Concept Extraction to Identify Adverse Drug Reactions in Medical Forums: A Comparison of Algorithms

Alejandro Metke-Jimenez Sarvnaz Karimi
CSIRO, Australia
Email: {alejandro.metke,sarvnaz.karimi}@csiro.au

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 in the social media text. Most of the existing studies use dictionary-based methods which are not evaluated independently from the overall signal detection task.

We compare different approaches to automatically identify and normalise medical concepts in consumer reviews in medical forums. Specifically, we implement several dictionary-based methods popular in the relevant literature, as well as a method we suggest based on a state-of-the-art machine learning method for entity recognition. MetaMap, a popular biomedical concept extraction tool, is used as a baseline. Our evaluations were performed in a controlled setting on a common corpus which is a collection of medical forum posts annotated with concepts and linked to controlled vocabularies such as MedDRA and SNOMED CT.

To our knowledge, our study is the first to systematically examine the effect of popular concept extraction methods in the area of signal detection for adverse reactions. We show that the choice of algorithm or controlled vocabulary has a significant impact on concept extraction, which will impact the overall signal detection process. We also show that our proposed machine learning approach significantly outperforms all the other methods in identification of both adverse reactions and drugs, even when trained with a relatively small set of annotated text.

1 Introduction

Adverse Drug Reactions (ADRs), also known as drug side effects, are a major concern for public health, costing health care systems worldwide millions of dollars [Hug et al., 2012, Ehsani et al., 2006, Roughead and Semple, 2009]. 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 of information that might help identify *signals* of potential ADRs. These signals can then be used to select which cases warrant a more thorough review. These assessments are performed by regulatory agencies such as the Food and Drug Administration (FDA) in the United States and the Therapeutic Goods Administration (TGA) in Australia, and intend to establish a causal effect between the drug and the ADR. If a causal link is found, and depending on the severity of the ADR, it will be added to the drug's label or it might even trigger a removal of the drug from the market if it is considered life-threatening.

Social media has been identified as a potential source of information that could be used to find signals of potential ADRs [Benton et al., 2011]. A public opinion survey conducted by The Pew Research Center's Global Attitudes Project in 2009 [Fox and Jones, 2009] showed that 61% of American adults looked for health information online, 41% had read about someone else's experience, and 30% were actively creating new content. These numbers give a strong indication about the growing importance of social media in the area of health.

Several attempts at extracting ADR signals from social media have shown promising results [Benton et al., 2011, Leaman et al., 2010, Yang et al., 2012, Liu and Chen, 2013]. However, all of these techniques first need to identify concepts of interest, such as mentions of adverse effects, in the social media text which is unstructured and noisy. Most current approaches use medical concept identification techniques based on dictionary lookup, but do not evaluate this step independently [Metke-Jimenez et al., 2014]. This step is critical because errors can affect the subsequent stages of the signal detection process. The problem of concept identification and normalisation — linking the identified concepts to their corresponding concepts in controlled vocabularies — has been studied extensively in the context of Natural Language Processing (NLP) of clinical notes, but these techniques have not been used in the context of ADR mining from social media. The main reasons for this are the lack of publicly available corpora with gold-standard annotations and the difficulty in identifying specific concepts, such as adverse reactions, in lay people language. The noisy nature of social media text makes concept identification a hard problem. For example, in the corpus used in this work the drug Lipitor is spelled in seven different ways (Lipitor, Liptor, Lipitol, Lipiltor, Liptior, Lipior and Litpitor) and it is also written using different case combinations (e.g. Lipitor, LIPITOR and lipitor).

The contributions of this paper are twofold:

1. Several existing concept identification and normal-

isation methods are evaluated in the domain of adverse effect discovery from social media, including the dictionary-based methods applied in the recent ADR mining literature, as well as a state-of-the-art machine learning method that has been used successfully in similar tasks in other domains; and,

2. A variety of evaluation metrics are used, and in some cases proposed, to compare the effectiveness of different methods, including the statistical analysis of performance improvement.

2 Background

This section starts by clarifying the terminology that is used throughout the paper. Then, a brief introduction to the controlled vocabularies that we refer to in the literature and our experiments is given.

2.1 Terminology Clarification

Some of the terminology used in this paper has been used inconsistently in the literature. We use the terms concept and entity in free text interchangeably. Concept recognition and concept identification are also treated as synonyms. We refer to concept extraction as a process of concept identification followed by normalisation / mapping to controlled vocabularies.

We also note that the methods we refer to as dictionary-based are also known as lexicon-based or lexicon lookup.

2.2 Controlled vocabularies

Controlled vocabularies are typically used to identify medical concepts in free text. This section provides background on controlled vocabularies that are commonly used in the relevant literature. Note that some of these resources are really taxonomies or ontologies, but are used as controlled vocabularies in the context of this paper.

MedDRA The Medical Dictionary for Regulatory Activities[1] is a thesaurus of ADRs used internationally by regulatory agencies and pharmaceutical companies to consistently code ADR reports.

Before MedDRA, the FDA had developed the Coding Symbols for a Thesaurus of Adverse Reaction Terms (COSTART), which is now obsolete.

CHV The Consumer Health Vocabulary[2] provides a list of health terms used by lay people, including frequent misspellings. For example, it links both *lung tumor* and *lung tumour* to *lung neoplasm*.

SNOMED CT The Systematized Nomenclature of Medicine - Clinical Terms[3] is a large ontology of medical concepts that has been recommended as the reference terminology for clinical information systems in countries such as Australia, the United Kingdom, Canada, and the United States [Lee et al., 2013]. It includes formal definitions, codes, terms, and synonyms for more than 300,000 medical concepts. Several versions of the ontology exist, including an international version and several country-specific versions that extend the international version to add local content and synonyms.

UMLS The Unified Medical Language System[4] is a collection of several health and biomedical controlled vocabularies, including MedDRA, SNOMED CT, and CHV. Terms in the controlled vocabularies are mapped to UMLS concepts. It also provides a semantic network that contains semantic types linked to each other through semantic relationships. Each UMLS concept is assigned one or more semantic types.

AMT The Australian Medicines Terminology[5] is an extension of the Australian version of SNOMED CT that provides unique codes and accurate, standardised names that unambiguously identify all commonly used medicines in Australia.

3 Related Work

Although there is a large body of literature on generic information extraction from formal text such as news and social media, especially Twitter, there is limited work on the specific area of ADR detection. ADR signal detection has been studied in spontaneous reporting systems [Bate and Evans, 2009], medical case reports [Gurulingappa et al., 2012], and Electronic Health Records [Friedman, 2009]. A comprehensive survey of text and data mining techniques used for ADR signal detection from several sources, including social media, can be found in [Karimi et al., 2015b]. Below, we review the most relevant ADR extraction techniques used in social media. We also review the state of the art in medical concept identification and normalisation in the context of clinical notes.

