Benchmark analysis of baseline methods for ADR detection in social media – A technical report from Web-RADR, WP2b-ADE detection

Web/RADR&WP2b. Responsible author Johan Ellenius/UMC, 2016-01-11

Summary

The purpose of this technical report is to describe the performance of already existing methods that exist within the Web-RADR consortium for detection of ADRs in social media posts, in order to establish a base line for further reference. A predictive algorithm developed by Epidemico that scores social media posts with regard to the likelihood that they contain so called proto-AEs is not evaluated here.

A previously developed algorithm for named entity recognition of medication names developed by UMC in narratives in spontaneous reports was adapted to analysis of tweets and evaluated on a large data set of tweets containing 16,268 mentions of medication names.

By combining a drug dictionary with global scope for dictionary lookup with a method for word sense disambiguation, it was possible to correctly detect 81% of mentioned medicines in tweets while maintaining an acceptable precision equal to 0.89. Without disambiguation, 83% of all names could have been detected but the precision would then be unacceptable low, 0.41.

A brief analysis of causes of errors points to potential measures to improve performance: Introduction of a medication name vernacular dictionary to account for how medicines are referenced may improve recall, and the introduction of semantic analysis in word sense disambiguation has the potential to improve precision. Recall can only be improved to a lesser extent by the introduction of semantic analysis.

Because of our method for gold standard classification, the presented results should be interpreted as indicative and not definite.

Introduction

The purpose of this technical report is to describe the *current* methods that exist within the consortium for the purpose of ADR detection and evaluate its performance. The focus of WP2b is on developing methods for automatic detection of medication names, events and ADRs in social media. UMC has developed methods for named entity recognition (NER) of medication names and events in spontaneous report narratives in VigiBase. Epidemico has developed methods for analysis of complete posts (not for NER in individual phrases).

This report will focus on the evaluation of NER of medication names as developed by UMC. The methods developed by Epidemico are incorporated as key components in the processing pipeline in Web-RADR. However, their focus has not yet been on NER of individual word phrases but rather on analysis of complete posts. Therefore, this technical report will only focus on the adaptation and evaluation of the NER algorithm for medication names developed by UMC.

Named Entity Recognition of medication names

Detecting medication names in social media content is an essential computational step when performing NER of ADRs. The by far most common context in which such methods have been developed and evaluated is clinical narratives c.f. (1-8).

In an early study from 1996, Evans et al combined lexical resources, including the Unified Medical Language System (UMLS) with regular-expression-based pattern matching to extract drug names and dosages from discharge summaries of a US teaching hospital (1). They report an overall accuracy of 80%, pooling true positives and true negatives. Sirohi and Peissig (2) used drug lexicons derived from the Marshfield Clinic's drugs database, and employed a frequency dictionary of words in English language, to eliminate false positives. This increased precision from 6.9% to 54.6% with only a slight reduction in recall, from 96.4% to 95.8%.

Gold et al used RxNorm entries in the UMLS for dictionary and developed a tool for extraction of medication information from discharge summaries (3) . They reported a precision of 94.1% and a recall of 82.5%.

Xu et al proposed MedEx, a system for extraction of medication information in clinical narratives, including e.g. name, strength, route, frequency, form and duration, using a lexicon file of drug names derived from RxNorm (7). They evaluated performance of drug name extraction in 50 discharge summaries from the EMR at Vanderbilt University Medical Center, and reported precision and recall equal to 95% and 92%, respectively (7).

The 2009 i2b2 workshop on natural language processing for clinical records provided an open challenge for the identification of medications, their names and related attributes on a set of discharge summaries from Partners Healthcare in the United States (6). The best performance for the drug name recognition task was achieved with an algorithm using conditional random fields (CRF) reported to have a precision, recall and F-score equal to 93%, 88% and 90% respectively (5). Doan et al demonstrated retrospectively, that performance on the 2009 i2b2 challenge data set could be further improved by an ensemble of individual classifiers (8).

A recent study by Polepalli Ramesh et al is to our knowledge the only example where natural language processing methods have been applied for automatic identification of medication names in spontaneous reports (9). In this study, a number of machine learning methods were utilized in order to automatically annotate medication names, adverse events and related information in the Food and Drug Administration's Reporting System (FAERS). For medication name entity recognition, an F-score equal to 0.83 was reported for their best system, a tagger using CRF and a set of input features including UMLS semantic type, syntactic, semantic, morphological, affix, negation and hedging features.

