Medication and Adverse Event Extraction from Noisy Text

Xiang Dai[1,2], Sarvnaz Karimi[1], and Cecile Paris[1]

[1]CSIRO Data61, Marsfield, NSW, Australia
[2]School of Information Technologies, University of Sydney
{dai.dai, sarvnaz.karimi, cecile.paris}@data61.csiro.au

Abstract

We investigate the problem of extracting mentions of medications and adverse drug events using sequence labelling and non-sequence labelling methods. We experiment with three different methods on two different datasets, one from a patient forum with noisy text and one containing narrative patient records. An analysis of the output from these methods are reported to identify what types of named entities are best identified using these methods and, more specifically, how well the discontinuous and overlapping entities that are prevalent in our forum dataset are identified. Our findings can guide studies to choose different methods based on the complexity of the named entities involved, in particular in text mining for pharmacovigilance.

1 Introduction

An Adverse Drug Reaction (ADR) is an injury occurring after a drug (medication) is used at the recommended dosage, for recommended symptoms. The practice of monitoring the ADRs of pharmaceutical products is known as *pharmacovigilance* [Alghabban, 2004]. Different from controlled clinical trials which are mainly conducted before drugs are licensed for use, pharmacovigilance is especially concerned with identifying previously unreported adverse reactions. Text mining over different sources of information, such as electronic health records and patient reports on health forums, can be one way of finding such potential adverse reactions. This is the area to which our work contributes. We note that when causality between an adverse reaction and a medication is not known, it is referred to as Adverse Drug Event (ADE) [Karimi et al., 2015c].

Extracting mentions of drugs and adverse events from social media is difficult for two main reasons. First, social media text contains colloquial language and typographical mistakes. References to the names of drugs, diseases, and ADEs are particularly prone to misspellings. Second, a medical concept can be considered both an ADE or a symptom in different contexts. For example, *aches and pains* is a symptom in the context of "*I am only taking 75 mg a day and it is wonderful for relieving all of my aches and pains*", while it is an ADE in another post complaining of *severe fatigue with aches and pains*. We cast this concept extraction problem as a supervised Named Entity Recognition (NER) task, with drugs and ADEs as entity types.

Named entity recognition itself has long been studied in different domains, including the biomedical domain. There are a number of challenges that are associated with how named entities are expressed in text. These include entities expressed as *multiwords*, especially if the entities are *overlapping*, *discontinuous*, or *nested*. Table 1 lists some examples of entity types of different complexities. Entities that consist of a discontinuous sequence of tokens are one of the more challenging ones to recognise correctly, yet they constitute a large portion (over 10%) of adverse event mentions in the forum posts that we studied.

We are interested in the extraction of two named entities from free-text: medications or drugs, and adverse drug events. We divide named entities into two sets of complex and simple entities. *Complex* entities are those belonging to one of the categories of *nested*, *discontinuous*, and *overlapping* entities. *Multiword* entities that are continuous are also more difficult to identify than the single word ones. We therefore refer to the entities that are single words, as *simple* regardless of what class of entity, drug or ADE, they represent. Extraction of complex entities is particularly important in the

Xiang Dai, Sarvnaz Karimi and Cecile Paris. 2017. Medication and Adverse Event Extraction from Noisy Text. In *Proceedings of Australasian Language Technology Association Workshop*, pages 79−87.

Complexity of Entity	**Entity**	**Sentence**
Simple and multiword (Continuous, non-overlapping)	1. *Disorentatation*, 2. *trouble brathing*, 3. *extreme hot*, 4. *readness and sweeling*, 5. *itching*, and 6. *abominal cramps*	Disorentatation, trouble brathing, extreme hot, readness and sweeling, itching, later abominal cramps.
Continuous, overlapping	*pain in knee* (overlapping with *pain in foot*)	pain in knee and foot.
Discontinuous, non-overlapping	*liver blood test mildly elevated*	My Liver blood test are also mildly elevated.
Discontinuous, overlapping	*pain in foot*	Pain in knee and foot.

Table 1: Examples of complexities in entities. Note the misspellings and irregular text in the sentences.

biomedical domain where these entities are more popular [Kilicoglu et al., 2016].