3.1 ADR Extraction from Social Media

Medical forums are online sites where people discuss their health concerns and share their experience with other patients or health professionals. Actively mining these forums could potentially reveal safety concerns regarding medications before regulators discover them through more passive methods via official channels such as health professionals.

Leaman et al. [2010] proposed to mine patients' comments on health related web sites, specifically DailyStrength[6], to find mentions of adverse drug events. They used a lexicon that combines COSTART and a few other sources to extract ADR-related information from text. In a preprocessing step, they break the

[1] http://www.meddra.org/
[2] http://www.consumerhealthvocab.org/
[3] http://www.ihtsdo.org/snomed-ct/
[4] http://www.nlm.nih.gov/research/umls
[5] http://www.nehta.gov.au/our-work/clinical-terminology/australian-medicines-terminology
[6] http://www.dailystrength.org/

posts into sentences, tokenise the sentences, run a Part-of-Speech (POS) tagger, remove stopwords, and stem the words using the Porter stemmer. Using a sliding window approach, they match the lexicon entries with the preprocessed text and then evaluate the matches against the manually annotated text. Their data was annotated with ADRs, beneficial effects, indications, and others. We evaluate a similar method without taking into account the similarity between the tokens.

Chee et al. [2011] applied Naïve Bayes and Support Vector Machine classifiers to identify drugs that could potentially become part of the watchlist of the US regulatory agency, the FDA. They used patients posts on Health and Wellness Yahoo! Groups. The text was processed to generate features for the classifiers. They had two sets of features: all the words from the posts, and only those words that matched their controlled vocabulary (that included MedDRA and a list of diseases). Misspellings were not fixed.

Benton et al. [2011] extracted potential ADRs from a number of different breast cancer forums (such as `breastcancer.org`) by using frequency counts of terms from a controlled vocabulary in their corpus and then using association rule mining to establish the relationship between the matching terms. Association rule mining is a data mining approach popular for mining ADRs from regulatory and administrative databases. The method by Benton et al. was an advancement on the approach used by Leaman et al. (2010), as they did not stop at just the extraction of interesting concepts, but also proposed a method to establish a relationship between the extracted terms.

Yang et al. [2012] studied signal detection from a medical forum called MedHelp using data mining approaches. They extended the existing association rule mining algorithms by adding "interestingness" and "impressiveness" metrics. They had to find mentions of ADRs in the text to process the forum data and calculate confidence and leverage. To do this, they used a sliding window and the CHV as a controlled vocabulary to match the terms.

None of these studies [Chee et al., 2011, Benton et al., 2011, Yang et al., 2012] evaluated the information extraction step on its own, which we will cover in this study.

Liu and Chen [2013] implemented a system called AZDrugMiner. Data was collected using a crawler and was then post-processed by removing any HTML tags and extracting text for further analysis. They then used an NLP tool called OpenNLP to break the text into sentences. To find the relevant parts of each sentence, for example mentions of a drug, they used MetaMap [Aronson, 2001], which maps text to UMLS concepts. After this stage, they extracted relations using co-occurrence analysis. They also used a tool called NegEx [Chapman et al., 2001] to identify negations in the text. This work uses MetaMap for the concept extraction step which we use as a baseline in our work.

Sampathkumari et al. [2014] proposed to used a machine learning approach, Hidden Markov Model, to extract relationships between drugs and their side effects in a medical forum called medications.com. For their concept recognition module they relied on a dictionary-lookup method.They created a dictionary of drug names from the drugs.com website and a dictionary of adverse drug effects from SIDER, a resource that lists side effect terminology. The concept recognition step was not evaluated on its own.

Metke-Jimenez et al. [2014] empirically evaluated a lexicon-based concept identification mechanism, similar to the ones reported in the existing literature, and tested different combinations of preprocessing techniques and controlled vocabularies using a manually annotated data set of medical forum posts from the AskaPatient website. The results showed that the best performing controlled vocabulary was the CHV, but the overall performance was quite poor. Our work has a similar goal but differs because we compare more methods, including a baseline method and a state of the art machine learning method. Also, the data set we use is larger and contains a wider variety of posts. The task we evaluate also includes *concept normalisation* which requires mapping the spans that were identified to a corresponding concept in a controlled vocabulary. Finally, we use more comprehensive metrics to compare the relative performance of the different techniques under evaluation.

3.2 Medical Concept Identification and Normalisation

The problem of medical concept identification and normalisation has been extensively studied by the clinical text mining community. Early work often relied on pattern matching rules — e.g., [Evans et al., 1996] — or used MetaMap as a tool to identify concepts using the UMLS Metathesaurus — e.g., [Jimeno et al., 2008].

More recently, several open challenges have bolstered the research in this area, including the i2b2 Medication Extraction Challenge [Uzuner et al., 2010], ShARe/CLEF eHealth Evaluation Lab 2013 and SemEval-2014.

In 2010, the i2b2 medication extraction challenge was introduced as an annotation exercise. Participating teams were given a small number of discharge summaries (10 per person) to annotate for mentions of medications, the way these medications were administered (dosage, duration, frequency, and route), as well as reasons for taking the medications. To complete this challenge some participating teams used automated methods as well as manual reviews. Mork et al. [2010], for example, used a combination of dictionary lookup (e.g., UMLS, RxTerms, DailyMed) and concept annotation tools (e.g., MetaMap) to find the concepts.

Task 1 of the ShARe/CLEF eHealth Evaluation Lab 2013 [Pradhan et al., 2013] used the ShARe corpus, which provides a collection of annotated, de-identified clinical reports from US intensive care units (version 2.5 of the MIMIC II database[7]).

The task was divided into two parts. The goal of part A was to identify spans that represent *disorders*, defined as any text that can be mapped to a SNOMED

[7] `http://mimic.physionet.org`

Table 2: Best reported scores in part B (SNOMED CT mapping) of ShARe/CLEF 2013 Task 1 and SemEval 2014 Task 7.

Task	Strictness	Accuracy
CLEF [Leaman et al., 2013]	Strict	0.589
SemEval [Zhang et al., 2014]		0.741
CLEF [Leaman et al., 2013]	Relaxed	0.895
SemEval [Zhang et al., 2014]		0.873

CT concept in the Disorder semantic group. The goal of Part B was to map these spans to SNOMED CT codes. Part A was evaluated using precision, recall, and F-Score (see Section 4.1 for the definition of these metrics). Evaluation for concept identification was divided into two categories: *strict* and *relaxed*. The strict version required that the annotations match exactly, while the relaxed version did not. Part B was evaluated using accuracy, which was defined as the number of pre-annotated spans with correctly generated codes divided by the total number of pre-annotated spans (note that in this paper this metric is referred to as *effectiveness*; see Section 4.2). The strict version considered the total number of pre-annotated spans to be the total number of entities in the gold standard. The relaxed version considered the total to be the number of strictly correct spans generated by the system in part A.

