Comparison of Drug Resources to Build a Dictionary of Medications

Jaya Chaturvedi[1], Angus Roberts[1]
[1]King's College London, United Kingdom

Abstract

Recognition and extraction of medications from electronic health record (EHR) documents is an important sub-task of clinical information extraction, as medications and their prescriptions are central to understanding the course of patient treatment, both in clinical care and in research. Medication recognition is usually considered to be a type of named entity recognition (NER). As with any NER task, lists of terms known to describe the entity of interest are of value in the recognition step. Additional value may be provided if these terms are derived from existing domain knowledge resources, allowing the text to be linked to these resources. However, such resources are not generally designed with NLP in mind. We compare different available medication resources and use these to create gazetteers of medication names for NLP purposes. We compare the terms in these gazetteers to both an existing gazetteer and a set of medication terms manually extracted from EHR documents, and draw conclusions about the suitability of different resources, and the way in which their terms are processed for inclusion in the gazetteers.

Introduction

An important component of Information Extraction (IE) is Named Entity Recognition (NER) [1]. NER is the process of locating named entities mentioned within the text and assigning them to categories such as names of persons, organisations, quantities etc [2]. NER systems serve as crucial pre-processing tools for tasks of IE and text processing and are often the starting points for such processes [3]. A common technique in NER is to use lists of terms used to denote entities, often referred to as gazetteers or entity dictionaries, against which mentions of those entities in the text are matched. NER systems may need extensive gazetteers, and the compilation of such a resource is often mentioned as a bottleneck in the designing of the systems [3]. Building and maintaining high-quality gazetteers is a time-consuming task [4]. Multiple methods have been proposed to overcome this problem by automatically extracting gazetteers from large amounts of texts [5–8]. However, these methods require statistical approaches and induction of patterns in order to extract a high-quality gazetteer [4].

Rather than creating a gazetteer by extracting terms from text, gazetteers may be constructed from knowledge resources such as ontologies, vocabularies and terminologies. Such resources relate terms to some underlying conceptual framework, often providing a unique identifier for each concept described. These knowledge resources provide both a source of terms, helping to overcome the aforementioned bottleneck, and a conceptual framework relating the entities to each other and other knowledge. Gazetteers built from these resources may thus support both the initial recognition of entity mentions within the text and additionally the linking of those mentions to identifiers used to describe entities, as well as to further world knowledge about the entities. It should be noted that the disambiguation of homonyms required for this linking would need to be resolved by any application. Such knowledge-resource backed gazetteers are especially useful in domains where there has been effort to standardise and codify the conceptual knowledge of the domain. Providing a link from textual mentions of entities to the domain knowledge can support downstream tasks and reasoning about these entities. Medicine is such a domain, with large numbers of knowledge resources available. However, there is a mismatch between the needs of NLP and the purposes for which these medical resources were made. None of them have been built for the purposes of NLP, and so the terms contained may not reflect usage in natural language. It is therefore essential to understand how useful these resources can be from an NLP perspective, and what steps need to be taken in order to adapt them for NLP purposes. This is one of the questions that we aim to answer through this paper, and further discuss whether such adaptation of the resources can be automated as well.

In this paper, we focus on an important category of NE in EHR documents, the medication name. Extracting medication information holds considerable value for a variety of research purposes, such as investigating treatment resistance, characterising prescribing patterns and polypharmacy (prescribing more than one medication), and studying the effects as well as side-effects of particular medications on the general health of particular groups of patients [15]. The paper focuses on the creation of a gazetteer using a number of pre-existing lists in order to create a unified dictionary of medication names that incorporates all the nuances of these different pre-existing lists. We describe the initial steps in the development of a gazetteer for medication names from a number of openly available

medication resources. The gazetteer is compared to an existing medication gazetteer, and to a corpus of EHR documents manually labelled for medications.

Methods

There are various medication knowledge resources, which have been constructed and made available for a variety of reasons, such as: use in primary care; use in secondary care; understanding the cost of prescriptions dispensed in the community per year; as well as understanding trends in prescription patterns for individual medication preparations [11]. In order to create a comprehensive medications gazetteer, we decided to incorporate openly available sources of medication names both from primary as well as secondary care, as the texts in secondary care EHRs often refer to medications prescribed by primary care physicians. The datasets that were incorporated were: datasets made available by NHS Digital for primary care prescribing [12], Prescription Cost Analysis (PCA) data which includes primary care prescribing data as well [11], pharmacy data obtained from the South London and Maudsley NHS Foundation Trust (SLaM), and DrugBank [14]. While the British National Formulary (BNF) is a well-known comprehensive source of medication names, since it is not openly available as a dataset, it has not been included in the analysis. We have also omitted resources that are not relevant to UK prescribing, such as US resources, as these use different medication terms. We give details of these different resources in the following paragraphs.

