Concept Identification and Normalisation for Adverse Drug Event Discovery in Medical Forums

Alejandro Metke-Jimenez and Sarvnaz Karimi

CSIRO, Australia
{alejandro.metke,sarvnaz.karimi}@csiro.au
http://aehrc.com

Abstract. Social media is becoming an increasingly important source of information to complement traditional pharmacovigilance methods. In order to identify signals of potential adverse drug reactions, it is necessary to first identify medical concepts and drugs in the text.
We evaluate different concept extraction techniques on medical forums and for the machine learning approaches we encode complex annotations using a scheme that showed good results in other domains.
Our study shows that the extended encoding scheme, although imperfect, still produces good results despite the complexities of social media. The comparison of techniques shows that the machine learning approach significantly outperforms the other approaches.

Keywords: Text Mining, Information Extraction, Ontology-based Text Normalisation, Drug Safety Adverse Drug Reaction Discovery

1 Introduction

Adverse Drug Reactions (ADRs) are a major concern for public health. An ADR is an injury caused by a medication that is administered at the recommended dosage, for recommended symptoms. The traditional pharmacovigilance methods have shown limitations that have prompted the search for alternative sources that might help identify *signals* of potential ADRs.

One of these sources is social media. However, it is first necessary to identify concepts of interest, such as mentions of adverse effects, in the text which is unstructured and noisy. This step is critical because errors can affect the subsequent stages of the signal detection process.

2 Background and related work

Although there is a large body of literature on generic information extraction from text such as news and social media, especially Twitter, there is limited work on the specific area of ADR detection. A comprehensive survey of text and data mining techniques used for ADR signal detection can be found in [1].

In this paper we are concerned with concept extraction which can be divided in two steps: identifying spans of text that represent a concept of interest, referred to as concept identification, and mapping the spans to the corresponding concepts in a chosen ontology, referred to as concept normalisation.

The problem of medical concept extraction has been extensively studied by the clinical text mining community. Most techniques used to extract ADRs from social media use dictionary-based approaches. A review of these approaches and the most commonly used lexicons can be found in [2].

More recently, machine learning techniques have been applied to extract ADRs from social media. In [3] the authors implemented a CRF classifier to detect mentions of ADRs in a corpus of Twitter and DailyStrength posts and reported improvements over dictionary-based approaches.

3 Problem formulation

Our goal is to evaluate the concept extraction task specifically on medical forums. Apart from the challenges that this type of data raises, such as dealing with misspellings and colloquial language, we also aim to evaluate techniques that are widely used to determine how well they perform against each other.

3.1 Concept identification

Concept identification consists of identifying spans of text that represent medical concepts. This task can be framed as a binary classification problem and evaluated using *precision*, *recall*, and *F-score*. In the strict version of the evaluation, the spans are required to match exactly. In the relaxed version the spans only need to overlap to be considered a positive match.

In order to consider the correct classification of negative examples we also evaluate the systems using *accuracy*. The set of negative examples is defined as all the spans that are created by all the systems under evaluation that are not part of the gold standard.

3.2 Concept normalisation

The normalisation step takes the spans that were identified in the identification step and maps them to a concept in an ontology. ADR spans are mapped to the *Clinical Finding* hierarchy of SNOMED CT and drug spans to concepts in the Australian Medicines Terminology (AMT).

Concept normalisation is often evaluated using a metric referred to as accuracy. To avoid confusion with the metric used in the first part of the task, we refer to this metric as effectiveness, which is defined as

$$\text{Effectiveness}_{strict} = \frac{n_{TP} \cap n_{correct}}{t_g} \quad \text{Effectiveness}_{relaxed} = \frac{n_{TP} \cap n_{correct}}{n_{TP}},$$

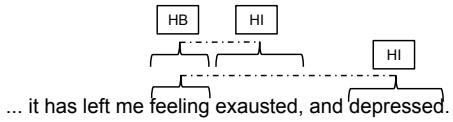

Fig. 1. A discontinuous, overlapping annotation using the extended BIO format.

where n_{TP} is the number of spans that match the gold standard exactly, $n_{correct}$ is the number of spans that were mapped to the correct concept in the corresponding ontology, and t_g is the total number of identified concepts or spans in the gold standard. The relaxed version only considers the spans that were correctly identified in the previous stage.

4 Dataset

In our experiments, we used an annotated corpus called CSIRO Adverse Drug Event Corpus (CADEC)[1]. This corpus is a collection of medical posts sourced from the medical forum AskaPatient. A detailed description of the corpus can be found in [4]. To develop and evaluate a machine learning approach, we divided the data into training and testing sets, using a 70/30 split.