All the best performing systems for part A (concept identification) used machine learning algorithms, including Conditional Random Fields (CRF) and Structural Support Vector Machines (SSVM) [Tang et al., 2013, Leaman et al., 2013, Gung, 2013]. Our work extends on the definition of this task by increasing the number of concepts to be identified, and tailoring it to the adverse effect signal detection area. We also target forum data which raises its own specific challenges due to language irregularities.

Task 7 at SemEval-2014 was a continuation of the CLEF 2013 task, but used more data for training and introduced a new test set [Pradhan et al., 2014]. The best scores obtained in these challenges are shown in Tables 1 and 2.

The main differences between the different implementations submitted to these challenges was the selection of features used as input to the machine learning algorithms. Table 3 shows some of the common features used by different systems. Note that not all systems reported their features.

Apart from these open challenges, a recent study by Ramesh et al. [2014] also proposed using supervised machine learning, including Naïve Bayes, support vector machines, and conditional random fields, to annotate ADR reports collected by the FDA for drugs and adverse effects and then reviewing the annotation using human annotators. The main goal of this study however was developing an annotated corpus of drug reviews and therefore different to our study in terms of the evaluations involved.

There is another line of studies that are also referred as *normalisation*, or more specifically social text normalisation, in the natural language processing domain. Studies such as [Hassan and Menezes, 2013, Ling et al., 2013, Chrupala, 2014] propose algorithms to restore the standard or formal form of non-standard words that appear frequently in social media text. For example, *chk* and *abt*, two abbreviations common in Twitter, may be normalised to *check* and *about*. In our work, we refer to normalisation as mapping specific medical concepts to biomedical ontologies and controlled vocabularies, which is different to transforming a given free text abbreviation to its formal equivalent.

Table 3: Some of the common features used in machine learning approaches to disorder span identification.

Feature	Description
Bag of words	The words that surround each token.
POS tags	The part of speech tag assigned to the token.
Word shape	Indicates the shape of the token, for example, if the token is composed of only lower case letters, upper case letters, or a combination.
Type of notes	The corpus includes different types of clinical notes (e.g. discharge summaries, radiology reports, etc.). This feature indicates the type of note that contains the token.
Section information	Indicates the section of the note that contains the token (e.g. Past Medical History).
Semantic mapping	The concept assigned to a token by an existing tool, such as MetaMap or cTAKEs.

4 Problem Formulation

Our goal is to evaluate the concept identification and normalisation step independently from the overall task of signal detection in free-text. We restrict our task to focus on social media, specifically medical forums. Apart from the challenges that this data type raises, such as dealing with misspellings and colloquial language, we also aim to evaluate of concept identification techniques that are widely used in the literature to determine how well they perform in comparison to each other. Since in the specific application of ADR signal detection, linking the concepts to a standard vocabulary provides another level of knowledge that can be utilised, we also evaluate this step which we call normalisation. In this section, we formally describe these two parts of the task: *concept identification* and *concept normalisation*, and their evaluation metrics.

4.1 Concept Identification

Concept identification consists of identifying spans of text that represent medical concepts, specifically drugs and ADRs. The latter is more challenging because the same medical concept can be considered an ADR, a symptom, or a disease, depending on the context in

Table 1: Best reported scores in part A (concept identification) of ShARe/CLEF 2013 Task 1 and SemEval 2014 Task 7.

Task	Strictness	Precision	Recall	F-Score
CLEF [Tang et al., 2013]	Strict	0.800	0.706	0.750
SemEval [Zhang et al., 2014]		0.843	0.786	0.813
CLEF [Tang et al., 2013]	Relaxed	0.925	0.827	0.873
SemEval [Zhang et al., 2014]		0.916	0.907	0.911

which it is used. We avoid dealing with this complexity and therefore define the goal of the task to be the identification of any span of text that could represent a drug or an ADR, disregarding the context.

Spans can be continuous or discontinuous. Spans cannot overlap each other, except when several discontinuous spans share a common fragment. Figure 1 shows some examples of these different span types. In the presence of potentially overlapping spans, the annotators were asked to select the longest one.

Concept identification can be framed as a binary classification problem and evaluated using precision, recall, and F-score as defined below

$$\text{Precision} = \frac{n_{TP}}{n_{TP}+n_{FP}},$$

$$\text{Recall} = \frac{n_{TP}}{n_{TP}+n_{FN}},$$

$$\text{F-Score} = 2 \times \frac{precision \times recall}{precision + recall},$$

where n_{TP} is the number of matching spans, n_{FP} is the number of spans reported by the system that are not part of the gold standard, and n_{FN} is the number of spans in the gold standard that were not reported by the system. 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 this case, however, only one to one mappings are allowed, i.e. a span can only be mapped to one other span.

These metrics do not consider the correct classification of negative examples [Sokolova and Lapalme, 2009]. In order to measure the overall effectiveness of each system, we propose to use accuracy, which is defined as

$$\text{Accuracy} = \frac{n_{TP} + n_{TN}}{n_{TP} + n_{FN} + n_{FP} + n_{TN}},$$

where n_{TN} is the number of spans that are not in the gold standard that were not generated by the implementation under evaluation. Notice that in this task, any span that is not part of the gold standard is considered an incorrect span; negative examples are not explicitly enumerated. Given that the total number of negative examples is extremely large and that we are interested in comparing several methods, the set of negative examples is defined as all the spans that are created by all the methods under evaluation that are not part of the gold standard.

4.2 Concept normalisation

The normalisation step takes the spans that were identified in the identification step and maps them to a concept in an ontology or controlled vocabulary. For example, all three mentions of medications *Pethidine*, *Demerol*, and *Meperidine* are all mapped to one, *Pethidine*. This step helps to find the links to concepts that are semantically similar or identical.

In our setting, ADR spans are mapped to the *Clinical Finding* hierarchy of SNOMED CT and in the case of drugs to a representative concept in AMT.

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

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

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

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. Notice that the relaxed effectiveness metric only considers the spans that were correctly identified in the concept identification stage, therefore a system that performs very poorly overall can still get a very high score on this metric.

5 Dataset

In our experiments, we used a publicly available annotated corpus called CSIRO Adverse Drug Event Corpus (CADEC)[8]. This corpus is a collection of medical posts sourced from the medical forum AskaPatient[9]. The forum is organised by drug names and allows consumers to post reviews on the medications that they are consuming in natural language. Figure 2 shows a sample from the AskaPatient website on Voltaren. For each post shown in one row, CADEC only contains two free-text columns: side effects and comments.

