Extracting Information about Medication Use from Veterinary Discussions

Haibo Ding
School of Computing
University of Utah
Salt Lake City, UT 84112
hbding@cs.utah.edu

Ellen Riloff
School of Computing
University of Utah
Salt Lake City, UT 84112
riloff@cs.utah.edu

Abstract

Our research aims to extract information about medication use from veterinary discussion forums. We introduce the task of categorizing information about medication use to determine whether a doctor has *prescribed* medication, *changed protocols*, *observed effects*, or *stopped* use of a medication. First, we create a medication detector for informal veterinary texts and show that features derived from the Web can be very powerful. Second, we create classifiers to categorize each medication mention with respect to six categories. We demonstrate that this task benefits from a rich linguistic feature set, domain-specific semantic features produced by a weakly supervised semantic tagger, and balanced self-training.

1 Introduction

Natural language processing holds great promise for automatically extracting empirical data about medications, from the perspective of both doctors and patients. A wealth of information about the administration and effectiveness of medications lies within unstructured text, including medical records created by health care professionals (e.g., discharge summaries) as well as informal texts written by medical practitioners and patients (e.g., Web forums).

Previous work has been done on detecting medication terms and recognizing relations between medications and other medical concepts, such as diseases and symptoms. Our research explores a new problem: identifying and categorizing contexts involving the administration of medications, which we call *medication use categorization*. For each mentioned medication, we want to know whether it was used in a patient's care, and if so, what actions or observations are being reported. Our task aims to distinguish between contexts where a doctor *prescribed* a medication, *changed* the protocol of a medication (e.g., dosage or frequency), *stopped* use of a medication, *observed effects* produced by the medication, or *is asking a question* about a medication. Distinguishing these contexts is an important step toward being able to extract empirical data about medication use, such as effectiveness, success under different protocols, and adverse events.

Our research studies veterinary discussion forums, which often contain informal vocabulary such as shortened and abbreviated medication terms (e.g. *"pred"* instead of *"prednisone"*, or *"abx"* for *"antibiotics"*). The first part of our research addresses the problem of medication detection from informal text. We create an effective medication detector using supervised learning with linguistic features as well as contextual features acquired from the Web. We show that the Web context features substantially improve recall, and yield an effective medication detector even with small amounts of training data.

Second, we design supervised classifiers for medication use categorization. We incorporate a rich set of contextual, syntactic, and sentential features as well as a semantic tagger trained for the veterinary domain with bootstrapped learning over a large set of unannotated veterinary texts. We demonstrate additional performance gains by using balanced self-training with the unannotated texts.

2 Related Work

Previous work on extracting medication information from text has primarily focused on clinical medical text, such as discharge summaries (e.g., (Doan and Xu, 2010; Halgrim et al., 2010; Doan et al., 2012; Tang et al., 2013; Segura-Bedmar et al.,

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 1452–1458,
Denver, Colorado, May 31 – June 5, 2015.

2013)). The Third and Fourth i2b2 Shared Tasks included medication detection from clinical texts (Uzuner et al., 2010; Uzuner et al., 2011), and the Fourth i2b2 Shared Task also included relation classification between treatments (including medications), problems, and tests. Recently, there has been growing interest in extracting medication information from other types of text, such as Twitter, online health forums, and drug review sites (e.g., (Leaman et al., 2010; Bian et al., 2012; Liu et al., 2013; Liu and Chen, 2013; Yates and Goharian, 2013; Segura-Bedmar et al., 2014)). Much of this research is geared toward identifying adverse drug events or drug-drug interactions.

Many methods have been used for medication extraction, including rule based approaches (Levin et al., 2007; Xu et al., 2010), machine learning (Patrick and Li, 2010; Doan and Xu, 2010; Tang et al., 2013), and hybrid methods (Halgrim et al., 2010; Meystre et al., 2010). Rule based and hybrid approaches typically rely on manually created lexicons and rules. RxNorm (Nelson et al., 2011; Liu et al., 2005) is a large knowledge base containing generic and brand names of drugs and it is often used as a component in these systems. We compare our results with the MedEx system (Xu et al., 2010), which uses RxNorm coupled with manually defined rules.

To our knowledge, classifying medication mentions with respect to administration use categories has not yet been studied. A novel aspect of our work is also the use of Web Context features for medication detection. Similar Web features have been exploited for fine-grained person classification (Giuliano, 2009), while we demonstrate that they can be highly beneficial for medical concept extraction.

