Journal of Biomedical Informatics 46 (2013) 1088–1098

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts ☆

Shaodian Zhang *, Noémie Elhadad

Department of Biomedical Informatics, Columbia University, 622 W. 168th Street, VC-5, New York, NY 10032, USA

ARTICLE INFO

Article history:
Received 20 February 2013
Accepted 7 August 2013
Available online 15 August 2013

Keywords:
Natural language processing
Named entity recognition
Distributional semantics
UMLS
Chunking

ABSTRACT

Named entity recognition is a crucial component of biomedical natural language processing, enabling information extraction and ultimately reasoning over and knowledge discovery from text. Much progress has been made in the design of rule-based and supervised tools, but they are often genre and task dependent. As such, adapting them to different genres of text or identifying new types of entities requires major effort in re-annotation or rule development. In this paper, we propose an unsupervised approach to extracting named entities from biomedical text. We describe a stepwise solution to tackle the challenges of entity boundary detection and entity type classification without relying on any handcrafted rules, heuristics, or annotated data. A noun phrase chunker followed by a filter based on inverse document frequency extracts candidate entities from free text. Classification of candidate entities into categories of interest is carried out by leveraging principles from distributional semantics. Experiments show that our system, especially the entity classification step, yields competitive results on two popular biomedical datasets of clinical notes and biological literature, and outperforms a baseline dictionary match approach. Detailed error analysis provides a road map for future work.

1. Introduction

An overwhelming amount of health and biomedical text is becoming available with the recent adoption of electronic health records, the growing number of biomedical publications, and the exploding prevalence of health information online. At the same time, in the research community, significant efforts have been devoted to creating standard terminologies and knowledge bases hence facilitating extraction of information from and reasoning over raw data. The bottleneck of biomedical information processing thus has shifted from where to collect data and resources to how to make use of the knowledge resources and build scalable models to process large amounts of text. Since much of the data is recorded in narrative and unstructured form, like in clinical notes and biomedical publications, the quality of basic natural language processing (NLP) tools has a critical impact on the performance of higher-level tasks such as information retrieval, information extraction, and knowledge discovery. Biomedical named-entity recognition (BM-NER),[1] sometimes referred to as biomedical concept identification or concept mapping, is a key step in biomedical language processing: terms (either single words or multiple words) of interest are identified and mapped to a pre-defined set of semantic categories. Examples of BM-NER systems include extracting clinical information from radiology reports [1–3], identifying diseases and drug names in discharge summaries [4–6], detecting gene and protein mentions in biomedical paper abstracts [7–9].

In the general domain, named-entity recognition (NER) focuses on identifying names of persons, locations, and organizations in news articles, reports, and even tweets. Thanks to the availability of annotated corpora, supervised learning methods have been widely adopted and prevail unsupervised ones. Such state-of-the-art NER systems have achieved performance as high as human annotators [10,11]. On their side, BM-NER are getting better with the advant of more annotated corpora to learn from. Recent supervised systems could efficiently find gene names and clinical problems from certain type of texts with above 0.8 F score [6,12,13,14]. Traditional ways of tackling BM-NER range from dictionary matching, heuristic rules, to supervised Hidden Markov Models (HMMs)/Conditional Random Fields (CRFs)-based sequence labeling. The first two approaches do not require training data, but usually involve ad hoc rules and assumptions that may

* Corresponding author.
E-mail addresses: shaodian@dbmi.columbia.edu (S. Zhang), noemie@dbmi.columbia.edu (N. Elhadad).

[1] In this paper, without further explanation, "biomedical entity", "entity", and "named entity" are all referring to biomedical entities.

1532-0464/$ - see front matter © 2013 The Authors. Published by Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2013.08.004

limit the type of entities and texts to which they could apply. CRF-based labelers have yielded high performance in sequence learning tasks, and are the state of the art for some biological and medical entity recognition tasks. However, the supervised nature of CRF relies on a fairly large amount of training data which must be annotated by humans. As a result, it is only applicable in a limited number of settings.

In this paper, we provide a stepwise unsupervised solution to biomedical named-entity recognition. Our approach does not rely on hand-built rules or examples of annotated entities, so it can be adapted to different semantic categories and text genres easily. Instead of supervision, the entity recognition leverages terminologies, shallow syntactic knowledge (noun phrase chunking), and corpus statistics (inverse document frequency and context vectors). Experimental results demonstrate that our method yields competitive results on two popular datasets of different genres, clinical notes and biomedical literature, respectively, and different corresponding entity types. An implementation of the methods described in this paper is available at http://people.dbmi.columbia.edu/~szhang/ner.html.