Practice level prescribing data from primary care General Practitioners (GP) were obtained from an openly available data source made accessible by NHS Digital [12]. This data included a list of all medicines, dressings and appliances that were being prescribed and dispensed by the practices in England each month. The data covers all NHS prescriptions that were written in England and dispensed in the community within England. It also includes prescriptions that were written in England but dispensed outside England [12]. The data is not limited to prescriptions written by GPs, but also includes ones written by other non-medical prescribers attached to the practices, such as nurses and pharmacists. The medication names were linked to a BNF code. The list required some cleaning up before being used for comparison since the medication names were linked with different dose preparations, such as "Mebeverine HCl_Tab 135mg" which had to be stripped down to "Mebeverine HCl" or just "Mebeverine" for a more normalised format of comparison.

Prescription Cost Analysis (PCA) is another data source which provides information on medications used in primary care. Along with the names of medications, PCA also provides information on the number of items and Net Ingredient Cost (NIC) of all prescriptions that were dispensed in the community in England [11]. The PCA consists of medication names listed as BNF chemical names and linked to a BNF code. It also contains the individual preparation names linked with BNF codes as well. These were the two main categories that were used in the comparison with the other medication sources.

SLaM pharmacy data were obtained by requesting the list of prescribed medications from the pharmacy staff. The list of medication names provided by them were compiled based on the medications that are requested for purchase by clinicians at the hospital. The list provided by them required some pre-processing before being used for comparison, since the medication names were combined with the different dose and route preparations, such as "Buprenorphine 2 mg Sublingual Tablets", which had to be stripped down to just "Buprenorphine" for effective comparison.

DrugBank is a comprehensive and freely available resource of Food and Drug Administration (FDA) approved medication names and detailed information associated with the medication names [14]. It combines detailed chemical and pharmacological details of medications with comprehensive medication target information (such as structure and pathway), however only the medication names were utilised for the comparisons. While this is a US resource, it was included to gauge its similarities with other UK resources.

In this study, we considered mental health EHRs from the South London and Maudsley NHS Foundation Trust (SLaM), one of the largest mental health care providers in Western Europe. SLaM serves around 1.36 million residents of four south London boroughs (Lambeth, Southwark, Lewisham and Croydon) and provides comprehensive secondary, and a range of tertiary, mental health care services. All clinical records in SLaM services have been electronic since 2006 (including imported legacy data from earlier years for some services) and have been made available for research since 2008 following the establishment of the Clinical Record Interactive Search (CRIS) platform. CRIS has been described in detail [9]. CRIS currently accesses about 30 million case notes and correspondence, with an average of 90 documents per patient [13] and there are 11,551,563 medication mentions across 202,447 patient records. CRIS is routinely processed by an NER application to extract medication entities [10]. A handcrafted gazetteer is used for this purpose, which we will refer to as the CRIS medications gazetteer. The CRIS medication gazetteer has been created using information from the BNF, as well as a custom list of common

misspellings that occur within the clinical notes and has been re-iterated over the years making it a comprehensive list of medications.

The medication names from the other resources were compared to the CRIS medications gazetteer. Since the main objective is to use the combined gazetteer in a lookup function to identify medication name mentions within clinical notes, each medication source was also compared to a list of medications manually extracted from EHR documents. We will refer to these medication names as the manually extracted medication names. To construct the list of manually extracted medication names, medication names were manually extracted by annotation of CRIS documents by two medical students. The documents examined were from patients with diagnoses of schizophrenia, depression, dementia, and stress/anxiety [15]. To make the manual annotation task feasible, it was decided that documents for 50 patients would be extracted for each diagnosis group and for each text source. Each patient had 1-2 documents each. These manual annotations included generic categories of medications as well such as "antipsychotics" or "painkillers" which were excluded from the list before any comparisons were made [15]. We A summary of the number of documents that were annotated, and the number of annotations is provided in the table below (Table 1).

Table 1. Number of annotations of medication names

Cohort	No. of annotations
Schizophrenia	945
Depression	577
Dementia	464
Stress and anxiety	186

Since the different medication term resources had medication information written in quite varying formats, such as medication names along with the dose preparation, route preparation or medication combinations, and as we are only interested in matching the drug name at this stage, some pre-processing was required in order to normalise terms for a more accurate comparison. We therefore tokenised the medication names to facilitate normalisation. Normal forms were obtained by splitting the names in the original list into their first tokens only, splitting them wherever there was a '-', '/', '(', or a space. A summary of the different data sources used can be seen in the table below (Table 2).