CADEC includes reviews on 12 drugs, a total of 1250 forum posts. These reviews were manually annotated with a set of tags such as drug name, and disease

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

[9] http://www.askapatient.com

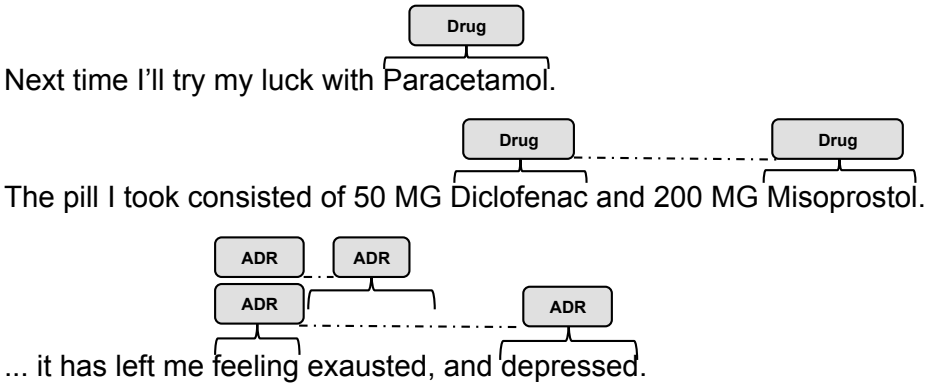


Figure 1: Span type examples from our dataset. From top to bottom: a sentence with a continuous annotation; a sentence with a discontinuous annotation; and a sentence with multiple discontinuous annotations that share a common fragment.

Table 4: The concepts annotated in the CADEC corpus.

Tag	Description
Drug	A mention of a medicine or drug. Medicinal products and trade names are included, but not drug classes (such as NSAIDs).
ADR	Mentions of adverse drug reactions clearly associated with the drug referenced by the post.
Disease	A mention of a disease that is the reason for the patient taking the drug.
Symptom	A mention of a symptom that is the reason for the patient taking the drug.
Finding	Any other mention of a clinical finding that does not fit into the previous categories, for example, the mention of a disease that is not the reason for the patient taking the drug.

Table 5: Number of documents and span types in the training and test sets.

	Training	Test	Total
Documents	875	375	1250
Continuous spans	5702	2350	8052
Discontinuous, non-overlapping spans	57	37	94
Discontinuous, overlapping spans	688	281	969
Total spans	6447	2668	9115

name as shown in Table 4. An expert clinical terminologist then mapped these spans to concepts in MedDRA, SNOMED CT and AMT. When no corresponding concept was available in the ontologies to represent the span, the value *concept_less* was assigned. A detailed description of the corpus, including the annotation guidelines, can be found in [Karimi et al., 2015a].

To develop and evaluate a machine learning approach, we divided the data into training and testing sets, using a 70/30 split. Unlike some previous work such as [Sampathkumari et al., 2014], we do not use k-fold cross-validation to avoid potential bias that may be introduced due to the nature of social media text [Karimi et al., 2015c]. Table 5 shows the number of documents and span types in each set.

6 Methods

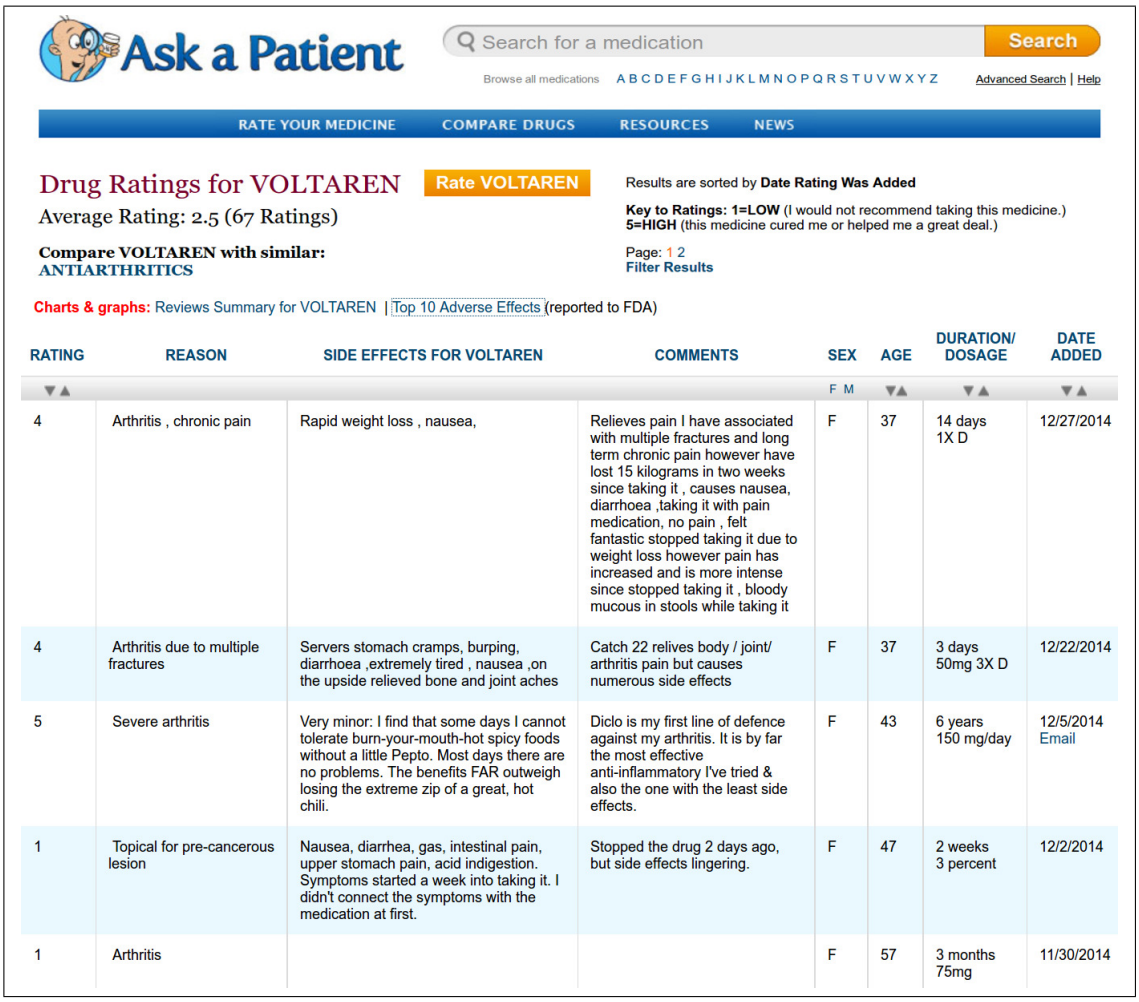
There are existing tools that are capable of extracting medical concepts from free text. One of these tools, MetaMap, is used as a baseline for the evaluation. The performance is expected to be poor mainly because MetaMap was not designed to work with social media text, which presents several challenges such as irregularities, including misspellings, colloquial phrases, and even novel phrases.

6.1 Dictionary-based Approaches

As discussed in Section 3.1, most existing approaches to ADR mining in social media use dictionary-based techniques based on pattern matching rules or sliding windows to identify drugs and adverse effects in noisy text. However, these techniques have never been evaluated independently of the overall task, nor been systematically compared to each other under one setting. This is in part because no standard testing set was publicly available previously.

