Classifying Message Board Posts with an Extracted Lexicon of Patient Attributes

Ruihong Huang and Ellen Riloff
School of Computing
University of Utah
Salt Lake City, UT 84112
{huangrh, riloff}@cs.utah.edu

Abstract

The goal of our research is to distinguish veterinary message board posts that describe a case involving a specific patient from posts that ask a general question. We create a text classifier that incorporates automatically generated attribute lists for veterinary patients to tackle this problem. Using a small amount of annotated data, we train an information extraction (IE) system to identify veterinary patient attributes. We then apply the IE system to a large collection of unannotated texts to produce a lexicon of veterinary patient attribute terms. Our experimental results show that using the learned attribute lists to encode patient information in the text classifier yields improved performance on this task.

1 Introduction

Our research focuses on the problem of classifying message board posts in the domain of veterinary medicine. Most of the posts in our corpus discuss a case involving a specific patient, which we will call *patient-specific* posts. But there are also posts that ask a general question, for example to seek advice about different medications, information about new procedures, or how to perform a test. Our goal is to distinguish the patient-specific posts from general posts so that they can be automatically routed to different message board folders.

Distinguishing patient-specific posts from general posts is a challenging problem for two reasons. First, virtually any medical topic can appear in either type of post, so the vocabulary is very similar. Second, a highly skewed distribution exists between patient-specific posts and general posts. Almost 90% of the posts in our data are about specific patients.

With such a highly skewed distribution, it would seem logical to focus on recognizing instances of the minority class. But the distinguishing characteristic of a general post is the *absence* of a patient. Two nearly identical posts belong in different categories if one mentions a patient and the other does not. Consequently, our aim is to create features that identify references to a specific patient and use these to more accurately distinguish the two types of posts.

Our research explores the use of information extraction (IE) techniques to automatically identify common attributes of veterinary patients, which we use to encode patient information in a text classifier. Our approach involves three phases. First, we train a conditional random fields (CRF) tagger to identify seven common types of attributes that are often ascribed to veterinary patients: SPECIES/BREED, NAME, AGE, GENDER, WEIGHT, POSSESSOR, and DISEASE/SYMPTOM. Second, we apply the CRF tagger to a large set of unannotated message board posts, collect its extractions, and harvest the most frequently extracted terms to create a *Veterinary Patient Attribute (VPA) Lexicon*.

Finally, we define three types of features that exploit the harvested VPA lexicon. These features represent the patient attribute terms, types, and combinations of them to help the classifier determine whether a post is discussing a specific patient. We conduct experiments which show that the extracted patient attribute information improves text classification performance on this task.

2 Related Work

Our work demonstrates the use of information extraction techniques to benefit a text classification application. There has been a great deal of research on text classification (e.g., (Borko and Bernick, 1963; Hoyle, 1973; Joachims, 1998; Nigam et al., 2000; Sebastiani, 2002)), which most commonly has used bag-of-word features. Researchers have also investigated clustering (Baker and McCallum, 1998), Latent Semantic Indexing (LSI) (Zelikovitz and Hirsh, 2001), Latent Dirichlet Allocation (LDA) (Br et al., 2008) and string kernels (Lodhi et al., 2001). Information extraction techniques have been used previously to create richer features for event-based text classification (Riloff and Lehnert, 1994) and web page classification (Furnkranz et al., 1998). Semantic information has also been incorporated for text classification. However, most previous work relies on existing semantic resources, such as Wordnet (Scott and Stan, 1998; Bloehdorn and Hotho, 2006) or Wikipedia (Wang et al., 2009).

There is also a rich history of automatic lexicon induction from text corpora (e.g., (Roark and Charniak, 1998; Riloff and Jones, 1999; McIntosh and Curran, 2009)), Wikipedia (e.g., (Vyas and Pantel, 2009)), and the Web (e.g., (Etzioni et al., 2005; Kozareva et al., 2008; Carlson et al., 2010)). The novel aspects of our work are in using an IE tagger to harvest a domain-specific lexicon from unannotated texts, and using the induced lexicon to encode domain-specific features for text classification.