In this paper, we evaluate three NER methods, one most popular and two most recent ones, for their capabilities in extracting the complex entities that exist in our noisy dataset of reports of adverse drug events. Our aim is to identify which of these methods more accurately extracts these entities, and whether the differences in complexity or type of entities guide what method to choose.

2 Related Work

Related studies are categorised into two: (1) methods for named entity recognition, both in general and in the biomedical area, and (2) approaches to concept extraction for pharmacovigilance. NER has a very long history in the area of natural language processing, going back to early 90s and information extraction tasks in the Message Understanding Conferences. Below we only review those methods that are directly relevant to our work. Concept extraction in the biomedical area also has been studied extensively, with some of the early work for biological concepts such as genes and proteins in the context of GENIA [Ohta et al., 2002]. We only review a subset of these, focusing on medications and their adverse events.

2.1 Named Entity Recognition

Named Entity Recognition (NER), the problem of identifying named entities in free-text, was originally and long focused on entity classes of person, location, organisation, and time. It then expanded to a large variety of entities, depending on the application domains, including biomedical [Ramakrishnan et al., 2008, Leaman and Gonzalez, 2008, Verspoor et al., 2012] and social media [Ritter et al., 2011].

NER is traditionally seen as a sequence labelling task. One of the most competitive models is Conditional Random Fields (CRF). It is applied in a number of NER systems, such as Stanford NER. Finkel et al. [2005], who propose one of the methods underlying Stanford NER, modify CRF by adding non-local structure using Gibbs sampling. This method maintains a sequence model structure and, at the same time, adds long-distance conditioning influences. This way, there is no need to enforce no-overlap constraints as some of the other NER methods do. We note that, in the biomedical domain, similar CRF-based NER tools have been developed that incorporate some of the biomedical ontologies. BANNER [Leaman and Gonzalez, 2008] is one example of such publicly available tools.

In one of the early studies where the complexity of named entities is specifically investigated, Downey et al. [2007] propose a Web NER method to locate a diverse set of entities that can be found from the web. They consider the task of NER as an n-gram detection of multiword units. Their method starts unsupervised and therefore does not bound itself to pre-defined entities. It then uses CRFs and Conditional Markov Models, and is tested on both simple and complex named entities. In that work, complex named entities are defined as entities such as names of books and movies, where then baseline systems used to fail. Their proposed method, called *LEX*, is particularly high performing for finding the continuous multiword entities on web pages since it is designed for such cases.

In the biomedical domain, entities can be complicated. Ramakrishnan et al. [2008] highlight the problem of dealing with *compound* entities which they define as those NEs that are composed of simpler entities, such as names of diseases, body parts, processes and substances. Another class of complex entities is *nested* entities which has received attention from the biomedical NLP community. Alex et al. [2007] study nested and discontinuous entities in the biomedical domain in the context of two corpora: GENIA [Ohta et al., 2002] and BioInfer [Pyysalo et al., 2007]. For example, the name of a DNA can be nested inside the name of an RNA protein. Alex et al. [2007] identify the problem to be how NEs used to be represented using the IOB (Inside, Outside, Beginning) representation. This representation does not allow tokens to belong to more than one entity. They experiment with extending this representation as well as with cascading and joint learning models. Their method, however, is unable to identify nested entities of the same type. Finkel and Manning [2009] propose a discriminative parsing-based method for nested named entity recognition, employing CRFs as its core.

Kilicoglu et al. [2016] identify the lack of training data for NER in biomedical domain for consumer health questions. They create an annotated corpus where entities can be ambiguous by having multiple types, being nested, multi-part or discontinuous. They recognise the problem of evaluating systems using some of these entity types and provide some recommendations on how to deal with them by using different entity representations.

Most recently, neural networks have been applied to the NER task. Crichton et al. [2017] investigate NER in biomedical area using convolutional neural networks. They experiment with a variety of models–such as single-task, multi-task, dependent multi-task and multi-output models–on 15 different datasets. They show that, on average, multi-task models were superior to single-task ones.