We implemented a method similar to the sliding window approach used by Yang et al. [2012], but using the Lucene search engine. The medical forum posts were indexed and every post became a *document*. Then, a controlled vocabulary was chosen and for each entry a phrase search was executed. No stemming or stop word removal were used and tokenisation was done using Lucene's standard tokeniser, a grammar-based tokeniser that implements the Word Break rules from the Unicode Text Segmentation algorithm. Any matches were transformed into spans. Figure 3 illustrates how this approach works when CHV is used as the controlled vocabulary. The process is the same when replacing other vocabularies.

Notice that the concept normalisation step is implicitly being done when the spans are created. In the event that two different concepts match the same identical span, the system always selects the concept with the lexicographically greater concept id.



Ask a Patient

Search for a medication | Search

Browse all medications A B C D E F G H I J K L M N O P Q R S T U V W X Y Z | Advanced Search | Help

RATE YOUR MEDICINE | COMPARE DRUGS | RESOURCES | NEWS

Drug Ratings for VOLTAREN | Rate VOLTAREN

Average Rating: 2.5 (67 Ratings)

Compare VOLTAREN with similar: ANTIARTHRITICS

Results are sorted by **Date Rating Was Added**

Key to Ratings: 1=LOW (I would not recommend taking this medicine.) **5=HIGH** (this medicine cured me or helped me a great deal.)

Page: 1 2

Filter Results

Charts & graphs: Reviews Summary for VOLTAREN | Top 10 Adverse Effects (reported to FDA)

RATING	REASON	SIDE EFFECTS FOR VOLTAREN	COMMENTS	SEX	AGE	DURATION/ DOSAGE	DATE ADDED
				F M			
4	Arthritis , chronic pain	Rapid weight loss , nausea,	Relieves pain I have associated with multiple fractures and long term chronic pain however have lost 15 kilograms in two weeks since taking it , causes nausea, diarrhoea ,taking it with pain medication, no pain , felt fantastic stopped taking it due to weight loss however pain has increased and is more intense since stopped taking it , bloody mucous in stools while taking it	F	37	14 days 1X D	12/27/2014
4	Arthritis due to multiple fractures	Servers stomach cramps, burping, diarrhoea ,extremely tired , nausea ,on the upside relieved bone and joint aches	Catch 22 relives body / joint/ arthritis pain but causes numerous side effects	F	37	3 days 50mg 3X D	12/22/2014
5	Severe arthritis	Very minor: I find that some days I cannot tolerate burn-your-mouth-hot spicy foods without a little Pepto. Most days there are no problems. The benefits FAR outweigh losing the extreme zip of a great, hot chili.	Diclo is my first line of defence against my arthritis. It is by far the most effective anti-inflammatory I've tried & also the one with the least side effects.	F	43	6 years 150 mg/day	12/5/2014 Email
1	Topical for pre-cancerous lesion	Nausea, diarrhea, gas, intestinal pain, upper stomach pain, acid indigestion. Symptoms started a week into taking it. I didn't connect the symptoms with the medication at first.	Stopped the drug 2 days ago, but side effects lingering.	F	47	2 weeks 3 percent	12/2/2014
1	Arthritis			F	57	3 months 75mg	11/30/2014

Figure 2: A screenshot of AskaPatient forum posts on Voltaren.

6.2 Machine Learning Approaches

Several machine learning approaches have been used successfully to do entity recognition in natural language text. For example, CRF classifiers have been used to identify medical concepts in Electronic Health Records (EHRs); however, social media text has very different characteristics and is typically noisier. Even though these techniques *learn* from the data, this does not necessarily mean that the performance will be comparable.

To implement this approach, we used the CRF classifier from the Stanford NER suite [Finkel et al., 2005][10]. A CRF classifier takes as input different features that are derived from the text. The features used in our implementation are listed in Table 6.

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 common in the most NLP applications, which assigns a B to the first token in a span, an I to all the other contiguous tokens in the span, and an O to all other tokens that do not belong to any span. This format does not support the notion of discontinuous spans and several solutions have been proposed in previous research to overcome this limitation. One of these is to treat discontinuous spans as several continuous spans and after classification use additional machine learning techniques to correctly reassemble them. Another alternative is to extend the BIO format with additional tags to represent the discontinuous spans. The latter approach has proved more successful in the CLEF and SemEval tasks and therefore has been used in our implementation.

Table 6: The features used in our CRF implementation.

Feature	Description
Bag of words	The words that surround each token.
N-grams	Creates features from letter n-grams (substrings of the word). In this implementation n was set to 6.
Word shape	Indicates the shape of the token, for example, if the token is composed of only lower case letters, upper case letters, or a combination.

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 4 shows an example of these types of an-

[10] http://nlp.stanford.edu/software/CRF-NER.shtml

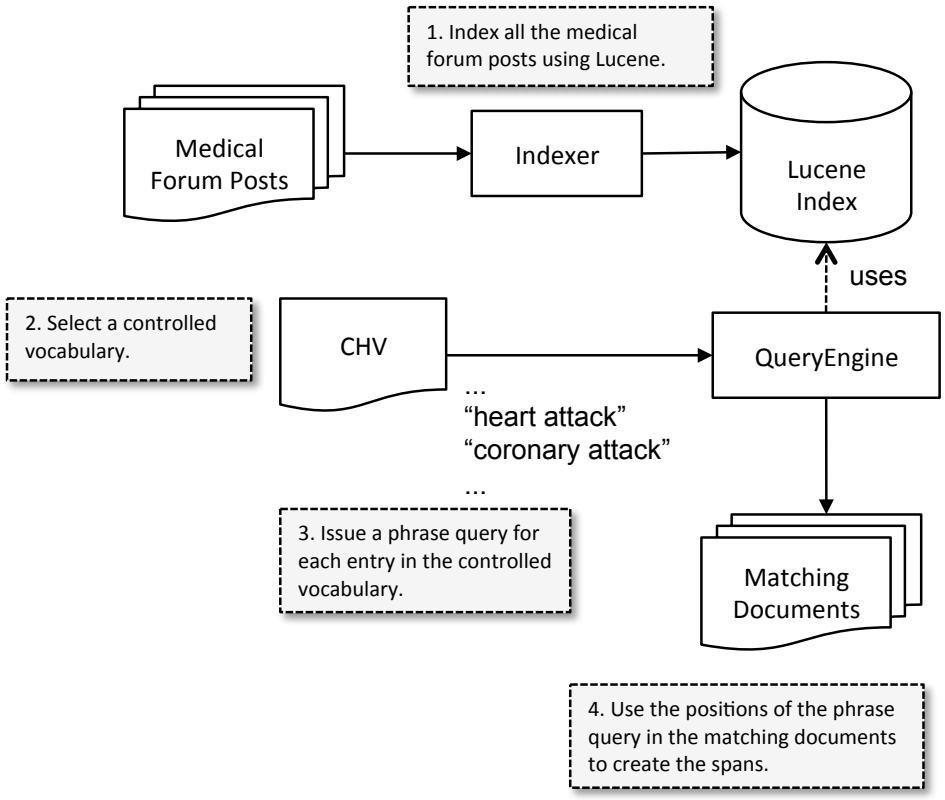


Figure 3: Diagram illustrating how the Lucene dictionary-based implementations work.

