Artificial Intelligence in Medicine 65 (2015) 145–153

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

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

Boosting drug named entity recognition using an aggregate classifier

Ioannis Korkontzelos[a,*], Dimitrios Piliouras[a], Andrew W. Dowsey[b,c], Sophia Ananiadou[a]

[a] National Centre for Text Mining (NaCTeM), School of Computer Science, The University of Manchester, Manchester Institute of Biotechnology, 131 Princess Street, Manchester M1 7DN, United Kingdom
[b] Centre for Endocrinology and Diabetes, Institute of Human Development, Faculty of Medical and Human Sciences, The University of Manchester, Manchester, United Kingdom
[c] Centre for Advanced Discovery and Experimental Therapeutics (CADET), The University of Manchester and Central Manchester University Hospitals NHS Foundation Trust, Manchester Academic Health Sciences Centre, Oxford Road, Manchester M13 9WL, United Kingdom

ARTICLE INFO

Keywords:
Named entity annotation sparsity
Gold-standard vs. silver-standard annotations
Named entity recogniser aggregation
Genetic-programming-evolved string-similarity patterns
Drug named entity recognition

ABSTRACT

Objective: Drug named entity recognition (NER) is a critical step for complex biomedical NLP tasks such as the extraction of pharmacogenomic, pharmacodynamic and pharmacokinetic parameters. Large quantities of high quality training data are almost always a prerequisite for employing supervised machine-learning techniques to achieve high classification performance. However, the human labour needed to produce and maintain such resources is a significant limitation. In this study, we improve the performance of drug NER without relying exclusively on manual annotations.
Methods: We perform drug NER using either a small gold-standard corpus (120 abstracts) or no corpus at all. In our approach, we develop a *voting system* to combine a number of heterogeneous models, based on dictionary knowledge, gold-standard corpora and silver annotations, to enhance performance. To improve recall, we employed genetic programming to evolve 11 regular-expression patterns that capture common drug suffixes and used them as an extra means for recognition.
Materials: Our approach uses a dictionary of drug names, i.e. DrugBank, a small manually annotated corpus, i.e. the pharmacokinetic corpus, and a part of the UKPMC database, as raw biomedical text. Gold-standard and silver annotated data are used to train maximum entropy and multinomial logistic regression classifiers.
Results: Aggregating drug NER methods, based on gold-standard annotations, dictionary knowledge and patterns, improved the performance on models trained on gold-standard annotations, only, achieving a maximum *F*-score of 95%. In addition, combining models trained on silver annotations, dictionary knowledge and patterns are shown to achieve comparable performance to models trained exclusively on gold-standard data. The main reason appears to be the morphological similarities shared among drug names.
Conclusion: We conclude that gold-standard data are not a hard requirement for drug NER. Combining heterogeneous models build on dictionary knowledge can achieve similar or comparable classification performance with that of the best performing model trained on gold-standard annotations.

1. Introduction

Named entity recognition (NER) is the task of identifying members of various semantic classes, such as *persons*, *mountains* and *vehicles* in raw text. In biomedicine, NER is concerned with classes such as *proteins*, *genes*, *diseases*, *drugs*, *organs*, *DNA sequences*, *RNA sequences* and possibly others [1]. Drugs (as pharmaceutical products) are special types of chemical substances highly relevant for biomedical research. A simplistic and naive approach to NER is to directly match textual expressions found in a relevant lexical repository against raw text. Even though this technique can sometimes work well, often it suffers from certain limitations. Firstly, its accuracy heavily depends on the completeness of the dictionary. However, as terminology is constantly evolving, especially in bio-related disciplines, producing a complete lexical repository is not feasible. Secondly, direct string matching overlooks term ambiguity and variability [2]. On one hand, ambiguous dictionary entries refer to multiple semantic types (term ambiguity), and

[*] Corresponding author. Tel.: +44 0161 306 3094.
E-mail addresses: Ioannis.Korkontzelos@manchester.ac.uk (I. Korkontzelos), piliourd@cs.man.ac.uk (D. Piliouras), Andrew.Dowsey@manchester.ac.uk (A.W. Dowsey), Sophia.Ananiadou@manchester.ac.uk (S. Ananiadou).

http://dx.doi.org/10.1016/j.artmed.2015.05.007
0933-3657/© 2015 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

therefore contextual information needs to be considered for disambiguation. On the other hand, several slightly different tokens may refer to the same semantic type (term variability). Typically, to address these issues, statistical learning models are deployed for NER.

In such approaches, NER is formalised as a classification task in which an input expression is either classified as an entity or not. Supervised learning methods are reported to achieve superior performance than unsupervised ones, but previously annotated data are essential for training supervised models [2]. Data annotated by human curators are of high quality and guarantee best results in exchange for the cost of manual effort. For these reasons, they are also known as gold-standard data. Due to the cost of manual annotations, corpora for NER are often of limited size and for particular domains.

Drugs are referred to by their chemical name, generic name or brand name. Since the chemical name is typically complex and a brand name may not exclusively identify a drug once the relevant patents expire, a unique non-proprietary name for the active ingredient is devised for standardised scientific reporting and labelling. This generic name is negotiated when the drug is approved for use, as the nomenclature is tightly regulated by the World Health Organization (WHO) and local agencies such as the U.S. Food and Drug Administration (FDA) and the European Medicines Agency. Several criteria are assessed, such as ensuring the drug action fits the naming scheme, ease of pronunciation and translation, and differentiation from other drug names to avoid transcription and reproduction errors during prescription [3]. Since the naming scheme, assessment criteria and cross-border synchrony have developed organically over the years, there is neither a definitive dictionary nor syntax of drug names.