Table 2. Sources of medication name data (no duplicates)

Source	Total number of medications (pre-tokenisation)	Total number of medications (post-tokenisation)
NHS Digital - GP prescribing data	1957	1528
Prescription Cost Analysis	1442	1202
SLaM pharmacy	679	559
DrugBank	6176	5466
Manually extracted medication names	315	265
CRIS medications gazetteer	4170	3718

A comparison of the different sources to the manually extracted medication names and the CRIS medications gazetteer was done before any clean-up (pre-tokenisation) and then against the normalised term after the tokenisation (post-tokenisation). A quick analysis was carried out on how many medication names compared with the manually extracted medication names after the tokenisation, and further ongoing work is looking into what kind of information was lost in this process, in order to assess whether crucial information was being missed because of the tokenisation.

Due to the repetitions between the sources for NHS GP prescriptions and PCA, and the need to have a combined medication list for both primary and secondary care, a comparison was also run between a list created by combining the NHS GP prescription dataset and the SLaM pharmacy list (GP + SLaM) to understand if this would be comprehensive enough.

Results

A comparison of the different sources to the manually extracted medication names and to the CRIS medications gazetteer, both pre- and post- tokenisation, was done to shed some light on which sources matched the most. When comparing the GP + SLaM list with the CRIS medications gazetteer, 56% of the medication names matched pre-

tokenisation which went up to 87% after the tokenisation, using lenient rules of the first token of the medication name to match with any part of the medication name in the comparison list. The remaining 13% that did not match after tokenisation appear to be names that did not make sense after being split, for example the first token being a word such as "activated" from "activated charcoal" or "conjugated" from "conjugated oestrogens". Another example would be words such as "co" from "co-tenidone", which were rendered meaningless after the split. Amongst all the medications in the GP list, 43% had been split in order to match the CRIS gazetteer. Amongst all the medications in the SLaM list, 48% of them were split in order to match the CRIS gazetteer. Out of the ones that did not match with the CRIS gazetteer, 180 of them were split, 77% of all non-matches. The reason for this might be the same as mentioned previously, where the medication names are incomplete or have lost meaning after the split. Amongst manually extracted medication names, 77% of the annotations matched with the GP + SLaM list. A summary of comparisons between the other sources is in Table 3 and Table 4 below. Annotations in the tables refer to the manually extracted medication names.

Table 3. Summary of comparison between different sources (pre-tokenisation)

	CRIS gaz.	**Annotations**	**PCA**	**DrugBank**	**SLaM**	**NHS_GP**	**GP + SLaM**
CRIS gaz.		5%	19%	21%	19%	24%	35%
Annotations**	64%		36%	46%	50%	32%	56%
PCA	56%	8%		48%	38%	24%	42%
DrugBank	14%	2%	11%		11%	5%	11%
SLaM	77%	15%	54%	63%		34%	100%
NHS_GP	52%	5%	18%	15%	18%		100%
GP + SLaM	56%	7%	23%	26%	39%	74%	

Table 4. Summary of comparison between different sources (post-tokenisation)

	CRIS gaz.	**Annotations**	**PCA**	**DrugBank**	**SLaM**	**NHS_GP**	**GP + SLaM**
CRIS gaz.		7%	30%	28%	16%	37%	44%
Annotations**	79%		62%	60%	33%	60%	72%
PCA	83%	14%		81%	37%	44%	63%
DrugBank	18%	3%	19%		9%	10%	14%
SLaM	83%	16%	76%	74%		42%	100%
NHS_GP	88%	13%	39%	38%	19%		100%
GP + SLaM	87%	13%	49%	47%	41%	84%	

As noticed by comparing Table 3 and Table 4, there is a substantial increase in the percentage of matches amongst the different sources when they are compared post-tokenisation, based on the first token of the medication name only. This was further investigated by looking at the individual medication names that were split in order to match and understand why they matched after the split. Amongst the medication names that matched on their first token, 35% of them were extensions to the names, such as "aclidinium bromide" in the GP prescriptions list not matching with the CRIS gazetteer, but "aclidinium" matching. Another common situation is the medication name containing a suffix of "sr"/ "xl"/ "mr"/ "cr", each of which have different meanings: "sr" refers to sustained release, "xl" is prolonged release, "mr" is modified release, and "cr" is to controlled release. Whilst eliminating part of the medication name might increase the number of matches, these cases require further investigation and consultation from a clinician to understand the impact of this elimination, and how using such a partial name in a gazetteer, might introduce ambiguity into the meaning of the term, and potentially change the meaning of the medication mention captured.