Table 7: The number of true positives (TP), false positives (FP), and false negative (FN) spans that are created from the process of transforming the ground truth spans into the extended BIO format and back. This is equivalent to having a perfect classifier.

Set	TP	FP	FN	Total
Training	6325	122	66	6513
Test	2618	50	26	2694
Total	8943	172	92	9207

notations and how they are represented in the extended BIO format.

Notice, however, that there is an obvious limitation with this approach: if several discontinuous spans occur in the same sentence, then it is impossible to represent them unambiguously. In order to determine how this limitation might affect the performance of the CRF approach in the social media dataset, a round trip transformation was performed, using the gold standard annotations, that is, the gold standard was transformed into the extended BIO format representation and then back to the original format. This is equivalent to having a perfect CRF classifier. Table 7 shows the number of correct (TP), incorrect (FP), and spurious (FN) spans created by the round trip process. In practice, the limitations of this format do not have a significant impact on the overall performance. Additional techniques to deal with ambiguous cases were not pursued and are left as future work.

One of the differences between this approach and the dictionary-based approaches is that the CRF classifier only identifies the spans that refer to drugs or ADRs and does not map them to the corresponding concepts. Therefore, the second part of the task has to be implemented independently.

Two approaches were explored. The first one is based on a traditional search method using the Vector Space Model (VSM). The Lucene search engine was used for this purpose. The target ontology was indexed by creating a document for each term and storing the corresponding concept id. This means that a concept with multiple synonyms generates multiple documents in the index. In this case, stemming and stop word removal were used. Then, the text of each span was used to query the index. When the span included multiple tokens the query was not required to match all of them. 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 [McBride et al., 2012], 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 [Sevenster et al., 2012]. It also supports other features such as spell checking and filtering based on hierarchies in the ontology. We used version 2.3.0 of Ontoserver, which is publicly available at `http://ontoserver.csiro.au:8080/`. The text in each span was used as a query. The parameters were set so that all terms were not required and, when dealing with SNOMED CT, the results were filtered so that

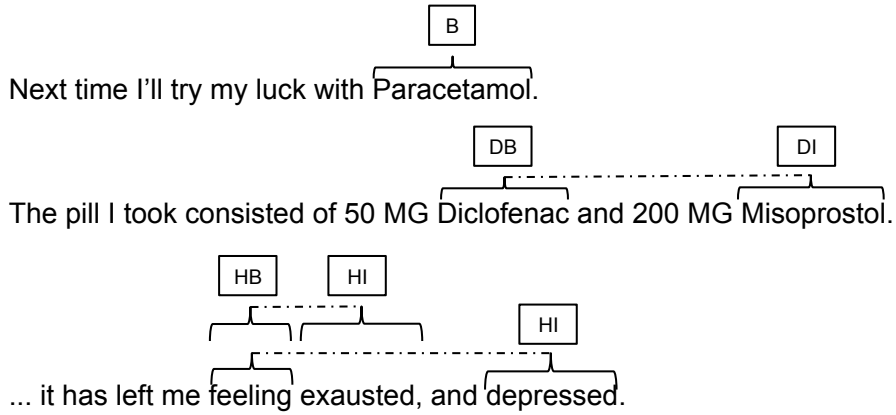


Figure 4: Examples of how annotation types are represented using the extended BIO format. The O annotations are not shown.

Table 8: A summary of the different methods that were evaluated.

Method	**Description**
MetaMap	The baseline method. MetaMap was used to identify and normalise the concepts in the social media text.
VSM + UMLS	A dictionary-based approach based on sliding window that uses the UMLS as the underlying controlled vocabulary.
VSM + CHV	A dictionary-based method based on sliding window that uses CHV, list of colloquial health terms.
VSM + SCT	A dictionary-based approach based on sliding window that uses SNOMED CT as the underlying controlled vocabulary. This implementation is used to identify ADRs.
VSM + AMT	A dictionary-based approach based on sliding window that uses AMT as the underlying controlled vocabulary. This implementation is used to identify drugs.
CRF + VSM	A mixed approach that uses a CRF classifier to identify the concept spans and a VSM implementation to map these spans to concepts in a controlled vocabulary (SNOMED CT for ADRs and AMT for drugs).
CRF + Ontoserver	A mixed approach that uses a CRF classifier to identify the concept spans and Ontoserver to map these spans to concepts in a controlled vocabulary.

only those concepts that belong to the Clinical Finding hierarchy were returned. When a query returned no results, the span was annotated as *concept_less*.

A summary of all the methods that were implemented is shown in Table 8.

6.3 Statistical Significance

To determine if the improvements obtained with any two different methods were statistically significant, we used McNemar's test [Davis et al., 2012]. This test is applied to paired nominal data using a 2×2 contingency table to determine if row and column marginal frequencies are equal. The contingency table is shown in Table 9, where A is the number of correct predictions by both methods; B is the number of correct predictions by Method 1 where Method 2 produced an incorrect prediction; C is the number of correct predictions by method 2 where method 1 produced an incorrect prediction; and D is the number of incorrect predictions by both methods.

Table 9: The contingency table used as input to McNemar's test, used to test statistical significance.

		Method 2	
		Correct	Wrong
Method 1	Correct	**A**	**B**
	Wrong	**C**	**D**

7 Results and Discussion

The results of the concept identification task are shown in Table 10. There are several noteworthy results. First, 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. Also, notice that in some cases the overall ranking provided by the F-Score value is different from the ranking provided by the accuracy value. In particular, when dealing with ADR identification, the MetaMap implementation has a higher accuracy than the VSM+UMLS implementation despite its precision, recall and F-Score being much lower. This happens because the VSM+UMLS implementation, despite producing more correct spans than the MetaMap implementation, also produces many more incorrect spans.

The task of identifying drugs is considerably different from the task of identifying ADRs because it usually involves less ambiguity. For example, trade products usually have no synonyms and therefore limit the number of ways a person can refer to them (this of course does not rule out misspellings, which are common in social media). Because of this, intuitively, this task should be easier than the task of identifying ADRs. The results show that the CRF implementation indeed performs better in this task that in the ADR identification task. Note also that MetaMap obtains very low preci-

Table 10: Evaluation results of the concept identification task, sorted by accuracy. Statistical significant difference with the next best performing method is indicated with $*$ ($p < 0.01$).

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

sion and recall. This is because the tool was not designed to identify drugs. Also, 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, that indicate that drug classes should be excluded. If the drug classes are mentioned frequently and are part of the underlying controlled vocabularies then this will create many false positives. In contrast, the CRF implementation is capable of identifying some of the common drug classes that are not annotated in the training set and is able to avoid creating false positives in most cases.