Liu et al. [2017] investigate entity recognition from the Informatics on the Integrating Biology and the Bedside (i2b2) corpora (2010, 2012, and 2014 NLP challenges) using Long-Short Term Memory (LSTM). For the concept extraction task of the i2b2 challenge 2010, they show improvements over CRF methods using an LSTM-based method that uses character embeddings. NEs are represented using BIOES (B-beginning of an entity, I-insider of an entity, O-outsider of an entity, E-end of an entity, S-a single-token entity). This is close to one of the methods we use in our work. However they do not provide any insight on extracting complex entities (if there were any) in their datasets.

2.2 Concept Extraction for Pharmacovigilance

Safety signal detection for pharmacovigilance from medical literature, electronic health records and medical forums have been studied in the past decade [Kuhn et al., 2010, Leaman et al., 2010, Benton et al., 2011, Liu and Chen, 2013, Karimi et al., 2015c, Henriksson et al., 2015, Zhao et al., 2015, Pierce et al., 2017]. One of the problems in generating such signals from text is the extraction of relevant concepts, such as medications, adverse events, and patient information. Named entity recognition therefore has been studied as one of the methods to extract these relevant concepts.

Nikfarjam et al. [2015] investigate ADE extraction from a medical forum, DailyStrength, and Twitter, using Conditional Random Fields (CRFs). To train the CRFs, they use word embeddings as one of the features created based on the forum and Twitter data. The entity representation method in their study is IOB (Inside, Outside, and Beginning).

Sarker and Gonzalez [2015] consider ADE extraction as a classification task using multiple corpora: a patient forum, Twitter, and medical case reports. They use these datasets for training classifiers, including SVM, Maximum Entropy and Naïve Bayes. The combination of different corpora for training leads to improved classification accuracy. Klein et al. [2017] also extend the work by providing classification baselines as well as making the Twitter data available for further research. This dataset however does not contain discontinuous, overlapping or nested entities and therefore is not used in our study. Another problem with using the Twitter dataset is that it changes as based on the availability of tweets at the time of crawling, making it difficult for comparisons of the reported results in the literature.

[Karimi et al., 2015b, Metke-Jimenez and Karimi, 2016] investigate both dictionary-based and machine learning approaches based on CRFs for the identification of medical concepts, including drugs and ADEs. Their CRF models outperform most lexicon-based methods popular in this domain on a corpus from medical forums, called CADEC [Karimi et al., 2015a]. They argue that some of the discontinuous entities in the corpus are responsible for a portion of their errors.

Cocos et al. [2017] develop an BiLSTM model that used word embeddings from a large Twitter corpus to identify ADEs in tweets. They compare their method to CRFs (from CRFSuite software [Okazaki, 2007]) and lexicon-based methods showing improvements using BiLSTMs. We note improvements are seen in recall values, as opposed to higher precision that is achieved using CRFs. Their experiments show that static semantic information learned from a large, generic dataset through word embeddings is the only setting in BiLSTM that leads to higher F-Score. Since their focus is not on identifying complex entities such as discontinuous ones, they use a straightforward IO schema (Inside, Outside) for representing the named entities. Their error analysis identifies constituent phrases (multiword ADEs) as one source of errors made by their method.

3 Datasets

We use two datasets for our evaluations: *CADEC* [Karimi et al., 2015a] and the *i2b2 2009 medication challenge* [Uzuner et al., 2010]. CADEC consists of 1,250 posts from the medical forum AskaPatient. These posts were manually annotated by medical experts and a clinical terminologist for drugs, ADEs, diseases, symptoms, and findings. Among all 9,111 annotated entities, there are 1,800 drug entities and 6,318 ADE entities. The rest of the entities are ignored in this study.

Table 2 lists the overall statistics of the entities in the two datasets. In CADEC, many of the entities, especially ADEs, consist of discontinuous and overlapping spans. Different from other NER corpus, where the longest possible spans are identified as single entities, CADEC has many fine-grained entities, each of which can be referred to a specific medical concept in medicine terminology vocabularies. For example, in the sentence *Pain in hip, lower back, knees & elbow*, there are four ADEs: *Pain in lower back*, *Pain in knees*, *Pain in elbow*, and *Pain in hip*, corresponding to four different concepts in SNOMED Clinical Terms (SNOMED CT) [1].