We have not identified any studies where NER of medication names were developed and evaluated for a wide range of medicines. A common approach is to focus on a narrow range of medications with the objective to capture all ADRs mentioned in relation to them.

UMC has developed a Named Entity Recognition (NER) method for detecting mentions of medication names in spontaneous report narratives (Submitted for publication). The algorithm (*NER-d, spontaneous reports*) consists of three core components in a processing pipeline: 'Dictionary lookup', 'Overlap Resolution' and 'Disambiguation'.

Dictionary lookup consists of matching all words and phrases in the narratives with a dictionary consisting of all medication names extracted from WHO Drug Dictionaries, the world's most comprehensive drug dictionary. Overlapping dictionary hits were resolved into a single phrase using procedural logic. Disambiguation of identified phrases were performed using a logistic regression model predicting the likelihood that a hit constitutes an actual medication and compared with a threshold in order to perform a classification ('Medication'/'Not a Medication'). The algorithm was built on the Apache clinical Text Analysis and Knowledge Extraction System (cTAKES) framework.

NER-d, spontaneous reports, was evaluated on 333 spontaneous reports in English not previously used in the development of the algorithm. 90% of all occurrences of medication names were correctly detected (recall). The probability that a phrase classified as being a medication names actually was correct was equal to 91% (precision). As a comparison, annotations by the WHO Drug Dictionaries alone not using disambiguation yielded a recall of 94% and a precision of 19%. It was concluded that effective medication name entity recognition in an international context requires a global dictionary to ensure high recall, which in turn requires sophisticated analytics to maintain acceptable precision.

Existing data

During recent years Epidemico has collected data (tweets) from Twitter using 900 product names as search terms, in various projects. The extracted tweets were automatically annotated by an algorithm developed by Epidemico that assigns a score to all posts, indicating the likelihood that the tweet is mentioning a medication name and an adverse drug reaction ('Proto-AE'). The posts with high indicator scores were manually curated/validated and mapped to specific drugs and reactions, represented by text strings (for example "Gardasil" and "Memory impairment").

Study Objective

The objective of the present study is to *adapt* our current method for NER of medication names in spontaneous reports to social media (Twitter) and *evaluate* its performance, using a data set made available by Epidemico.

Method

Datasets

Epidemico was willing to share their dataset of tweets described above with UMC for the purpose of adapting and evaluating *NER-d, spontaneous reports* for analysis of tweets.

Tweets being obvious duplicates (exact verbatim copies and retweets) were excluded, as were posts mentioning unspecific medication names that could not be matched to a medicinal product, e.g. ‘vaccine’, ‘unknown drug’ or ‘flu shot’, since the aim of the algorithm is to identify only specific medication names. Around 70,000 posts remained after this step of exclusion.

One part of this validated and mapped dataset was used for training of the logistic regression model of the disambiguation component, and another part for evaluation, with the validated drugs used as the gold standard annotations. The following principles were applied for sampling the data set and for the selection of training and test sets:

- The dates when the tweets were posted were inspected and a cut-off date was selected that roughly divided the complete data set in two equal halves.
- Tweets that were used for training (*training data set*) were sampled randomly from the earlier time period, and tweets that were included in the *test data set* were sampled randomly from the later time period.

The rationale for this choice is that it most closely resembles a realistic situation where the data used for developing the method originates from a time before the method is used. It also significantly reduces the likelihood that duplicate postings inadvertently ends up in training and test sets, which might introduce a positive bias in the method’s classification performance when evaluated on the test data set.

Gold standard annotation

Manual annotation of medication names was previously done by Epidemico. Verbatim matches of phrases in tweets with actual product names were NOT considered medications if a phrase had another meaning (e.g. referring to the musical term “intermezzo” and not the medication “intermezzo”); non-specific medication mentioning such as “vaccination”; typos; non-comprehensible alerts; or metaphorical mentions, e.g. “She is morphine, queen of my vaccine” would not result in any valid medication gold standard annotation (Reference: epidemic, MedWatcher Social – Coding and Curation overview).

The manually annotated sets obtained by UMC from Epidemico did not contain the character locations (spans) where annotated phrase were found, just the mapped medication name.