Table 11 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. Overall, however, the results are quite poor. This highlights two 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. For example, when using the VSM+CHV implementation, even if the term in the text appears in CHV, if this term is mapped to an incorrect concept in SNOMED CT, the implementation will produce an incorrect result. Even though this issue has not been explored in depth, some of the potential problems include mappings to concepts that are now inactive (and therefore will never appear in the gold standard) and mappings to concepts in other versions of SNOMED CT (for example SNOMED US, which shares a common subset with SNOMED AU but also includes some local concepts that will not appear in the Australian version).

For example, in the MetaMap implementation the concept 366981002 (Pain) is returned as the top concept for several spans and this concept has been replaced in the current version with 22253000 (Pain). It may be possible to automatically replace an inactive concept with the current concept that replaced it; however, this option was not attempted and is left as future work.

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 at all because a map between the other controlled vocabularies and AMT does not currently exist.

Finally, a strict evaluation of the full task was carried out, where a span was only considered correct if it matched the gold standard span exactly and was annotated with the same concept. This evaluation is important because a good free text annotation system will not only need to identify relevant spans but also annotate them correctly. The results for the full evaluation are shown in Table 12. The best performing system overall was the CRF implementation using Ontoserver for concept normalisation.

8 Conclusions and Future Work

Pharmacovigilance has passed the era where it would only rely on manual reports of potential drug adverse effects. Actively detecting signals of adverse drug reactions through automated methods of text mining consumer reviews is one of the emerging areas.

We conducted an empirical evaluation of different methods to automatically identify and normalise medical concepts in the domain of adverse drug reaction

Table 12: Results of the evaluation of the full task applied to ADRs, sorted by accuracy. Statistical significant difference with the next best performing method is indicated with $*$ ($p < 0.01$).

Entities	Name	Precision	Recall	F-Score	Accuracy
ADRs	VSM+UMLS	0.088	0.104	0.095	0.363
	MetaMap	0.041	0.029	0.034	0.468*
	VSM+CHV	0.218	0.106	0.143	0.590*
	CRF+VSM	0.564	0.327	0.414	0.702*
	VSM+AMT	0.572	0.332	0.420	0.706
	CRF+Ontoserver	0.771	0.376	0.506	0.764*
Drugs	VSM+UMLS	0.000	0.000	0.000	0.461
	VSM+AMT	0.163	0.758	0.269	0.605*
	VSM+CHV	0.000	0.000	0.000	0.814*
	MetaMap	0.000	0.000	0.000	0.814
	CRF+VSM	0.988	0.749	0.852	0.975*
	CRF+Ontoserver	0.988	0.773	0.867	0.977

Table 11: Results of the evaluation of the concept normalisation task. Baseline is MetaMap.

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

detection in medical forums. It included several methods commonly used in the ADR mining literature, as well as state-of-the-art machine learning methods that have been used in other domains. To our knowledge this is the first study to systematically compare the most common concept identification and normalisation approaches used in adverse effect mining from social media under a controlled setting. This is an important step in ADR signal detection which determines the effectiveness of automated systems in this domain.

The experimental results showed that the CRF implementation combined with Ontoserver outperformed all the other methods that were evaluated, including MetaMap and the dictionary-based methods. We believe that the availability of the new CADEC corpus and the empirical results shown in this paper will benefit other researchers working on ADR mining methods.

In the future, we plan to improve the CRF method with additional features, specifically domain specific features that are likely to improve recognition of ADRs in text.

Regarding the concept normalisation task, the results showed that there is still room for improvement. There are two avenues to explore. The concept normalisation could also be evaluated completely independently by using the spans in the gold standard as input. Second, to the best of our knowledge, existing concept normalisation implementations, including the ones implemented in this work, 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.

Regarding the evaluation, there are several ways that the current methods could be extended. In many use cases, generating the exact same span as in the gold standard is not relevant. However, the current definition of a *relaxed* match is too loose and might not work appropriately in certain situations. For example, when spans tend to be long and include multiple tokens, a single token overlap constitutes a positive relaxed match. An improvement over the relaxed matching criteria would be to consider the extent of the match by establishing a threshold based on either a ratio of characters or tokens that are required for the overlap to be considered a valid match. Using a high threshold would ensure that the systems under this relaxed evaluation are only producing spans with minimal differences (such as including prepositions before a noun or adjacent punctuation symbols, for example). Several metrics that could be adapted for this scenario have been proposed in the area of passage retrieval [Wade and Allan, 2005].

When evaluating the concept normalisation task, the current evaluation method only considers a span to be correct if it is assigned the same concept found in the gold standard. However, considering that the annotations in the CADEC corpus come from an ontology, if the span is annotated with a concept that is very close to the concept in the gold standard, say a parent concept, then considering the span to be completely wrong

seems too severe. Modifying the evaluation metric to consider this 'semantic distance' will likely give a better sense of the performance of the systems under evaluation.

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

A. Aronson. Effective mapping of biomedical text to the UMLS metathesaurus: The MetaMap program. In *AMIA Annual Symposium*, pages 17–21, 2001.

A. Bate and S. Evans. Quantitative signal detection using spontaneous ADR reporting. *Pharmacoepidemiology and Drug Safety*, 18(6):427–436, 2009.

A. Benton, L. Ungar, S. Hill, S. Hennessy, J. Mao, A. Chung, C. Leonard, and J. Holmes. Identifying potential adverse effects using the web: A new approach to medical hypothesis generation. *Journal of Biomedical Informatics*, 44(6):989–996, 2011.

W. Chapman, W. Bridewell, P. Hanbury, G. Cooper, and B. Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, 34(5):301–310, 2001.

B. Chee, R. Berlin, and B. Schatz. Predicting adverse drug events from personal health messages. In *AMIA Annual Symposium*, pages 217–226, Washington, DC, 2011.

G. Chrupala. Normalizing tweets with edit scripts and recurrent neural embeddings. In *The 52nd Annual Meeting of the Association for Computational Linguistics*, pages 680–686, Baltimore, MD, 2014.

K. Davis, C. Staes, J. Duncan, S. Igo, and J. C. Facelli. Identification of pneumonia and influenza deaths using the death certificate pipeline. *BMC Medical Informatics and Decision Making*, 12(1):37, 2012.

J. Ehsani, T. Jackson, and S. Duckett. The incidence and cost of adverse events in Victorian hospitals 2003–04. *The Medical Journal of Australia*, 184(11):551–555, 2006.

D. Evans, N. Brownlow, W. Hersh, and E. Campbell. Automating concept identification in the electronic medical record: An experiment in extracting dosage information. In *AMIA Annual Fall Symposium*, pages 388–392, Washington, DC, 1996.

J. R. Finkel, T. Grenager, and C. 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.

S. Fox and S. Jones. The social life of health information. *Washington, DC: Pew Internet & American Life Project*, pages 2009–12, 2009.