In this study, we investigate methods for achieving high performance in drug name recognition in cases where either very limited or no gold-standard training data is available. Our proposed method employs a *voting system* able to combine predictions from a number of diverse recognisers. Moreover, genetic programming is used to evolve string-similarity patterns based on common suffixes of single-token drug names occurring in the DrugBank database [4]. Subsequently, these patterns are used to compile regular expressions in order to generalise dictionary entries in an effort to increase coverage and tagging accuracy.

We compare the performance of our method with several state-of-the-art NER approaches in recognising manually annotated drug names in the PK corpus [5]. Where no gold-standard data is available, the proposed method is shown to achieve competitive performance. In particular, the performance achieved without gold-standard data is comparable with the performance of the model aware of gold-standard annotations.

The rest of this paper is organised as follows: Section 2 summarises previous work on drug NER and methods for dealing with data sparsity in general NER. Section 3 describes the dictionaries and data used in our experiments, as well as the experimental methodology followed. Sections 4 and 5 present and discuss the experiments and their results. Finally, section 6 concludes the paper.

2. Related work

NER is a large, well-studied field of natural language processing (NLP) [6]. Most publications address it as a supervised task, i.e. the procedure of training a model on annotated data and then applying it to new text. In the past, several evaluation challenges have taken place on recognising entities of the general domain [7–10] as well as scientific domains [2,11,12]. In contrast, research related with Drug NER is limited [13–15]. Very recently, an evaluation challenge that focussed exclusively on drug name recognition and drug-drug interactions has taken place [16].

As a result of the collaborative annotation of a large biomedical corpus project [17], a large-scale biomedical silver standard corpus has been produced. It contains annotations resulting from the harmonisation of named entities (NEs) automatically recognised by five different tools, namely, Whatizit [18], Peregrine [19], GeNO [20], MetaMap [21] and I2E [22]. Apart from names of chemicals and drugs, proteins, genes, diseases and species names were also tagged by these tools in the 174,999 MEDLINE abstracts comprising the corpus. Approximately half a million NE annotations for each semantic category are contained in the resulting harmonised corpus which is publicly available. It has been used for the 2 annotation challenges [23].

Dictionaries and ontologies have been used extensively as the basis to generate patterns and rules for NER. Tsuruoka et al. [24] used logistic regression to learn a string similarity measure from a dictionary, useful for soft-string matching. Kolarik et al. [25] used lexico-syntactic patterns to extract terms. Patterns are similar to the ones introduced in [26] and contain drug names and directly related drug annotation terms found in DrugBank. Then, these patterns were applied to MEDLINE abstracts, to add annotations of pharmacological effects of drugs. Similar methods have also been applied for recognising drug-disease interactions [27] and interactions between compounds and drug-metabolising enzymes [28]. Hettne et al. [29] developed a rule-based method intended for term filtering and disambiguation. They identify names of drugs and small molecules by incorporating several dictionaries such as the UMLS (nlm.nih.gov/research/umls, accessed: 15 April 2015), MeSH (nlm.nih.gov/mesh, accessed: 15 April 2015), ChEBI (www.ebi.ac.uk/chebi, accessed: 15 April 2015), DrugBank (drugbank.ca, accessed: 15 April 2015), KEGG (www.genome.jp/kegg, accessed: 15 April 2015), HMDB (hmdb.ca, accessed: 15 April 2015) and ChemIDplus (chem.sis.nlm.nih.gov/chemidplus, accessed: 15 April 2015). An earlier system, EDGAR [30], extracts genes, drugs and relationships between them using existing ontologies and standard NLP tools such as part-of-speech taggers and syntactic parsers.

A popular means of dealing with data sparsity in NER is to generate data semi-automatically or fully automatically. Although, the resulting data is of lower quality than gold-standard annotations, supervised learners can benefit largely from large volumes of data, since they are based on annotation statistics. Towards the same ultimate goal, our approach aims to overcome the restrictions of data sparsity or unavailability in the biomedical domain. Usami et al. [31] describe an approach for automatically acquiring large amounts of training data from a lexical database and raw text that relies on reference information and coordination analysis. Similarly, noisy training data was obtained by using a few manually annotated abstracts from FlyBase (flybase.org, accessed: 15 April 2015) [32,33]. The approach uses a bootstrapping method and context-based classifiers to increase the number of NE mentions in the original noisy training data. Even though they report high performance, their method requires some minimum curated seed data. Similarly, Thomas et al. [34] demonstrated the potential of distant learning in constructing a fully automated relation extraction process. They produced two distantly labelled corpora for protein–protein and drug–drug interaction extraction, with knowledge found in databases such as IntAct [35] for genes and DrugBank [4] for drugs.

Active learning is a framework that can be used for reducing the amount of human effort required to create a training corpus [36,37]. The most informative samples are chosen from a big pool of human annotations by a maximum likelihood model in an iterative and interactive manner. It has been shown that active learning can often drastically reduce the amount of training data