Medication names not actively monitored but nevertheless present in posts were also annotated by Epidemico in most cases. Because of this, we have chosen to denote this manual annotation as the gold Standard annotation even though it may not have been the purpose of Epidemico to use all annotated names for a purpose like in this study. Inconsistencies in the gold standard annotation may be a result of this circumstance. Therefore, results should be regarded as indicative and not definite.

The gold standard annotation dictionary

The gold standard annotation only contains a string representation of the name of the mentioned drug in the tweet. The gold Standard annotation dictionary maps all the drugs mentioned in the dataset to a list of medicinal product ids (obtained from WHO Drug Dictionaries), a list of substance ids and a VigiBase report count, indicating how many reports that exists in VigiBase that contains the product name. All unique products (e.g. “Ibumetin”) have unique medicinal products ids. Substances (for e.g. “Ibuprofen”) also have unique medicinal product ids. All medicinal products are mapped to a substance. The list of substance ids and the VigiBase report count are derived from the medicinal product id list.

The dictionary was built automatically with a name mapping between the string representation of the drug in the gold standard annotation and the WHO Drug Dictionaries. Tradenames were mapped to all medicinal products with a matched name. Substances were mapped to single substance medicinal products where this substance was the only ingredient.

Some entries that failed automatic mapping were manually mapped.

Adaptation of *NER-d, spontaneous reports* to social media

The named entity recognition algorithm used is based on the previously developed NERd-algoritm (reference unpublished report). An adaptation of this algorithm to social media/Twitter was performed while maintaining the overall computational structure of the algorithm.

The first component of NER-d consists of lookup of all words and phrases in two dictionaries. The drug dictionary was kept intact, but the frequency dictionary was changed to reflect the context of twitter

Drug dictionary

A drug dictionary with the purpose of being used by the NER-d algorithm was created from the WHO Drug Dictionaries, and includes medicinal product names, VigiBase report counts, lists of medicinal product ids, lists of substance ids and a Boolean variable isSubstance indicating if the medication name refers to a substance name. The list of medicinal product ids contains all ids of drugs with the same name. Substances are also included, and their medicinal product id list will contain single substance medicinal products where this substance is the only ingredient. The VigiBase report count and the substance ids list are derived from the medicinal product ids list.

Frequency dictionary for English

The NER-d algorithm also makes use of a word frequency dictionary indicating the frequency with which words occur in tweets. The dictionary is based on statistics collected from 72 million tweets (http://blog.luminoso.com/2015/09/01/wordfreq-open-source-and-open-data-about-word-frequencies/#more-659. License information: https://github.com/LuminosoInsight/wordfreq). All words with higher frequency then 1 per million words were included.

Overlap resolution

The dictionary lookup may result in clusters of annotations that are partly overlapping, for example when analyzing the text “I got 3 doses of Engerix-B as well as 2 boosters”, the following drug trade names were temporarily annotated against the WHO Drug Dictionaries: “Engerix”, “Engerix B”, and “B”. The objective of the overlap resolution annotator is to analyze such clusters and select the best annotation.

Inspection of the training data set led to the development of an algorithm for resolution of overlapping temporary annotations of trade names based on 1) the number of words that matches words verbatim in the dictionary, 2) the ratio of annotated words that match the total number of words verbatim in the dictionary entry, and if those criteria failed to resolve the overlap, 3) the number of reports in VigiBase containing the annotated drug. For substances, these criteria were applied in the order 2,1 3.

Medication name disambiguation

The last component in the pipeline is a logistic regression function trained to perform word sense disambiguation, to determine if a word phrase being identified in the preceding components should be annotated as a drug mention or not. Training was performed using the gold standard annotated data sets. Four separate trainings were performed: The training data set was split in three parts that each was used to train the model separately (cross validation). A fourth model was developed by using the full data set in training. This last model was used in the evaluation.

Three selected predictors (features) were used from the original NERd-implementation to train the model; “VigiBase report count”, “Substance” and “Word frequency”.

- “VigiBase report count” is the natural logarithm of the number of reports (+1 for numerical stability) in VigiBase according to the drug dictionary.
- “Substance” is a binary variable indicating whether the annotated term is a substance or not, according to the drug dictionary.
- “Word frequency” is the natural logarithm of the counts of the annotated words (+ 1 for numeric stability) in the frequency dictionary for English. If the annotated term contains multiple words, the lowest count is used.

