

The logo for DATA 61, featuring the text "DATA" above "61" in white, enclosed within a teal-colored hexagonal border.

DATA
61

Medication and Adverse Event Extraction from Noisy Text

Xiang Dai, Sarvnaz Karimi, and Cecile Paris

December 2017

www.data61.csiro.au



Pharmacovigilance: for safer drugs



- Pharmacovigilance relates to monitoring Adverse Drug Reactions (ADR), especially previously unreported reactions.
- An ADR is an injury occurring after a medication is used at the recommended dosage, for recommended symptoms.
- When causality between an ADR and a medication is not known, it is referred to as Adverse Drug Event (ADE).



Where is raw data from?



Official Channels have the issue of under-reporting (less than 10% of ADRs are reported ¹).

Unstructured data, such as medical reports or social network data, are becoming the primary data source.

The screenshot shows the top portion of the FDA MedWatch Voluntary Report form. At the top is the FDA U.S. Food & Drug Administration logo and a navigation bar with links for Home, Food, Drugs, Medical Devices, Radiation-Emitting Products, and Vaccines, Blood & Biologics. Below this is a progress indicator with six steps: 1. About Problem, 2. About Device, 3. About Product, 4. About Patient, 5. About Reporter, and 6. Review & Submit. The 'About Problem' section is active and contains the following text and form elements:

About Problem
** Required information*

What kind of problem was it?
(Check all that apply)

- Were hurt or had a bad side effect (including new or worsening symptoms)
- Used a product incorrectly which could have or led to a problem
- Noticed a problem with the quality of the product
- Had problems after switching from one product maker to another maker

Did any of the following happen?
(Check all that apply)

- Hospitalization - admitted or stayed longer
- Required help to prevent permanent harm (for medical devices only)
- Disability or health problem
- Birth defect
- Life-threatening
- Death (include date) (mm/dd/yyyy)
- Other serious/important medical incidents (please describe)

Date the problem occurred (mm/dd/yyyy):
mm/dd/yyyy

Tell us what happened and how it happened: *
(Include as many details as possible)

A large text input area is provided for the user to enter details.

¹Hazell and Shakir, 2006. Under-reporting of adverse drug reactions. Drug Safety.

Given a sentence, we want to identify drugs and ADEs



Example sentence

i was put on this medication because i was taking such high doses of **Advil** to quell the **inflammation in my neck and back muscles** that it caused me to have a **GI bleed**.

We solve this as an NER problem.

Concept Extraction for Pharmacovigilance



- Methods
 - ▶ Lexicon-based and rule-based methods.
 - ▶ Machine learning methods
 - Classification: whether a text contains an ADE.
 - NER: extract ADEs from free-text
- Challenges
 - ▶ Distinguish between ADEs and symptoms which are supposed to be cured by drugs.
 - I am only taking 75 mg a day and it is wonderful for relieving all of my **aches and pains**.
 - ▶ ADEs consist of complex spans.

Our focus: complex entities



Complexity of Entity	Samples from CADEC dataset
Continuous, non-overlapping	Disorientation, trouble breathing, extreme hot, redness and swelling, itching, later abdominal cramps.
Continuous, overlapping	pain in knee and foot.
Discontinuous, non-overlapping	My Liver blood test are also mildly elevated.
Discontinuous, overlapping	pain in knee and foot.

Our contribution



- We evaluate three NER methods for their capabilities in extracting the complex entities.
 - ▶ Sequence labeling approaches
 - CRF
 - BiLSTM ¹
 - ▶ Non-sequence labeling approach ²
- 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.

¹Dernoncourt et al., 2017. De-identification of patient notes with recurrent neural networks. JAMIA

²Xu et al., 2017. A local detection approach for named entity recognition and mention detection. ACL

Overall statistics of entities in datasets



	CADEC		i2b2
Text	Patient post		Discharge summary
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)

- In both datasets, drug names rarely have overlapping or discontinuous properties. In contrast, it is common for ADEs to be discontinuous or overlapped.