5 Methods

Most existing approaches to ADR mining in social media use dictionary-based techniques based on pattern matching rules or sliding windows. We implemented a sliding window approach using the Lucene search engine, without using stemming or removing stop words.

We also implemented a CRF classifier, similar to the one used in [3] but with fewer features, using the Stanford NER suite [5]. A CRF classifier takes as input different features that are derived from the text, such as the words that surround each token, letter n-grams and word shape features.

One of the challenges of dealing with discontinuous spans is representing them in a format that is suitable as input to the classifier. Continuous spans are typically represented using the standard Begin, Inside, Outside (BIO) chunking representation. This format does not support the notion of discontinuous spans and several solutions have been proposed to overcome this limitation. The most successful approach in tasks such as CLEF has been to extend the BIO format with additional tags to represent the discontinuous spans.

With the extended BIO format, the following additional tags are introduced: D{B, I} and H{B, I}. The first set of tags is used to represent discontinuous, non-overlapping spans. The second set of tags is used to represent discontinuous, overlapping spans that share one or more tokens (the H stands for *Head*, as in *head word*). Figure 1 shows an example of a complex span.

[1] `http://dx.doi.org/10.4225/08/5490FA2E01A90`

Table 1. Results of roundtrip transformation using extended BIO format.

Set	TP	FP	FN	Total	Precision	Recall	F-Score
Training	6325	122	66	6513	0.98107	0.98967	0.98535
Test	2618	50	26	2694	0.98125	0.99016	0.98569
Total	8943	172	92	9207	0.98113	0.98981	0.98545

One limitation of this approach is that it is impossible to represent several discontinuous spans in the same sentence unambiguously. To determine how this might affect the performance of the CRF approach with the CADEC dataset, a round trip transformation was done on the gold standard annotations and the results are shown in Table 1. This is equivalent to having a perfect classifier.

The CRF classifier only identifies relevant spans but does not map them to concepts. Two approaches were explored to achieve this mapping. The first one is based on the Vector Space Model (VSM) and was implemented using Lucene. The target ontology was indexed using stemming and removing stop words by creating a document for each term and storing the corresponding concept id. Then, the text of each span was used to query the index, without requiring all the tokens to match. The top ranked concept was assigned to the span and if the query returned no results then the span was annotated as *concept_less*.

The second approach uses Ontoserver, a terminology server developed at the Australian e-Health Research Centre, that given a free-text query returns the most relevant SNOMED CT and AMT concepts. Ontoserver uses a purpose-tuned retrieval function based on a multi-prefix matching algorithm [6].

To determine if the improvements obtained with any two different methods were statistically significant, we used McNemar's test.

6 Results and discussion

The results of the concept identification task are shown in Table 2. The CRF implementation outperforms MetaMap and all the dictionary-based implementations in all of the metrics that were considered, in both strict and relaxed modes, as expected.

Identifying drugs usually involves less ambiguity than identifying ADRs and therefore better results were expected in this task. The results show that the CRF indeed performs better in this task that in the ADR identification task. Note also that most of the dictionary-based implementations achieve good recall but low precision; this is likely due to some of the constraints in the annotation guidelines, for example, drug classes are excluded. The CRF is capable of learning these constraints while the dictionary-based approaches are not.

Table 3 shows the results of the concept normalisation task. In this case the strict metric is more relevant, because some implementations can achieve a very high score in the relaxed version despite having a very poor overall performance. The results show that Ontoserver outperforms the other approaches when normalising ADRs. Overall, however, the results are quite poor. This highlights two

Table 2. Evaluation results of the concept identification task, sorted by accuracy. Statistical significant difference with the next method is indicated with $*$ ($p < 0.01$).

ADRs						Drugs					
Type	**Method**	**Pr**	**Rc**	**Fs**	**Ac**	**Type**	**Method**	**Pr**	**Rc**	**Fs**	**Ac**
	UMLS	0.264	0.392	0.316	0.454		UMLS	0.160	0.882	0.271	0.546
	MetaMap	0.105	0.080	0.091	0.485*		AMT	0.160	0.775	0.266	0.589*
Strict	CHV	0.457	0.370	0.409	0.656*	Strict	MetaMap	0.022	0.021	0.021	0.816*
	SCT	0.498	0.352	0.412	0.678*		CHV	0.468	0.856	0.605	0.893*
	CRF	0.644	0.565	0.602	0.760*		CRF	0.943	0.840	0.889	0.980*
	UMLS	0.454	0.674	0.543	0.635		UMLS	0.168	0.923	0.284	0.554
	CHV	0.747	0.605	0.669	0.807*		AMT	0.173	0.837	0.287	0.601*
Relaxed	MetaMap	0.794	0.605	0.687	0.822*	Relaxed	MetaMap	0.145	0.139	0.142	0.839*
	SCT	0.818	0.578	0.677	0.822		CHV	0.489	0.893	0.632	0.900*
	CRF	0.908	0.797	0.849	0.909*		CRF	0.979	0.872	0.923	0.986*

Table 3. Results of the evaluation of the concept normalisation task.