Determination of classification threshold

Output from the logistic regression model was compared with a decision threshold to perform classification. An optimal threshold was determined by calculating the performance for a range of different values on the training set and selecting the threshold that yielded the best performance in terms of F- score.

Evaluation methodology

The adapted NER-d algorithm was used to analyse the content in the test data set and perform NER of medication names which was compared to the gold standard classification of medication names.

The gold standard annotations in the data sets do not contain indications of on what character positions in the text the medication were found (location), only the mapped phrases that supposedly originates from another medication registry (RxNORM?).

Every annotation either performed by the NER-d algorithm or present as a gold standard (GS) annotation will be classified as either one of the following categories:

- ‘True Positive classification’ (TP): A phrase annotated by NER-d that was identified in the GS and where both refers to the same substance
- ‘True Negative classification’ (TN): A phrase in a tweet that was matched by a product in the drug dictionary, but that was correctly identified as not being a drug by the disambiguation component.
- ‘False Negative classification’ (FN): A GS annotated phrase not identified by NER-d
- ‘False positive classification’ (FP): A phrase annotated by NER-d referring to a specific substance was not present in the GS.

Classification performance was calculated in terms of recall (number of True positive classifications among all actual medication names as classified in the gold standard), precision (the probability that a word classified as a medication was correct) and F-score as the harmonic mean between recall and precision.

Result

Training of the logistic regression model in the disambiguation component

Parameters of the logistic regression model in the disambiguation component when trained on the three different subsets of the training data set is presented in Table 1. Note the low variability in the regression coefficients in the three different subsets of data.

Feature	Training data set - Complete	Training data set – Subset 1	Training data set – Subset 2	Training data set – Subset 3
VigiBase report count	0.48	0.48	0.49	0.52
Substance	-0.97	-1.04	-1.04	-1.04
Word frequency	-0.85	-0.85	-0.89	-0.82
Bias	-1.53	-1.49	-1.56	-1.90

Table 1. Regression coefficients in the logistic regression model when trained on the complete, and three subsets of the training data set

Optimal thresholds for classification based on applying the F-score criterion were 0.5, 0.5 and 0.45 for the three subsets of data, respectively. A threshold equal to 0.5 was selected for computing classification when applied to the test set.

Evaluation of classification performance

The classification performance of NER-d when applied to the three partitions of the *training* data set is presented as the three top most curves in Figure 1 and 2. Note the similarity between the three resulting models in terms of their classification performance. This indicates that the model is stable. The lower curve in Figure 1 represents the performance when evaluated on the *test set*. Table 2 presents numerically the performance of the adapted NER-d when applied to the *test set*.

Note that the performance on the test set is our best estimate on what can be expected to be the classification performance when applied to new yet unseen tweets. For all further presentation and discussion of results, we are referring to the results when evaluated on the *test set*.

Threshold	TP	TN	FP	FN	Precision	Recall	F-score
0.00	13477	0	19485	2791	0.41	0.83	0.55
0.10	13456	16353	3132	2812	0.81	0.83	0.82
0.20	13395	17443	2042	2873	0.87	0.82	0.85

0.30	13351	17620	1865	2917	0.88	0.82	0.85
0.40	13208	17717	1768	3060	0.88	0.81	0.85
0.50	**13154**	**17866**	**1619**	**3114**	**0.89**	**0.81**	**0.85**
0.60	12933	18020	1465	3335	0.90	0.80	0.84
0.70	12140	18576	909	4128	0.93	0.75	0.83
0.80	11051	18764	721	5217	0.94	0.68	0.79
0.90	8551	19008	477	7717	0.95	0.53	0.68
1.00	0	19485	0	16268	1.00	0.00	0.00

Table 2. Classification performance of NER-d adapted to social media for varying thresholds when applied to the test data set. TP and FN represent number of True Positive and False Negative classifications, respectively. TN and FP represent number of True Negative and False Positive classifications, respectively.

The classification obtained when using the threshold equal to 0.50 is indicated in the table and resulted in a recall and precision equal to 0.81 and 0.89, respectively.

Note that the recall without any disambiguation (obtained when using threshold = 0.00) is equal to 0.83. In this case, the probability that any given detected medication name by NER-d is correct is equal to 0.41. This illustrates the value of disambiguation: by accepting a slightly lower recall, disambiguation provides a significant improvement in precision so that the probability of that a detected medication name is correct is increased to 0.89.