The i2b2 medication extraction challenge dataset focuses on the identification of medications and medication-related information, such as their dosages, modes of administration, frequencies, durations, and reasons for administration, in discharge summaries. In our work, we identify only medication mentions. These include names, brand names, generics, and collective names of prescription substances, over-the-counter medications, and other biological substances.

There is a total of 1,249 discharge summaries in the i2b2 dataset. The i2b2 organisers first released a detailed annotation guideline along with a small set of ten annotated summaries. They then challenged the participants to collectively develop the gold standard annotations for 251 summaries which were used as test data. This was to encourage the development of unsupervised or semi-supervised methods. Here, we combine all these 261 summaries as labelled data.

4 Methods

4.1 Extended BIO Representation

To deal with discontinuous and overlapping spans, we use an extended version of the standard BIO chunking representation proposed by [Metke-Jimenez and Karimi, 2016]. In this representation, four additional prefixes are used: DB, DI, HB, and HI. The following details all the prefixes used in our work together with examples from our data.

O Outside concept. All tokens outside the concepts in which we are interested are labelled as *O*.

B- Begin of concept, for continuous and non-overlapping spans.

I- Continuation of concept, for continuous and non-overlapping spans.

DB- Begin of concept, for discontinuous and non-overlapping spans. For example, in the sentence *every joint in my body is in pain*, the ADE *joint pain* is a discontinuous span, so the label for the token *joint* is *DB-ADE*.

DI- Continuation of concept, for discontinuous and non-overlapping spans. The label for the token *pain* in the previous example is *DI-ADE*.

HB- Begin of concept, for discontinuous and overlapping spans that share one or more tokens with other concepts. For example, in the sentence *it has left me feeling exausted,*

[1] https://www.snomed.org/snomed-ct

	CADEC		i2b2
Entity	Drug	ADE	Drug
All	1800	6318	8850
Discontinuous, non-overlapping	1 (0.05)	82 (1.30)	0 (0.00)
Discontinuous, overlapping	1 (0.05)	593 (9.38)	0 (0.00)
Continuous, non-overlapping	1797 (99.83)	5311 (84.06)	8850 (100.00)
Continuous, overlapping	1 (0.05)	332 (5.25)	0 (0.00)
Multiword	141 (7.83)	4574 (72.40)	2181 (24.64)
Single word	1659 (92.17)	1744 (27.60)	6669 (75.36)

Table 2: Overall statistics of the number of entities and their breakdown based on their complexity in the datasets. Numbers in brackets are percentages.

and depressed, two ADEs *feeling exausted* and *feeling depressed* overlap and share one common token *feeling*, so the label of token *feeling* is *HB-ADE*.

HI- Continuation of concept, for discontinuous and overlapping spans. The label for the token *exausted* and *depressed* in the previous example are both *HI-ADE*.

4.2 Sequence Labelling: CRF and Bi-LSTM Model

We used two different sequence labelling methods: one based on conditional random fields as implemented in Stanford NER (version 3.8.0) [Finkel et al., 2005] and another based on a deep learning method implemented in NeuroNER [Dernoncourt et al., 2017a]. NeuroNER uses Bidirectional LSTM (Bi-LSTM) neural networks. It contains three layers: a character-enhanced token-embedding layer, a label prediction layer, and a label sequence optimisation layer [Dernoncourt et al., 2017b]. BiLSTMs are known to take into account context from both left and right of a token, and they can handle sequences of variable size.

Word embeddings can be provided as input to NeuroNER. To create the embedding, the documents are first tokenised using the spaCy tokeniser [2]. A token will be taken as input of the word embedding layer, and its vector representation will be generated as the output.

4.3 Non-Sequence Labelling

A recent work by Xu et al. [2017] propose a non-sequence labelling based on the FOFE (Fixed-Size Ordinally-Forgetting Encoding) representation [Zhang et al., 2015] which they call FOFE-NER. This is using a local detection approach where the left and right contexts of tokens created using FOFE are represented to a deep feed-forward neural network. This method is very powerful in capturing immediate dependencies in the tokens and therefore should recognise multiword entities well. We directly apply this method to our problem with all the features that are proposed including character and word level features, to examine its effectiveness in our problem.