ADRs				Drugs			
Strict		**Relaxed**		**Strict**		**Relaxed**	
Method	**Ef**	**Method**	**Ef**	**Method**	**Ef**	**Method**	**Ef**
MetaMap	0.029	MetaMap	0.363	MetaMap	0.000	MetaMap	0.000
UMLS	0.105	UMLS	0.266	UMLS	0.000	UMLS	0.000
CHV	0.106	CHV	0.287	CHV	0.000	CHV	0.000
CRF+VSM	0.327	CRF+VSM	0.578	CRF+VSM	0.749	CRF+VSM	0.891
SCT	0.332	SCT	0.943	CRF+Onto	0.773	CRF+Onto	0.920
CRF+Onto	0.376	CRF+Onto	0.666	AMT	0.758	AMT	0.978

important aspects of the task. First, it is inherently difficult to map colloquial language to ontologies that contain more formal terms. Second, because in this task the goal is to map the spans to SNOMED CT concepts, the quality of the results when using approaches that rely on other controlled vocabularies will depend on the quality of the mappings between those vocabularies and SNOMED CT.

It was also expected that the different methods would perform better when normalising drugs than when normalising ADRs. For most implementations this turned out to be true, except for the dictionary-based methods that are not based on AMT. These methods were unable to normalise any concepts because maps between the other controlled vocabularies and AMT do not currently exist.

7 Conclusions and future work

Pharmacovigilance should no longer rely only on manual reports of potential drug adverse effects. One viable alternative is actively detecting signals of adverse drug reactions in social media through text mining.

We conducted an empirical evaluation of different methods to automatically extract concepts from medical forums. We explored the implications of representing complex annotations in a format suitable for use with machine learning

methods. Finally, we proposed and implemented two concept normalisation techniques that we used in conjunction with our machine learning implementation.

We showed that there is some ambiguity when using the extended BIO format to represent the complex annotations, but the impact on the overall performance is not substantial. The experimental results showed that the CRF implementation combined with Ontoserver outperformed all the other methods that were evaluated. Even though these results show that machine learning methods perform better than simple dictionary-based methods, they also highlight the complexities in mapping the spans of text to concepts in an underlying ontology or controlled vocabulary.

Regarding future work, existing concept normalisation implementations in social media do not make use of the context of the spans. We believe more advanced methods may benefit from having access not only to the text in the span but also to the surrounding tokens and previously identified concepts.

Acknowledgements

AskaPatient kindly provided the data used in this study for research purposes only. Ethics approval for this project was obtained from the CSIRO ethics committee, which classified the work as low risk (CSIRO Ecosciences #07613).

References

1. Sarvnaz Karimi, Chen Wang, Alejandro Metke-Jimenez, Raj Gaire, and Cecile Paris. Text and data mining techniques in adverse drug reaction detection. *ACM Computing Surveys*, 47(4):56, 2015.
2. Abeed Sarker, Rachel Ginn, Azadeh Nikfarjam, Karen OConnor, Karen Smith, Swetha Jayaraman, Tejaswi Upadhaya, and Graciela Gonzalez. Utilizing social media data for pharmacovigilance: A review. *Journal of Biomedical Informatics*, 54:202 – 212, 2015.
3. Azadeh Nikfarjam, Abeed Sarker, Karen O'Connor, Rachel Ginn, and Graciela Gonzalez. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, 2015.
4. Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. CADEC: A corpus of adverse drug event annotations. *Journal of Biomedical Informatics*, 55:73–81, 2015.
5. Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by Gibbs sampling. In *The 43rd Annual Meeting On Association for Computational Linguistics*, pages 363–370, Ann Arbor, Michigan, 2005.
6. Merlijn Sevenster, Rob van Ommering, and Yuechen Qian. Algorithmic and user study of an autocompletion algorithm on a large medical vocabulary. *Journal of Biomedical Informatics*, 45(1):107–119, 2012.