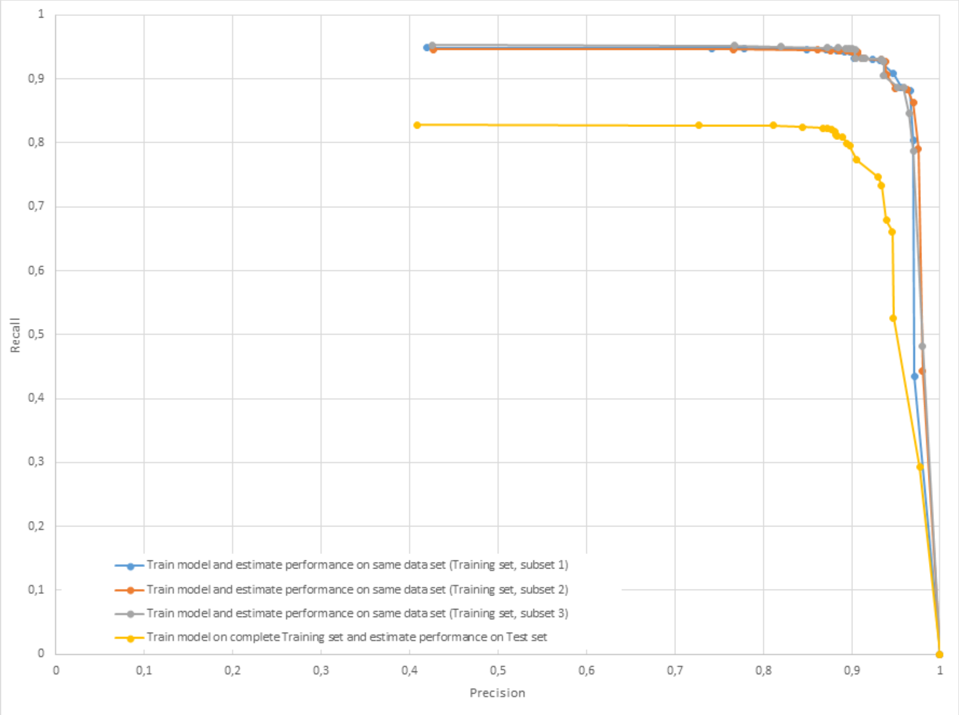

Figure 1. Recall and Precision of the logistic regression model when trained and evaluated on the same data sets (blue, orange and grey curves), and when

trained on the complete training set and being evaluated on the test set (yellow curve).

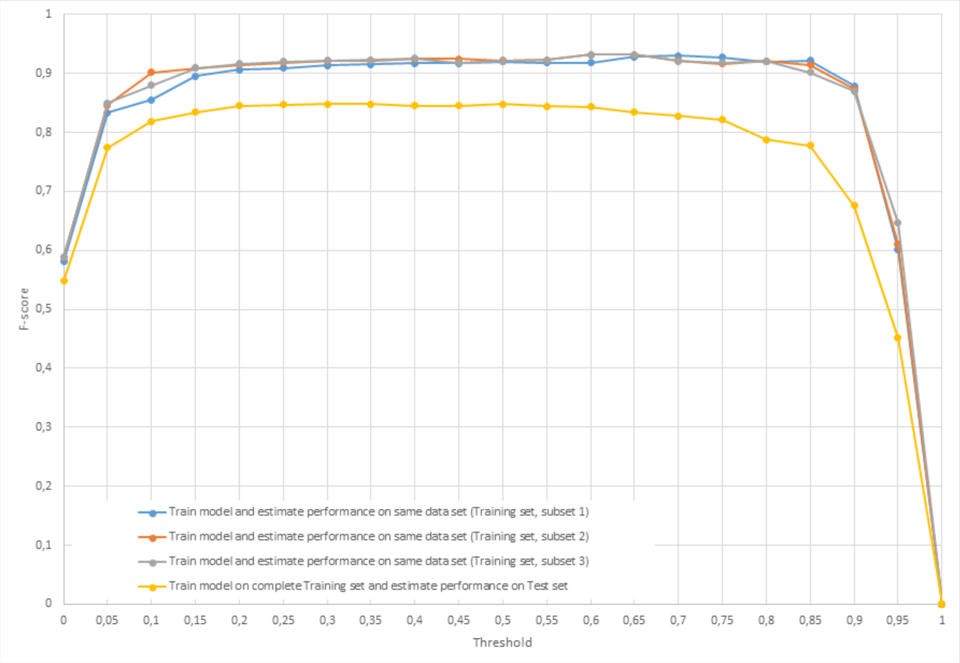

Figure 2. F-score of the logistic regression model when trained and evaluated on the same data sets (blue, orange and grey curves), and when trained on the complete training set and being evaluated on the test set (yellow curve).