Discussion and Conclusion

This analysis and construction of the ideal medication resource is still ongoing, and the process undertaken so far highlights the complex nature of creating such a comprehensive gazetteer. Further analysis is being carried out on how the tokenisation of the medication names affects the quality of the data, and whether important information is being missed or misclassified by such tokenisation. For example, omitting HCl from the medication term Mebeverine HCl might change the meaning of the medication name. Another important consideration is the process of updating such a gazetteer, and if it can be automated. If it is intended to be used as part of an NLP application to assist in NER on an ongoing basis, timely updates to the gazetteer will be required in order to keep up to date changes in the underlying resources, such as the introductions of any new medications. It will not be very practical to go through extensive pre-processing every time the list needs updating, and bearing this in mind, the construction of the gazetteer should be as straightforward as possible, so that updating it is not a cumbersome process. The findings from the completion of construction of such a gazetteer will bring us one step closer to understanding the steps required to convert a resource not meant for NLP into one that is well-suited for NLP purposes. Once complete, this gazetteer will be made openly available for other researchers to use.

References

1. Zamin N, Oxley A. Building a corpus-derived gazetteer for named entity recognition. In: Communications in Computer and Information Science. 2011. p. 73–80.
2. Dozier C, Kondadadi R, Light M, Vachher A, Veeramachaneni S, Wudali R. Named entity recognition and resolution in legal text. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2010. p. 27–43.
3. Mikheev A, Moens M, Grover C, Moens M, ac Uk E. Named Entity Recognition without Gazetteers. 1999.
4. Ichi Kazama J', Torisawa K. Exploiting Wikipedia as External Knowledge for Named Entity Recognition [Internet]. Association for Computational Linguistics; 2007 [cited 2020 Jan 16]. Available from: http://hyperestraier.sourceforge.net/index.html
5. Riloff E, Jones R. Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping [Internet]. 1999 [cited 2020 Jan 20]. Available from: www.aaai.org
6. Thelen M, Riloff E. A Bootstrapping Method for Learning Semantic Lexicons using Extraction Pattern Contexts. 2002.
7. Etzioni O, Cafarella M, Downey D, Popescu AM, Shaked T, Soderland S, et al. Unsupervised named-entity extraction from the Web: An experimental study. Artif Intell. 2005 Jun;165(1):91–134.
8. Shinzato K, Sekine S, Yoshinaga N, Torisawa K. Constructing Dictionaries for Named Entity Recognition on Specific Domains from the Web.
9. Perera G, Broadbent M, Callard F, Chang C-K, Downs J, Dutta R, et al. Cohort profile of the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) Case Register: current status and recent enhancement of an Electronic Mental Health Record-derived data resource. BMJ Open [Internet]. 2016 Mar 1 [cited 2018 Jul 31];6(3):e008721. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26932138
10. Kadra G, Stewart R, Shetty H, Jackson RG, Greenwood MA, Roberts A, et al. Extracting antipsychotic polypharmacy data from electronic health records: developing and evaluating a novel process. BMC Psychiatry [Internet]. 2015 Jul 22 [cited 2018 Nov 10];15:166. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26198696
11. NHS Digital. Prescription Cost Analysis [Internet]. Prescription Cost Analysis: England 2018. 2018 [cited 2020 Jan 17]. Available from: https://digital.nhs.uk/data-and-information/publications/statistical/prescription-cost-analysis/2018
12. NHS Digital. Practice Level Prescribing [Internet]. Practice Level Prescribing Data. 2019 [cited 2020 Jan 16]. Available from: https://digital.nhs.uk/data-and-information/publications/statistical/practice-level-prescribing-data/august-2019
13. Velupillai S, Suominen H, Liakata M, Roberts A, Shah AD, Morley K, et al. Using clinical Natural Language Processing for health outcomes research: Overview and actionable suggestions for future advances. 2018 [cited 2019 Nov 20]; Available from: https://doi.org/10.1016/j.jbi.2018.10.005
14. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: A major update to the DrugBank database for 2018. Nucleic Acids Res. 2018 Jan 1;46(D1):D1074–82.
15. Chaturvedi J, Viani N, Sanyal J, Tytherleigh C, Hassan I, Baird K, et al. Development of a Corpus Annotated with Medications and their Attributes in Psychiatric Health Records. In: Language Resources and Evaluation Conference (Submitted). 2020.