3 Task Description and Data Set

We divide our task into two subproblems: (1) *Medication Detection* aims to identify words corresponding to non-food substances used to treat patients (e.g., drugs, potassium supplements), and (2) *Medication Use Categorization* aims to classify medication mentions based on actions and observations related to their administration and to identify question contexts. We assign each medication mention to one of the six categories below.

Rx: The text indicates that a doctor prescribed the medication for a patient, or that a patient is taking (or has previously taken) the medication.
Example: "*I started the dog on abx.*"
Change: A change in the administration of the medication was made (e.g., dosage, route, frequency).
Example: "*I increased the pred to 5mg.*"
Stop: Use of the medication was discontinued.
Example: "*We took the cat off metacam.*"
Effect: The text reports a positive or negative effect from the medication on a patient.
Example: "*The dog responded well to Vetsulin.*"
Question: A question is asked about the medication.
Example: "*Do you think we should consider lasix?*"
Other: None of the above. This category primarily covers contexts not describing patient use.
Example: "*Aranesp is expensive.*"

Our data consists of discussion forums from the Veterinary Information Network (VIN), which is a web portal (www.vin.com) that hosts message boards for veterinary professionals to discuss cases and issues in their practice. To produce gold standard annotations, we collected the initial post of 500 randomly selected threads from VIN forums about cardiology/pulmonology, endocrinology, and feline internal medicine. We defined annotation guidelines to identify the minimum span of medication mentions.[1] Two people independently annotated 50 texts, and we measured their inter-annotator agreement (IAA) using Cohen's kappa (κ) statistic. For medication detection, their IAA score was κ = .96.

For the medication use categories, we measured IAA in two ways. First, we measured agreement on all of the words labeled as a medication by at least one annotator, yielding $\kappa = 0.80$. Second, we measured agreement only on the words labeled as a medication by both annotators (to more directly assess agreement on the six categories), yielding $\kappa = .92$. Finally, the annotators labeled an additional 450 texts, producing a gold standard set of 500 labeled texts. Of the annotated medication mentions, 93% have one word and 6% have two words. The frequency of each category is shown below.

Rx	Question	Effect	Change	Stop	Other
908	289	181	52	53	470

[1] Dosage and duration terms were not included.

4 Medication Detection

Detecting medication terms in discussion forums is challenging because of their informal nature. As we will show in Section 4.1, dictionary look-up from lexicons is not sufficient. Therefore the first part of our research aims to create an effective medication detector for these informal veterinary texts. We used the Stanford CoreNLP tools (Manning et al., 2014) for lemmatization, POS tagging and parsing, and created a SVM classifier with a linear kernel using SVMlin (Sindhwani and Keerthi, 2006). The classifier labels each token as a medication term or not a medication term. Adjacent medication tokens are then combined into a single medication mention. We designed three types of features:

Word Features include the medication word's string, lemma, and part-of-speech tag. Since drugs often have common affixes (e.g., *"-sone"* is a common suffix for corticosteroids), we also defined features for character sequences of length 2-4 at the beginning and end of a word.

Local Context Features represent the word preceding and the word following each medication term. We replace numbers with the symbol "CD". We also defined features to represent the syntactic dependency relations linked to the medication word using the Stanford Parser (De Marneffe et al., 2006).

Web Context Features capture information from web sites that mention a term, which provides external context beyond the information available in the training texts. During training, we issued a Google query for each unique word in our training data and collected the title and text snippets of the top 10 retrieved documents. We then defined binary-valued features to represent all of the words in the retrieved texts.[2] We store the results of each query so that additional queries are needed only for previously unseen words.

4.1 Medication Detection Results

We conducted 10-fold cross-validation experiments on our data set to evaluate our medication detector.

First, we created three baselines to assess the difficulty of medication detection for this data. The first row of Table 1 shows the performance of a veterinary thesaurus manually created by the VIN.[3] We extracted all of the words in the entries categorized as *Pharmacologic Substance* and label all instances of those words as medication terms. The VIN thesaurus achieved high precision but only 51% recall. Some reasons for the low coverage include abbreviations, misspellings, general terms (e.g., *"drug"*), and pronouns that refer to medications (which are annotated in our data). The second row shows the results of MedEx (Xu et al., 2010), which uses the RxNorm drug lexicon and ranked in second place for the 2009 i2b2 Medication Extraction challenge. MedEx's low precision is primarily due to labeling chemical substances (e.g., *"glucose"*) as medications, but in our data they are often test results (e.g., *"the cat's glucose level..."*). The third row shows the results of creating a Training Lexicon by collecting all nouns annotated as medications in the training data. We labeled all instances of these nouns as medication terms in the test data, which produced slightly higher recall and precision than MedEx.