Examples of tweets

A set of tweets were inspected and illustrative examples were selected and are presented below.

Correctly identified tweets (True Positives)

Table 3 presents examples of tweets with correctly annotated medication names. The total number of True Positives was 13,154.

Class	Gold standard	System	NER-d score	Other annotations in the post			
				Class	Gold standard	System	NER-d score
1) “@user i developed lupus like symptoms. so i stopped remicade, i did orencia, rituxan, xeljanz, now, mtx.”							
TP	methothrexate	mtx	0.88				
TP	orencia	orencia	0.94				
TP	remicade	remicade	0.98				
TP	rituxan	rituxan	0.94				
TP	xeljanz	xeljanz	0.89				

Table 3: Examples of true positive examples.

Correctly discarded non-medication names (True Negatives)

WHO Drug Dictionaries from which the drug dictionary was extracted, contains medication tradenames that have ambiguous meaning, such as “super”, “take”, “your” and “pain”. In most cases these will be discarded as non-medicines. Example 1 in Table 4 presents two typical examples where “today” and “headache” are discarded as medication name mentionings (“today” is also a spermicide containing nonoxinol and “headache” is a product marketed in Korea containing Flunarizine dihydrochloride).

Class	Gold standard	System	NER-d score	Other annotations in the post			
				Class	Gold standard	System	NER-d score
1) “today's oxycontin use is not recreational, btw; i am awaiting physio referral for a neck problem. it is bad today - headache, tingly arm etc..”							
TN	-	today	0. 01	TP	oxycontin	oxycontin	0.97
TN	-	headache	0.02				
2) “lama/laba combo umeclidinium/vilanterol increased fev1 more v vilanterol or tiotropium (but not umeclidinium) alone url”							
TN	-	vilanterol	0.12	FP	-	tiotropium	0.73
TN	-	umeclidinium	0.08				
TN	-	alone	0.00	FN	fluticasone furoate and vilanterol	-	-
TN	-	lama	0.06				
TN	-	combo	0.05				

Table 4: Examples of true negative examples

Example 2 in Table 4 illustrates inconsistencies with the gold standard annotation in the data set. The medication names “vilanterol” and “umeclinidium”, clearly medications, received NER-d scores equal to 0.12 and 0.08, and were classified as non-medications. However, the gold standard annotation includes the entire combination “fluticasone furoate and vilanterol”, resulting in the individual medications being classified as TN. This thus resulted in two TN:s and one FN, instead of just one FN for the whole combination. Also note that “tiotropium” was not gold standard classified as a medication name and because of this, “tiotropium” was classified as a FP instead of TP that would have been correct.

Falsely classified word phrases as medication names (False Positives)
This class of erroneously classified word phrases is exemplified by the tweets in Table 5.

Class	Gold standard	System	NER-d score	Other annotations in the post			
				Class	Gold standard	System	NER-d score
1) “@user i was on tysabri, but it gave me really bad side effects. neuro switched me to gilenya what has your neuro advised?”							
FP	-	neuro	0.64	TP	tysabri	tysabri	0.98
				TP	gilenya	gilenya	0.96
				TN	-	your	0.00
2) “ok, the halo is totally gone now, and i can feel some of the sumatriptan high/nausea kicking in. head hurts a bit, but it is not awful.”							
FP	-	nausea	0.57	TP	sumatriptan succinate	sumatriptan	0.91
				TN	-	head	0.00
				TN	-	halo	0.03