5 Experiments

We experiment using the three methods (CRF, Bi-LSTM, and non-sequence labelling) on the two datasets (CADEC, i2b2 2009) in two settings: (1) an overall comparison of the methods; and (2) an in-depth comparison based on the complexity of the named entities.

These methods are employed using a strategy called one-vs-all in which we train separate models for each entity type. For example, a model to identify *drug* is created yielding only two kinds of results: *drug* and *not-a-drug*.

Different methods of generating word embeddings using Wikipedia, MEDLINE, same corpus, and random embeddings are investigated. In our experiments they all generate similar results given our embeddings were used in a dynamic setting. Dynamic embeddings are re-calculated during the training phase. This is in line with the findings reported in [Karimi et al., 2017]. In the experimental results, we report on random word embeddings.

To evaluate, we run 10-fold cross-validation and report the average scores. Evaluation metrics used here are precision, recall, and F-score, all calculated based on the exact matches of extracted entities with the gold data.

[2] `https://spacy.io/` (Version 2.0, accessed 15 Nov 2017)

Dataset	Entity	Method	Precision	Recall	F-Score
CADEC	Drug	CRF	**95.1** ± **2.5**	79.1 ± 16.1	85.5 ± 11.4
		Bi-LSTM	92.9 ± 2.0	**92.2** ± **1.5**	**92.5** ± **0.7**†
		Non-sequence Labelling	88.6 ± 4.9	89.8 ± 1.8	89.1 ± 2.1
	ADE	CRF	67.5 ± 5.1	57.7 ± 2.9	62.1 ± 3.6
		Bi-LSTM	**73.4** ± **3.9**	**64.9** ± **4.4**	**68.7** ± **2.1**†
		Non-sequence Labelling	62.9 ± 3.8	61.6 ± 1.8	62.1 ± 1.0
i2b2	Drug	CRF	**93.5** ± **1.0**	85.7 ± 2.5	89.4 ± 1.7
		Bi-LSTM	93.2 ± 1.2	89.7 ± 1.3	**91.4** ± **0.6** †
		Non-sequence Labelling	84.4 ± 4.2	**90.2** ± **2.4**	87.1 ± 2.7

Table 3: Effectiveness of the different methodologies with their standard deviations over 10-fold cross-validations. Significant differences are shown with a † (p-value$<$ 0.05).

		Bi-LSTM	
		✓	✗
CRF	✓	2811	498
	✗	885	2124

		Bi-LSTM	
		✓	✗
Non-Seq	✓	2315	198
	✗	1321	2484

		CRF	
		✓	✗
Non-Seq	✓	2333	180
	✗	916	2889

Table 4: Comparisons of different methods on the CADEC dataset (ADE entities) based on extracted entities that were correct (✓) or incorrect (✗).

		Bi-LSTM	
		✓	✗
CRF	✓	7378	233
	✗	498	741

		Bi-LSTM	
		✓	✗
Non-Seq	✓	7561	406
	✗	315	568

		CRF	
		✓	✗
Non-Seq	✓	7317	649
	✗	296	588

Table 5: Comparisons of different methods on the i2b2 dataset (drug entities) based on extracted entities that were correct (✓) or incorrect (✗).

5.1 Sequence Labelling versus Non-Sequence Labelling NER

We compare the two sequence labelling methods, CRF model using Stanford NER and Bi-LSTM-based NER implemented in NeuroNER. We also compare them with a non-sequence labelling method [Xu et al., 2017] that uses the FOFE representation [Zhang et al., 2015] (FOFE-NER). Results are shown in Table 3.

On the CADEC dataset, NeuroNER using Bi-LSTMs perform best for both drug and ADE entities. The only exception is precision for drugs, where the CRF model outperform Bi-LSTM by 2.2%. This is not however statistically significant (Kruskal-Wallis H-test and T-test). For the i2b2 dataset, again the Bi-LSTM's overall F-score is higher than that of the other two methods, except that CRF and Non-sequence labelling methods show slightly higher results in precision and recall, respectively. The differences, again, are not statistically significant.