C. Friedman. Discovering novel adverse drug events using natural language processing and mining of the Electronic Health Record. In *Proceedings Of 12th Conference On Artificial Intelligence in Medicine: Artificial Intelligence in Medicine*, pages 1–5, Verona, Italy, 2009.

J. Gung. Using relations for identification and normalization of disorders: Team CLEAR in the ShARe/CLEF 2013 eHealth evaluation lab. In *The CLEF Workshop on Cross-Language Evaluation of Methods, Applications, and Resources for EHealth Document Analysis*, page 7 pages, Valencia, Spain, 2013.

H. Gurulingappa, A. Mateen-Rajput, and L. Toldo. Extraction of potential adverse drug events from medical case reports. *Journal Of Biomedical Semantics*, 3(1):15, 2012.

H. Hassan and A. Menezes. Social text normalization using contextual graph random walks. In *The 51st Annual Meeting of the Association for Computational Linguistics*, pages 1577–1586, Sofia, Bulgaria, 2013.

B. Hug, C. Keohane, D. Seger, C. Yoon, and D. Bates. The costs of adverse drug events in community hospitals. *Joint Commission Journal on Quality and Patient Safety*, 38(3):120–126, 2012.

A. Jimeno, E. Jimenez-Ruiz, V. Lee, S. Gaudan, R. Berlanga, and D. Rebholz-Schuhmann. Assessment of disease named entity recognition on a corpus of annotated sentences. *BMC Bioinformatics*, 9(Suppl 3):S3, 2008.

S. Karimi, A. Metke-Jimenez, M. Kemp, and C. Wang. CADEC: A corpus of adverse drug event annotations. *Journal of Biomedical Informatics*, (to appear), 2015a.

S. Karimi, C. Wang, A. Metke-Jimenez, R. Gaire, and C. Paris. Text and data mining techniques in adverse drug reaction detection. *ACM Computing Surveys*, (Accepted January 2015, to appear), 2015b.

S. Karimi, J. Yin, and J. Baum. Evaluation methods for statistically dependent text. *Computational Linguistics*, (Accepted November 2014, to appear), 2015c.

R. Leaman, L. Wojtulewicz, R. Sullivan, A. Skariah, J. Yang, and G. Gonzalez. Towards internet-age pharmacovigilance: Extracting adverse drug reactions from user posts to health-related social networks. In *The Workshop on Biomedical Natural Language Processing*, pages 117–125, Uppsala, Sweden, 2010.

R. Leaman, R. Khare, and Z. Lu. NCBI at 2013 ShARe/CLEF eHealth shared task: Disorder normalization in clinical notes with DNorm. In *The CLEF Workshop on Cross-Language Evaluation Of Methods, Applications, and Resources for EHealth Document Analysis*, page 9 pages, Valencia, Spain, 2013.

D. Lee, R. Cornet, F. Lau, and N. De Keizer. A survey of SNOMED CT implementations. *Journal Of Biomedical Informatics*, 46(1):87–96, 2013.

W. Ling, C. Dyer, W. A. Black, and I. Trancoso. Paraphrasing 4 microblog normalization. In *Conference on Empirical Methods in Natural Language Processing*, pages 73–84, Seattle, Washington, 2013.

X. Liu and H. Chen. AZDrugminer: An information extraction system for mining patient-reported adverse drug events in online patient forums. In *The International Conference on Smart Health*, pages 134–150, Beijing, China, 2013.

S. McBride, M. Lawley, H. Leroux, and S. Gibson. Using Australian Medicines Terminology (AMT) and SNOMED CT-AU to better support clinical research. *Studies in Health Technology and Informatics*, 178:144–149, 2012.

A. Metke-Jimenez, S. Karimi, and C. Paris. Evaluation of text-processing algorithms for adverse drug event extraction from social media. In *The 1st International Workshop on Social Media Retrieval and Analysis*, pages 15–20, Gold Coast, Australia, 2014.

J. G. Mork, O. Bodenreider, D. Demner-Fushman, R. I. Dogan, F.-M. Lang, Z. Lu, A. Neveol, L. Peters, S. E. Shooshan, and A. R. Aronson. Extracting Rx information from clinical narrative. *Journal of the American Medical Informatics Association*, 17(5):536–539, 2010.

S. Pradhan, N. Elhadad, B. South, D. Martinez, L. Christensen, A. Vogel, H. Suominen, W. Chapman, and G. Savova. Task 1: ShARe/CLEF eHealth evaluation lab 2013. In *The CLEF Workshop on Cross-Language Evaluation of Methods, Applications, and Resources for EHealth Document Analysis*, page 6 pages, Valencia, Spain, 2013.

S. Pradhan, N. Elhadad, W. Chapman, S. Manandhar, and G. Savova. SemEval-2014 task 7: Analysis of clinical text. In *The 8th International Workshop on Semantic Evaluation*, pages 54–62, Dublin, Ireland, 2014.

B. P. Ramesh, S. M. Belknap, Z. Li, N. Frid, D. P. West, and H. Yu. Automatically recognizing medication and adverse event information from food and drug administration's adverse event reporting system narratives. *JMIR Medical Informatics*, 2(1):e10, 2014.

E. Roughead and S. Semple. Medication safety in acute care in Australia: Where are we now? part 1: A review of the extent and causes of medication problems 2002-2008. *Australia and New Zealand Health Policy*, 6(18):12 pages, 2009.

H. Sampathkumari, X.-W. Chen, and B. Luo. Mining adverse drug reactions from online healthcare forums using Hidden Markov Model. *BMC Medical Informatics and Decision Making*, 14(91):18 pages, 2014.

M. Sevenster, R. van Ommering, and Y. Qian. Algorithmic and user study of an autocompletion algorithm on a large medical vocabulary. *Journal of Biomedical Informatics*, 45(1):107–119, 2012.

M. Sokolova and G. Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, 2009.

B. Tang, Y. Wu, M. Jiang, J. C. Denny, and H. Xu. Recognizing and encoding disorder concepts in clinical text using machine learning and vector space model. In *The CLEF Workshop on Cross-Language Evaluation Of Methods, Applications, and Resources for EHealth Document Analysis*, page 8 pages, Valencia, Spain, 2013.

O. Uzuner, I. Solti, F. X. F, and E. Cadag. Community annotation experiment for ground truth generation for the i2b2 medication challenge. *Journal of the American Medical Informatics Association*, 2010.

C. Wade and J. Allan. Passage retrieval and evaluation. IR 396, University of Massachusetts, 2005.

C. Yang, L. Jiang, H. Yang, and X. Tang. Detecting signals of adverse drug reactions from health consumer contributed content in social media. In *ACM SIGKDD Workshop on Health Informatics*, page 8 pages, Beijing, China, 2012.

Y. Zhang, J. Wang, B. Tang, Y. Wu, M. Jiang, Y. Chen, and H. Xu. UTH_CCB: A report for SemEval 2014–task 7 analysis of clinical text. In *The 8th International Workshop on Semantic Evaluation*, pages 802–806, Dublin, Ireland, 2014.