3) “ugh. #remicade is kicking my ass. hoping the big increase in #dose kicks the #behcet's ass though! #prednisone & #prednisone #eyedrops #suck”							
FP	-	prednisone	0.96	TP	remicade	remicade	0.98
				TN	-	ass	0.06
				TN	-	big	0.00
4) “back to clobetasol and prednisone. #skinasthma #sensitiveskin #highmaintenancegirl #clobetasol url”							
FP	-	clobetasol	0.71	TP	prednisone	prednisone	0.96
5) “bad luck is being up all night with the shits and finding a strip of imodium on the floor in the morning”							
FP	-	imodium	0.94	FN	imodium multi-sympton	-	-
				TN	-	with	0.00
6) “i need alcohol, cocaine and demerol. caffeine, nicotine, vicodin, heroin addicted.”							
FP	-	heroin	0.50	TP	vicodin	vicodin	0.96
				TP	demerol	demerol	0.95
FP	-	nicotine	0.92	TN	-	caffeine	0.40
				TN	-	cocaine	0.33
				TN	-	alcohol	0.29
7) “tylenol codeine works so fastt but the side effects are pure awful i am ssoo dizzy??”							
FP	-	codeine	0.73	FN	tylenol 3	-	-
FP	-	tylenol	0.95	TN	-	pure	0.01

Table 5: Examples of false positives

Example 1 and 2 in Table 5 are typical examples of false positive classifications. Example 3 is an example where the system correctly identified prednisone, but due to a possible inconsistency in the gold standard classification, this word phrase was classified as a false positive classification. Examples 4 and 6 are also of this kind.

Examples 5 and 7 illustrates GS annotations of combined products while the system finds one or two products with single substances.

Undetected medicines (False Negatives)

There were 3,114 false negatives in the test set. In this group, mentions of vaccines seemed to be particularly difficult to detect, as illustrated in example 1 in Table 6 below.

Class	Gold standard	System	Ner-d score	Other annotations in the post			
				Class	Gold standard	System	Ner-d score
1) “@user @user @user i think gatty's doc has gotten every possible vaccination out of him though. yfever, typhoid and hep a+b i got.”							
FN	hepatitis a vaccine	-	-	TP	hepatitis b vaccine	hep	0.72
FN	yellow fever vaccine	-	-	TN	-	doc	0.05
FN	typhoid vaccine	typhoid	0.42				
2) “can not think of anything grosser than a toenail fungus treatment called "jublia"”							
FN	jublia	jublia	0.18	TN	-	fungus	0.23

3) “voltarol pain-eze extra strength. on tooth ache. wow. i feel like i am flying & i have been steaming for days ??”							
FN	voltarol	voltarol pain-eze	0.18	TN	-	extra	0.01
				TN	-	days	0.00
4) “seriously some one come to mines with an imodium my arse is on fire and i can not handle any more.. #shiteyarse”							
FN	imodium multi-symptom	-	-	FP	-	imodium	0.94
				TN	-	any	0.00
				TN	-	with	0.00
				TN	-	come	0.00
				TN	-	one	0.00
5) “@user okay. i just woke up from a nap coz i took sinutab kanina. my head hurts. sipon daw, sabi ni mommy.”							
FN	sinutab with codeine	-	-	FP	-	sinutab	0.75
				TN	-	i up	0.00
				TN	-	head	0.00
				TN	-	nap	0.00
6) “took two nurofen plus, at 0900, wide awake, appointment 1430 and most people are nervous, just i can not sleep :-(not going to be good :-(“							
FN	nurofen	-	-	FP	-	nurofen plus	0.74
				TN	-	two	0.00
				TN	-	sleep	0.01

Table 6. Examples of False Negatives

The second example found the medication “jublia” in the drug dictionary, but it was erroneously discarded as being a medication due to the low indicator score produced by the disambiguation algorithm. “Voltarol pain-eze” in example 3 is another example of this. Example 4 may be an inconsistent gold standard classification since “Imodium” in the text was gold standard classified as being “Imodium multi-symptom”. The word “Imodium” was detected by the system. However, it was classified as a False Negative since “Imodium multi-symptom” contains the additional substance of Simeticone. Both products contain Loperamide hydrochloride. Example 6 is similar.

Discussion

This study demonstrated that by combining a drug dictionary with global scope for dictionary lookup with a method for word sense disambiguation, it was possible to correctly detect 81% of mentioned medicines in tweets while maintaining a high precision equal to 0.89. Without disambiguation, the gain in recall would be modest but the precision would fall to 0.41, meaning that more than half of all detected medication names would in fact be false (i.e. non medications).