We then further breakdown these results to identify the overlap between the three methods in terms of correctly identifying NEs, incorrectly identifying them or missing them. This is to determine whether there is the potential in these systems to be used together if the errors they make are different. We show these confusion matrices in Tables 4 and 5. For CADEC ADEs, there was an almost equal number of both correct and both incorrect cases for all the combinations. This shows the difficulty of extracting the ADEs. The non-sequence labelling method has a larger number of incorrect entities extracted compared to the other two approaches. In the i2b2 dataset (drug entities only), non-sequence labelling makes fewer mistakes that the other systems. We can also infer that these systems are similar (correct or incorrect) in approximately 90% of the time. There is thus a ceiling of 10% that combining the two methods could improve the effectiveness.

5.2 Complex Named Entities

In both the CADEC and i2b2 datasets, drug names rarely have overlapping or discontinuous properties. In contrast, it is common for ADEs to be discontinuous or overlapped. 925 out of 6318 ADEs in CADEC are overlapped, while 675 ones are discontinuous. In this section, we focus on the analysis of the effectiveness of the different methods on identifying these discontinuous or overlapped ADEs.

For the first set of experiments, we separate entities based on their complexity: overlapping, discontinuous, continuous multiword, and simple (single word). We then evaluate the output of the NER systems only based on those entities that fall into those categories. Results are shown in Table 6. Note that we mix both entity types of drugs and ADEs. Surprisingly, the CRF method is more successful in identifying overlapping and discontinuous entities. We note that the evaluations are strict: if two entities are overlapped, we expect both to be found to consider the system successful. We use the same strict criteria for the discontinuous entities: If a system only finds one of the two entities, it is not. For CADEC, continuous multiwords are more successfully identified by the non-sequence labelling model, while Bi-LSTM outperforms the other methods for simple entities. Multiwords for the i2b2 dataset do not differentiate the three methods as much, with CRF being slightly better than BiLSTM. Simple entities are equally identifiable for all the three methods too, with a slight win for non-sequence labelling.

One problem for the non-sequence labelling method is its tendency to extract long strings as one entity. For example, in the sentence *HORRIBLE muscle pains, horrible back spasms, spasms in leg muscles, nausea, vomitting, pain so bad that I could hardly walk or sit*, the ADE entities are: (1) *HORRIBLE muscle pains*, (2) *horrible back spasms*, (3) *spasms in leg muscles*, (4) *nausea*, (5) *vomitting* and (6) *pain*. These should be extract as separate entities. However, the non-sequence labelling method extracts the sequence of them as one entity which contributes to both one false positive and several false negatives.

6 Conclusions and Future Work

We presented a comparison of three named entity recognition (NER) methods for extraction of medications and adverse drug events. This is a task important in the context of pharmacovigilance, especially to extract information from consumer reports in health forums. Three methods–CRF, Bi-LSTM, and a non-sequence labelling method–were chosen based on their popularity, availability, and recency, as well as representing three different approaches to the NER task. We compared these methods based on how they deal with complexity in named entities, that is how they handle entity overlaps and discontinuity, as well as multiwords. Our experiments showed that the non-sequence labelling method can best extract continuous multiword entities, while CRF using Stanford NER is more successful for discontinuous entities.

Our next steps are to verify these results using other biomedical datasets, with different entity and document types. We are also interested in comparing these methods on nested entities which CADEC and i2b2 2009 did not contain. There are other methods that we should investigate, including joint models. Incorporating medical ontologies such as SNOMED CT and MedDRA can also potentially inform a system that deal with biomedical concepts, and in particular adverse events. Other potential methods to investigate are those that incorporate syntactic and semantic parsing of the sentences as well as tree-structure of the entities into account [Finkel and Manning, 2009, Dinarelli and Rosset, 2011, 2012].

Acknowledgments

Authors would like to thank Anthony Nguyen and Hamed Hassanzadeh (CSIRO e-Health Research Centre) as well as Stephen Wan and Mac Kim (CSIRO Data61) for earlier discussions on this work.