Method	Precision	Recall	F
VIN thesaurus	**90.9**	51.3	65.6
MedEX	52.5	73.8	61.4
Training Lexicon	59.4	76.9	67.0
SVM Classifier			
Word Features	88.2	79.9	83.9
+ Local Context	89.7	81.2	85.3
+ Web Context	89.2	**86.1**	**87.6**

Table 1: Medication Detection Results

The last three rows in Table 1 show the results for our classifier. With only Word Features, the classifier produced an 83.9% F score, outperforming the baselines. Adding the Local Context Features yielded small gains in recall and precision. The Web Context Features further increased recall from 81% to 86%, raising the F score to 87.6%. We tried adding features for the VIN thesaurus and MedEx system, but they did not improve upon the results obtained with the Web Context features.

We observed that the Web Context Features can compensate for small amounts of training data. To demonstrate how powerful they are, we randomly selected 100 gold standard texts to use as a test

[2] We also tried different context windows but found that using the title and entire snippet achieved the best results.

[3] We used a version provided to us in 2013.

set, and trained classifiers using different amounts of training data. Figure 1 shows the results for classifiers using only the Word Features, Word and Local Context Features, and all features. The classifier with Web Context Features achieved an F score $> 70\%$ using only 10 training texts, and approached its best performance with just 100 training texts.

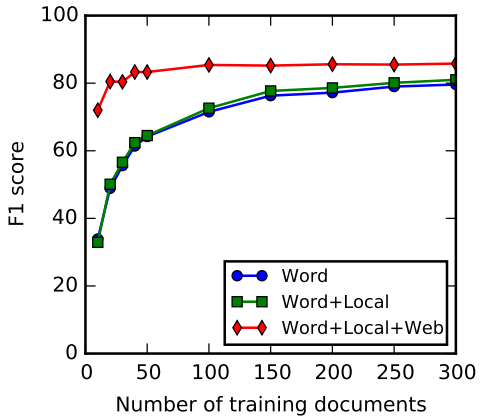

Figure 1: Learning curves with different feature sets

5 Medication Use Categorization

We tackled this problem by designing a supervised classifier with linguistic features, and incorporated a semantic tagger trained by bootstrapping on a large collection of veterinary text. We also used a balanced self-training method on unannotated veterinary texts to further improve performance.

First, we created a one-vs-the-rest binary SVM classifier for each category using scikit-learn (Pedregosa et al., 2011).[4] If an instance is labeled with multiple categories, we select the most confident one using the distance from the hyperplane. We designed three sets of features. **N-gram Features** represent a context window of size eight (+/-4) around the medication mention. We define features for lexical unigrams, lexical bigrams, lemma unigrams, and lemma bigrams. **Syntactic Features** capture verb phrases that participate in a dependency relation with the medication, using the Stanford parser. The third set of **Sentential Features** are for the *Question* and *Other* categories to recognize sentences that do not describe use of the medication on a patient, but ask questions, request guidance, describe hypothetical scenarios, etc. The sentential features consist of clause initial part-of-speech (POS) and lemma bigrams; whether the sentence ends with a question mark; whether the word "question" occurs in the same NP as the medication; whether the sentence contains the POS sequence <MD PRP>[5]; and whether the medication is separated by a comma from the ending question mark (for lists).

Semantic Tagging. We hypothesized that identifying semantic concepts might be beneficial. For example, the presence of an ANIMAL term suggests a patient, and a SYMPTOM term may indicate the reason for a prescription or an effect of medication use. First, we used WordNet (Miller, 1995) and identified synsets for 4 semantic classes: ANIMAL, DRUG, DISEASE/SYMPTOM, and HUMAN. We assigned any noun phrase with a head in these synsets to the corresponding semantic type. Next, we used a bootstrapping method (Huang and Riloff, 2010) to build domain-specific semantic taggers (**SemTaggers**) for the same four semantic classes as well as TEST, TREATMENT and OTHER. We used 10 seed words[6] for each category and 10,000 unlabeled veterinary forum texts for bootstrapping. Finally, we created **Semantic Features** for our medication use classifier. Each noun phrase tagged with a semantic class was replaced by a semantic type. Then we constructed features from pairs of adjacent terms in a context window of size eight (+/-4) around each medication mention. For example, the word sequence *"for a Boston terrier with diabetes"* would be transformed into *"for* ANIMAL *with* DISSYM*"* and the features for this context would be: <for ANIMAL>, <ANIMAL with>, and <with DISSYM>.

