/MET-2. *Proc. Message Understanding Conference MUC-7*.
- [62].Sekine, S., & Isahara, H. (2000, May). IREX: IR & IE Evaluation Project in Japanese. In *LREC* (pp. 1977-1980).
- [63].Tjong Kim Sang, E. F., & De Meulder, F. (2003, May). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4* (pp. 142-147). Association for Computational Linguistics.
- [64].Doddington, G. R., Mitchell, A., Przybocki, M. A., Ramshaw, L. A., Strassel, S., & Weischedel, R. M. (2004, May). The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation. In *LREC*.