Dataset	Complexity	Method	Precision	Recall	F-Score
CADEC	Overlapping	CRF	21.4	15.3	17.8
		Bi-LSTM	–	3.5	–
		Non-Seq.	–	0.0	–
	Discountinuous	CRF	21.4	1.5	2.8
		Bi-LSTM	13.8	2.3	3.9
		Non-Seq.	–	0.0	–
	Multiword	CRF	57.4	46.4	51.4
		Bi-LSTM	60.9	62.6	61.7
		Non-Seq.	62.3	66.5	64.3
	Simple	CRF	78.0	63	69.7
		Bi-LSTM	79.6	68.5	73.6
		Non-Seq.	61.9	51.8	56.4
i2b2	Multiword	CRF	91.1	82.6	86.7
		Bi-LSTM	85.0	86.8	85.9
		Non-Seq.	82.8	81.5	82.1
	Simple	CRF	94.4	86.8	90.4
		Bi-LSTM	94.9	89.6	92.2
		Non-Seq.	91.1	92.4	91.7

Table 6: Breakdown of the effectiveness of NER methods based on entity complexity.

References

B. Alex, B. Haddow, and C. Grover. Recognising nested named entities in biomedical text. In *BioNLP*, pages 65–72, Prague, Czech Republic, 2007.

A. Alghabban. *Dictionary Of Pharmacovigilance*. Pharmaceutical Press, 2004.

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. *J. Biomed. Inform.*, 44(6):989–996, 2011.

A. Cocos, A. Fiks, and A. Masino. Deep learning for pharmacovigilance: Recurrent neural network architectures for labeling adverse drug reactions in Twitter posts. *J Am Med Inform Assoc.*, 24(4):813–821, 2017.

G. Crichton, S. Pyysalo, B. Chiu, and A. Korhonen. A neural network multi-task learning approach to biomedical named entity recognition. *BMC bioinformatics*, 18(1):368, 2017.

F. Dernoncourt, J. Y. Lee, and P. Szolovits. NeuroNER: an easy-to-use program for named-entity recognition based on neural networks. In *EMNLP*, pages 97–102, Copenhagen, Denmark, 2017a.

F. Dernoncourt, J. Y. Lee, Ö. Uzuner, and P. Szolovits. De-identification of patient notes with recurrent neural networks. *J Am Med Inform Assoc.*, 24(3):596–606, 2017b.

M. Dinarelli and S. Rosset. Models cascade for tree-structured named entity detection. In *IJCNLP*, pages 1269–1278, Chiang Mai, Thailand, 2011.

M. Dinarelli and S. Rosset. Tree-structured named entity recognition on OCR data: Analysis, processing and results. In *LREC*, pages 1266–1272, Istanbul, Turkey, 2012.

D. Downey, M. Broadhead, and O. Etzioni. Locating complex named entities in web text. In *IJCAI*, pages 2733–2739, Hyderabad, India, 2007.

J. R. Finkel and C. Manning. Nested named entity recognition. In *EMNLP*, pages 141–150, Singapore, 2009.

J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by Gibbs sampling. In *ACL*, pages 363–370, Ann Arbor, MI, 2005.

A. Henriksson, M. Kvistab, H. Dalianis, and M. Duneld. Identifying adverse drug event information in clinical notes with distributional semantic representations of context. *J. Biomed. Inform.*, 57: 333–349, 2015.

S. Karimi, A. Metke-Jimenez, M. Kemp, and C. Wang. CADEC: A corpus of adverse drug event annotations. *J. Biomed. Inform.*, 55:73–81, 2015a.

S. Karimi, A. Metke-Jimenez, and A. Nguyen. CADEminer: A system for mining consumer reports on adverse drug side effects. In *ESAIR*, pages 47–50, Melbourne, Australia, 2015b.

S. Karimi, C. Wang, A. Metke-Jimenez, R. Gaire, and C. Paris. Text and data mining techniques in adverse drug reaction detection. *ACM Comput. Surv.*, 47(4): 56:1–56:39, May 2015c.

S. Karimi, X. Dai, H. Hassanzadeh, and A. Nguyen. Automatic diagnosis coding of radiology reports: A comparison of deep learning and conventional classification methods. In *BioNLP*, pages 328–332, Vancouver, Canada, 2017.

H. Kilicoglu, A. B. Abacha, Y. Mrabet, K. Roberts, L. Rodriguez, S. E. Shooshan, and D. Demner-Fushman. Annotating named entities in consumer health questions. In *LREC*, Portorož, Slovenia, 2016.