2. Background

There are two main steps of named entity recognition: detecting boundaries of entity mentions and classifying the mentions into predefined semantic categories. The task of entity linking or concept normalization, that is linking a term to a unique concept identifier in a terminology for instance is not typically part of NER, and as such is not the focus of this paper. With sequence labeling models like Hidden Markov Models (HMMs) and Conditional Random Fields (CRFs), the two tasks could be jointly handled taking advantage of the Markov property which models transitions between labels [15,16]. In an unsupervised framework, however, boundary detection and entity classification are typically conducted separately [17]. In this section we review related work from two perspectives, unsupervised named entity recognition and biomedical named entity recognition, and direct the reader to existing reviews of supervised approaches for NER in the general domain [17].

2.1. Unsupervised named entity recognition

The NLP community has invested a lot of efforts in unsupervised NER. Early work [18,19] relies on heuristic rules and lexical resources such as WordNet [20]. More recently, Alfonseca and Manandhar formulate named entity classification as a word sense disambiguation task and cluster words based on the words with which they co-occur frequently in online search results [21]. The context word frequency vector, which represents the semantics of words to be classified, is called "signature." Nadeau et al. present a system of retrieving entity lists by web page wrapper, followed by disambiguation through heuristic rules [22]. Sekine and Nobata give definitions and rule-based taggers for 200 categories of entities, as well as a standard taxonomy of general entities [23]. Shinyama and Sekine observe that named entities often appear synchronously in several news articles, whereas common nouns do not [24]. Exploiting this characteristic, they successfully obtained rare named entities with 90% accuracy just by comparing time series distributions of a word in two newspapers. This technique can be useful in combination with other NER methods.

The second category of methods is relatively new, and is essentially weakly supervised instead of unsupervised. Such methods use a bootstrapping-like technique to strengthen the models, starting from small sets of seed data or rules. The first notable work is done by Collins and Singer, in which a small set of handcrafted rules are predefined as seed rules [25]. The system iteratively labels the dataset based on current rules, and induces more rules with high precisions on found entities. Riloff and Jones introduce mutual bootstrapping that consists of growing a set of entities and a set of contexts in turn [26]. Several improvements and extensions were later proposed following this bootstrapping approach [27–29]. It is noteworthy that previous works in this category focus only on entity classification, which assume that the named entities have already been correctly extracted from the text.

It is interesting that in many ways, unsupervised named entity recognition systems are enlightened by previous works in word sense disambiguation, especially in classifying extracted entities. On the one hand, the bootstrapping framework in [25] was initially used by [30] for word sense disambiguation; on the other hand, the idea of classifying entities based on their context signatures [21] is also similar with distributional methods in word sense disambiguation [31], in which contexts of mentions are used to determine word senses.

2.2. Biomedical named entity recognition

There are two major reserach directions in BM-NER: finding gene, protein, and related biological or genetic terms, as well as finding disease, drug names, and other medical terms. We use biological NER and medical NER to denote these two research sub-domains respectively. The early NER systems in both fields are typically rule-based or lexicon-based [1,7,32–36], several of which are widely applied. MedLEE is a general natural language processor for clinical texts, encoding and mapping terms to a controlled vocabulary [1]; GENIES is a system extracting molecular pathways from journal articles, which is modified from MedLEE [35]; EDGAR is a natural language processing system that extracts information about drugs and genes relevant to cancer from the biomedical literature [34]; AbGene is one of the most successful NER systems for gene and protein [7]; MetaMap, developed by National Library of Medicine (NLM), is a tool discovering UMLS Metathesaurus concepts referred to in text [36]. Many of these systems highly resort to lexical knowledge resources such as GO [37] and UMLS [38]. More recently cTAKES provides concept identification and normalization to UMLS in clinical texts [39].

Recent years have witnesses the rise of data-driven methods in biomedical named entity recognition with the availability of annotated data. In biological NER, the release of the GENIA corpus [40] has pushed forward related research using various supervised learning models, including Support Vector Machines (SVMs) [41–43], Hidden Markov Models (HMMs) [44], and Conditional Random Fields (CRFs) [8,45]. The shared task of BioNLP/NLPBA 2004 used GENIA as dataset [46], and 9 teams submitted their NER systems to the event. In the first BioCreAtIvE challenge [47], gene mention identification was the first subtask of task1 [9]. Such shared tasks and workshops continued every year with new challenges, advancing the field with related information extraction tasks such as gene normalization[48] and bio-event extraction[49]. So far, state-of-the-art systems for these datasets are mostly supervised ones based on SVM [41] and CRF [8,45].

In the medical domain, the first publicly available corpus for NER evaluation was created in the i2b2 challenge 2010 [6]. 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[50]. Before the availability of i2b2 corpus, recent research also focus on evaluation on, extension to, and comparison with MetaMap and its predecessor MMTx. Meystre and Haug evaluate MMTx with a automatically retrieved clinical problem list [51]. Abacha and Zweigenbaum make modifications to MetaMap, and compare MetaMap with statistical based methods like CRF and SVMs[12,52]. Patrick et al. implement a fuzzy matcher which better maps terms to UMLS concepts [53]. Before