5.1 Medication Use Categorization Results

Table 2 shows the results for medication use classification, applied to the mentions identified by our medication detector (from Section 4). The N-gram

[4]Note that for medication detection we used SVMlin, but we switched to scikit-learn for the medication categorization because it supported additional types of classifiers that we wanted to try. Ultimately, however, the SVM performed best. We confirmed that SVM results from both toolkits were very similar.

[5]For question phrases such as "would he".

[6]We used the same seeds as (Huang and Riloff, 2010). However we added one semantic class, TREATMENT, so for this category we manually identified the 10 most frequent words in the unannotated texts that describe treatments.

Method	Rx			Question			Effect			Change			Stop			Other			Average		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
N-grams	69	74	71	75	65	69	69	37	48	76	44	56	45	23	31	50	54	52	64.0	49.6	55.9
+Sentential	68	73	71	79	71	**75**	74	41	53	85	45	59	49	32	38	51	54	53	67.8	52.7	59.3
+Syntactic	69	72	71	78	70	74	70	40	51	77	47	59	70	49	58	51	56	53	69.3	55.7	61.8
All+WordNet	69	73	71	80	70	74	73	43	54	86	49	63	72	39	54	50	54	52	71.7	54.6	62.0
All+SemTaggers	69	74	71	80	70	**75**	78	42	55	87	51	64	73	51	60	53	55	54	**73.2**	57.2	64.2
w/Balanced Self-Training	71	73	**72**	81	69	**75**	69	49	**57**	76	64	**70**	67	69	**68**	55	56	**56**	70.0	**63.5**	**66.6**

Table 2: Medication Use Categorization Results on detected medications (each cell shows Precision, Recall, F)

Method	Rx			Question			Effect			Change			Stop			Other			Average		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
N-grams	75	85	**80**	84	82	83	70	40	51	77	44	56	46	24	32	62	65	63	69.1	56.6	62.3
+Sentential	75	83	79	89	88	**89**	70	45	55	86	45	59	52	38	44	61	64	63	72.4	60.4	65.9
+Syntactic	76	82	79	88	87	88	70	44	54	77	57	59	80	55	65	61	65	63	75.4	63.4	68.9
All+WordNet	75	83	79	89	86	88	75	46	57	81	54	65	82	45	58	60	64	62	77.2	63.1	69.4
All+SemTaggers	76	85	**80**	90	87	88	79	47	59	89	56	68	81	57	67	65	65	**65**	**79.8**	65.9	72.2
w/Balanced Self-Training	78	80	79	90	86	88	75	54	**63**	76	71	**74**	78	80	**79**	66	65	**65**	77.2	**73.2**	**75.1**

Table 3: Medication Use Categorization Results on gold medications (each cell shows Precision, Recall, and F)

features alone yield an average F score of 55.9%. Both the Sentential features and Syntactic features (added cumulatively) further improve performance, raising the average F score to 61.8%. The next two rows show the effect of adding the semantic features. WordNet improves performance for **Effect** and **Change** but recall is lower for **Stop** and **Other**. In contrast, the SemTaggers improve performance across all categories, raising the F score to 64.2%. Our ablation studies show the ANIMAL class contributed most to the improvement.

In addition, we explored self-training to exploit unannotated texts. We applied the classifiers to 2,000 unlabeled veterinary texts, and used the newly labeled instances as additional training data. This did not improve performance, presumably because the most common categories dominated the new instances. We then explored a balanced self-training method that enforces an even distribution of the six categories in the new training instances. For this approach, we added exactly k new instances[7] for each class, where k was selected to be the size of the smallest set of newly labeled instances among the six categories. The last row of Table 2 shows that this balanced self-training approach improved the average F score from 64.2% to 66.6%.

[7]The most confident new instances were selected based on the differences between the scores for the winning class and the other classes.

Table 3 shows the results for medication use classification applied to gold standard medication mentions. The same trends hold: the *sentential* and *syntactic* features improve over n-grams, the *SemTagger* semantic features add value and outperform WordNet, and balanced self-training further improves performance. Overall performance increases from 66.6% to 75.1% F score with gold medications.

6 Conclusion

This research introduced a new task for classifying medication mentions with respect to their use in patient care. We created an effective medication detector for informal veterinary texts that exploits features derived from Web pages, and we created classifiers to recognize six medication use categories. These classifiers achieved precision $\geq$ 75% for all categories except Other, with recall ranging from 54% for Effects to 86% for Questions. This research is a first step toward NLP systems that can acquire empirical data about the administration and effectiveness of medications from unstructured text.

Acknowledgments

This material is based upon work supported by the National Science Foundation under grant IIS-1018314. We are very grateful to the Veterinary Information Network for providing samples of their data, and Ashequl Qadir for help annotating the data.