Segment representations



BIO or BIOES¹ schemas can not be used in discontinuous or overlapping scenarios. Here is a simple example with two entities: *pain in knee* and *pain in foot*.

- BIO

pain	in	knee	and	foot
<hr/>				
B	I	I?O?	O	O?I?

- BIOES

pain	in	knee	and	foot
<hr/>				
B	I	E?O?	O	O?E?

¹B: Beginning; I: Inside; O: Outside; E: Ending; S: Single

Extended BIO representation



DB-/DI- Begin/continuation of concept, for discontinuous and non-overlapping spans.

every joint in my body is in pain
O DB O O O O O DI

One entity: *joint pain*.

HB-/HI- Begin/continuation of concept, for discontinuous and overlapping spans.

it has left me feeling exhausted , and depressed
O O O O HB HI O O HI

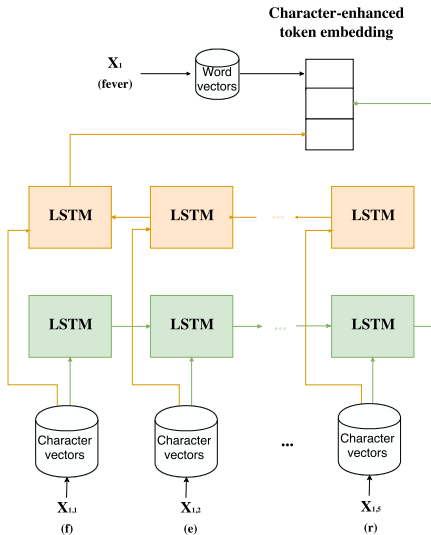
Two entities: *feeling exhausted* and *feeling depressed*.

Sequence labeling methods



1. Conditional random field (CRF)
2. Neural network based method
 - ▶ Character-enhanced token embedding layer
 - ▶ Label prediction layer: BiLSTM
 - ▶ Label sequence optimisation layer

Embedding layer



Prediction and optimisation layers

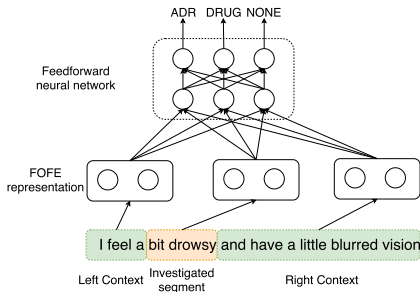


- The prediction BiLSTM layer takes as input the sequence of vectors, and outputs the probabilities for each token belonging to particular labels.
- Instead of selecting the label that has the highest probability for each token, optimisation layer models the **dependencies between two subsequent labels** through a transition probability matrix.

Non-sequence labeling approach



- Given a sentence $[t_1, t_2, t_3, \dots, t_n]$, all segments whose lengths are between 1 and k will be investigated.
- Apply a feedforward neural network on top of FOFE-based features ¹.



¹Zhang et al., 2015. The Fixed-Size Ordinally-Forgetting Encoding Method for Neural Network Language Models. ACL & IJCNLP

Overall effectiveness of different methods



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 Labeling	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 Labeling	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 Labeling	84.4 ± 4.2	90.2 ± 2.4	87.1 ± 2.7

- On both datasets, Bi-LSTMs perform better than two other methods.

Effectiveness of different methods based on entity complexity



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	

Conclusion



- A comparison of three NER methods for extraction of medications and ADEs.
 - ▶ BiLSTM: an overall effective model, but fails to directly extract discontinuous and overlapping entities.
 - ▶ CRF: more successful for discontinuous entities.
 - ▶ Non-sequence labeling method: tends to extract long strings as one entity due to its natural mechanism.
- Future work
 - ▶ Other methods to utilise syntactic structure
 - ▶ Other biomedical datasets
 - ▶ Nested entities