Narratives in spontaneous reports in VigiBase and tweets are very different data sources. For this reason, the originally developed *NER-d, spontaneous reports*, needed to be retrained on this new data, as well as incorporating an alternative database for word frequency. However, the basic idea of using word frequency and vigiBase report count to measure the likelihood that a word phrase refers to a common word and to a medication name, respectively, was still used for disambiguation. The performance of *NER-d, spontaneous reports* was somewhat higher than for the NER-d adapted to

Twitter: The recall was 0.90 and 0.81, for *NER-d, spontaneous reports* and *NER-d (Twitter)*, respectively. The precision was about the same for *NER-d, spontaneous reports* compared to *NER-d (Twitter)*: 0.91 versus 0.89.

The type of errors was not analysed in detail. However, the set of illustrative examples is useful in order to learn how the algorithm may be improved.

False positives
The first two examples of false positive errors in Table 5 are disambiguation errors. This occurs when the NER-d score is higher than the selected threshold 0.50. Possible ways to correct this error is to introduce more sophisticated analysis such as semantic analysis of surrounding words. A method that has the potential to do this is the Conditioned Random Fields (CRF) method that have been successfully applied in some methods for medication name entity recognition.

Correcting potential errors in the gold standard classification among false positives was outside the scope of this study.

False negatives
One cause of false negative error was due to the term was not present in the drug dictionary (e.g. example 1, Table 6). This issue can be addressed by including a vernacular dictionary of medication name in addition to our drug dictionary. This might be particularly relevant in the case of analysing social media content, where medications may be referred to in a great number of slang terms or synonyms.

Another cause of error regards disambiguation, where the NER-d score is below the selected threshold of 0.50. As has been described above, one potential remedy for this error is to include semantic analysis. However, the maximum available recall is 0.83 (when all word phrases matching a medication in the drug dictionary is classified as a mediation). The potential to improve recall by semantic analysis is thus rather limited.

A third cause of error is when only part of a medication name is detected, such as “sinutab” of “sinutab kanina”. This cause of error may also be addressed by including a medication name vernacular.

A fourth cause of Classification of phrases as being false negatives is in fact in the gold standard classification. Addressing these is however not in the scope of this study.

Conclusion

By combining a drug dictionary with global scope for dictionary lookup with a method for word sense disambiguation, it was possible to correctly detect 81% of mentioned medicines in tweets while maintaining an acceptable precision equal to 0.89. Without disambiguation, 83% of all names could have been detected but the precision would then be unacceptable low, 0.41.

A brief analysis of causes of errors points to potential measures to improve performance: Introduction of a medication name vernacular dictionary to account for how medicines are referenced may improve recall, and the introduction of semantic analysis in word sense disambiguation has the potential to improve precision. Recall can only be improved to a lesser extent by the introduction of semantic analysis.

Because of our method for gold standard classification, the presented results should be interpreted as indicative and not definite.

References

1. Evans DA, Brownlow ND, Hersh WR, Campbell EM. Automating concept identification in the electronic medical record: an experiment in extracting dosage information. *Proceedings : a conference of the American Medical Informatics Association / AMIA Annual Fall Symposium*. 1996:388-92.

2. Sirohi E. PP. Study of Effect of Drug Lexicons on Medication Extraction from Electronic Medical Records. Pacific Symposium on Biocomputing; 2005. p. 308-18.

3. Gold S, Elhadad N, Zhu X, Cimino JJ, Hripcsak G. Extracting structured medication event information from discharge summaries. *AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium*. 2008:237-41.

4. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of medical informatics*. 2008:128-44.

5. Patrick J, Li M. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. *Journal of the American Medical Informatics Association : JAMIA*. 2010;17(5):524-7.

6. Uzuner O, Solti I, Cadag E. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association : JAMIA*. 2010;17(5):514-8.

7. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx: a medication information extraction system for clinical narratives. *Journal of the American Medical Informatics Association : JAMIA*. 2010;17(1):19-24.

8. Doan S, Collier N, Xu H, Pham HD, Tu MP. Recognition of medication information from discharge summaries using ensembles of classifiers. *BMC medical informatics and decision making*. 2012;12:36.

9. Polepalli Ramesh B, Belknap SM, Li Z, Frid N, West DP, Yu H. Automatically Recognizing Medication and Adverse Event Information From Food and Drug Administration's Adverse Event Reporting System Narratives. *JMIR medical informatics*. 2014;2(1):e10.