References

Jiang Bian, Umit Topaloglu, and Fan Yu. 2012. Towards large-scale twitter mining for drug-related adverse events. In *Proceedings of the 2012 international workshop on Smart health and wellbeing*, pages 25–32. ACM.

Marie-Catherine De Marneffe, Bill MacCartney, Christopher D Manning, et al. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*.

Son Doan and Hua Xu. 2010. Recognizing medication related entities in hospital discharge summaries using support vector machine. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 259–266. Association for Computational Linguistics.

Son Doan, Nigel Collier, Hua Xu, Pham H Duy, and Tu M Phuong. 2012. Recognition of medication information from discharge summaries using ensembles of classifiers. *BMC medical informatics and decision making*, 12(1):36.

Claudio Giuliano. 2009. Fine-grained classification of named entities exploiting latent semantic kernels. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 201–209. Association for Computational Linguistics.

Scott Halgrim, Fei Xia, Imre Solti, Eithon Cadag, and Özlem Uzuner. 2010. Extracting medication information from discharge summaries. In *Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents*, pages 61–67. Association for Computational Linguistics.

Ruihong Huang and Ellen Riloff. 2010. Inducing domain-specific semantic class taggers from (almost) nothing. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 275–285. Association for Computational Linguistics.

Robert Leaman, Laura Wojtulewicz, Ryan Sullivan, Annie Skariah, Jian Yang, and Graciela Gonzalez. 2010. Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks. In *Proceedings of the 2010 workshop on biomedical natural language processing*, pages 117–125. Association for Computational Linguistics.

Matthew A Levin, Marina Krol, Ankur M Doshi, and David L Reich. 2007. Extraction and mapping of drug names from free text to a standardized nomenclature. In *AMIA Annual Symposium Proceedings*, volume 2007, page 438. American Medical Informatics Association.

Xiao Liu and Hsinchun Chen. 2013. Azdrugminer: an information extraction system for mining patient-reported adverse drug events in online patient forums. In *Smart Health*, pages 134–150. Springer.

Simon Liu, Wei Ma, Robin Moore, Vikraman Ganesan, and Stuart Nelson. 2005. Rxnorm: prescription for electronic drug information exchange. *IT professional*, 7(5):17–23.

Mei Liu, Ruichu Cai, Yong Hu, Michael E Matheny, Jingchun Sun, Jun Hu, and Hua Xu. 2013. Determining molecular predictors of adverse drug reactions with causality analysis based on structure learning. *Journal of the American Medical Informatics Association*, pages amiajnl–2013.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Stéphane M Meystre, Julien Thibault, Shuying Shen, John F Hurdle, and Brett R South. 2010. Textractor: a hybrid system for medications and reason for their prescription extraction from clinical text documents. *Journal of the American Medical Informatics Association*, 17(5):559–562.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Stuart J Nelson, Kelly Zeng, John Kilbourne, Tammy Powell, and Robin Moore. 2011. Normalized names for clinical drugs: Rxnorm at 6 years. *Journal of the American Medical Informatics Association*, 18(4):441–448.

Jon Patrick and Min Li. 2010. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. *Journal of the American Medical Informatics Association*, 17(5):524–527.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Isabel Segura-Bedmar, Paloma Martınez, and Marıa Herrero-Zazo. 2013. Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). *Proceedings of Semeval*, pages 341–350.

Isabel Segura-Bedmar, Santiago de la Pena, and Paloma Martınez. 2014. Extracting drug indications and adverse drug reactions from spanish health social media. *ACL 2014*, page 98.

Vikas Sindhwani and S Sathiya Keerthi. 2006. Large scale semi-supervised linear svms. In *Proceedings of*

the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pages 477–484. ACM.

Buzhou Tang, Hongxin Cao, Yonghui Wu, Min Jiang, and Hua Xu. 2013. Recognizing clinical entities in hospital discharge summaries using structural support vector machines with word representation features. *BMC medical informatics and decision making*, 13(Suppl 1):S1.

Özlem Uzuner, Imre Solti, and Eithon Cadag. 2010. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5):514–518.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*.

Hua Xu, Shane P Stenner, Son Doan, Kevin B Johnson, Lemuel R Waitman, and Joshua C Denny. 2010. Medex: a medication information extraction system for clinical narratives. *Journal of the American Medical Informatics Association*, 17(1):19–24.

Andrew Yates and Nazli Goharian. 2013. Adrtrace: detecting expected and unexpected adverse drug reactions from user reviews on social media sites. *Advances in Information Retrieval*, pages 816–819.