A. Klein, A. Sarker, M. Rouhizadeh, K. O'Connor, and G. Gonzalez. Detecting personal medication intake in Twitter: An annotated corpus and baseline classification system. In *BioNLP*, pages 136–142, Vancouver, Canada, 2017.

M. Kuhn, M. Campillos, I. Letunic, L. Jensen, and P. Bork. A side effect resource to capture phenotypic effects of drugs. *Mol. Syst. Biol.*, 6(1):Article 343, 2010.

R. Leaman and G. Gonzalez. BANNER: An executable survey of advances in biomedical named entity recognition. In *Pac. Symp. Biocomput.*, pages 652–663, Hawaii, 2008.

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 *BioNLP*, pages 117–125, Uppsala, Sweden, 2010.

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

Z. Liu, M. Yang, X. Wang, Q. Chen, B. Tang, Z. Wang, and H. Xu. Entity recognition from clinical texts via recurrent neural network. *BMC Med Inform Decis Mak.*, 17(2):53–61, 2017.

A. Metke-Jimenez and S. Karimi. Concept identification and normalisation for adverse drug event discovery in medical forums. In *workshop on Biomedical Data Integration and Discovery*, Kobe, Japan, 2016.

A. Nikfarjam, A. Sarker, K. O'Connor, R. E. Ginn, and G. Gonzalez. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *J Am Med Inform Assoc.*, 22(3):671–681, 2015.

T. Ohta, Y. Tateisi, and J.-D. Kim. The GENIA corpus: An annotated research abstract corpus in molecular biology domain. In *HLT*, pages 82–86, San Diego, CA, 2002.

N. Okazaki. CRFsuite: a fast implementation of conditional random fields (CRFs), 2007. URL `http://www.chokkan.org/software/crfsuite/`.

C. Pierce, K. Bouri, C. Pamer, S. Proestel, H. Rodriguez, H. Van Le, C. Freifeld, J. Brownstein, M. Walderhaug, R. Edwards, and N. Dasgupta. Evaluation of Facebook and Twitter monitoring to detect safety signals for medical products: An analysis of recent FDA safety alerts. *Drug Saf.*, 40(4): 317–331, 2017.

S. Pyysalo, F. Ginter, J. Heimonen, J. Björne, J. Boberg, J. Järvinen, and T. Salakoski. Bioinfer: a corpus for information extraction in the biomedical domain. *BMC bioinformatics*, 8(1):50, 2007.

C. Ramakrishnan, P. Mendes, R. d. Gama, G. Ferreira, and A. Sheth. Joint extraction of compound entities and relationships from biomedical literature. In *IEEE/WIC/ACM WI-IAT*, pages 398–401, 2008.

A. Ritter, S. Clark, Mausam, and O. Etzioni. Named entity recognition in tweets: An experimental study. In *EMNLP*, pages 1524–1534, Edinburgh, UK, 2011.

A. Sarker and G. Gonzalez. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *J. Biomed. Inform*, 53:196–207, 2015.

O. Uzuner, I. Solti, and E. Cadag. Extracting medication information from clinical text. *J Am Med Inform Assoc.*, 17(5):514–518, 2010.

K. Verspoor, K. B. Cohen, A. Lanfranchi, C. Warner, H. Johnson, C. Roeder, J. Choi, C. Funk, Y. Malenkiy, M. Eckert, N. Xue, W. Baumgartner, M. Bada, M. Palmer, and L. Hunter. A corpus of full-text journal articles is a robust evaluation tool for revealing differences in performance of biomedical natural language processing tools. *BMC Bioinformatics*, 13(207), 2012.

M. Xu, H. Jiang, and S. Watcharawittayakul. A local detection approach for named entity recognition and mention detection. In *ACL*, pages 1237–1247, Vancouver, Canada, 2017.

S. Zhang, H. Jiang, M. Xu, J. Hou, and L. Dai. The fixed-size ordinally-forgetting encoding method for neural network language models. In *ACL*, pages 495–500, Beijing, China, 2015.

J. Zhao, A. Henriksson, and H. Boström. Cascading adverse drug event detection in electronic health records. In *Data Science and Advanced Analytics*, pages 1–8, 2015.