3 Text Classification with Extracted Patient Attributes

This resesarch studies message board posts from the Veterinary Information Network (VIN), which is a web site (www.vin.com) for professionals in veterinary medicine. VIN hosts forums where veterinarians discuss medical issues, challenging cases, etc. We observed that patient-specific veterinary posts almost always include some basic facts about the patient, such as the animal's breed, age, or gender. It is also common to mention the patient's owner (e.g., *"a new client's cat"*) or a disease or symptom that the patient has (e.g., *"a diabetic cat"*). General posts almost never contain this information.

Although some of these terms can be found in existing resources such as Wordnet (Miller, 1990), our veterinary message board posts are filled with informal and unconventional vocabulary. For example, one might naively assume that *"male"* and *"female"* are sufficient to identify gender. But the gender of animals is often revealed by describing their spayed/neutered status, often indicated with shorthand notations. For example, *"m/n"* means male and neutered, *"fs"* means female spayed, *"castrated"* means neutered and implies male. Shorthand terms and informal jargon are also frequently used for breeds (e.g., *"doxy"* for dachsund, *"labx"* for labrador cross, *"gshep"* for German Shepherd) and ages (e.g., *"3-yr-old"*, *"3yo"*, *"3mo"*). A particularly creative age expression describes an animal as (say) *"a 1999 model"* (i.e., born in 1999). To recognize the idiosyncratic vocabulary in these texts, we use information extraction techniques to identify terms corresponding to seven attributes of veterinary patients: SPECIES/BREED, NAME, AGE, WEIGHT, GENDER, POSSESSOR, and DISEASE/SYMPTOM.

Figure 1 illustrates our overall approach, which consists of three steps. First, we train a sequential IE tagger to label veterinary patient attributes using supervised learning. Second, we apply the tagger to 10,000 unannotated message board posts to automatically create a Veterinary Patient Attribute (VPA) Lexicon. Third, we use the VPA Lexicon to encode patient attribute features in a document classifier.

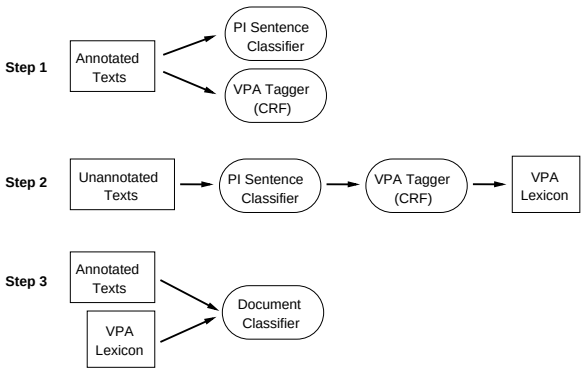

Figure 1: Flowchart for Creating a Patient-Specific vs. General Document Classifier

3.1 Patient Attribute Tagger

The first component of our system is a tagger that labels veterinary patient attributes. To train the tagger, we need texts labeled with patient attributes.

The message board posts can be long and tedious to read (i.e., they are often filled with medical history and test results), so manually annotating every word would be arduous. However, the patient is usually described at the beginning of a post, most commonly in 1-2 "introductory" sentences. Therefore we adopted a two stage process, both for manual and automatic tagging of patient attributes.

First, we created annotation guidelines to identify "patient introductory" (PI) sentences, which we defined as sentences that introduce a patient to the reader by providing a general (non-medical) description of the animal (e.g., *"I was presented with a m/n Siamese cat that is lethargic."*) We randomly selected 300 posts from our text collection and asked two human annotators to manually identify the PI sentences. We measured their inter-annotator agreement using Cohen's kappa (κ) and their agreement was κ=.93. The two annotators then adjudicated their differences to create our gold standard set of PI sentence annotations. 269 of the 300 posts contained at least one PI sentence , indicating that 89.7% of the posts mention a specific patient. The remaining 31 posts (10.3%) are general in nature.

Second, the annotators manually labeled the words in these PI sentences with respect to the 7 veterinary patient attributes. On 50 randomly selected texts, the annotators achieved an inter-annotator agreement of κ = .89. The remaining 250 posts were then annotated with patient attributes (in the PI sentences), providing us with gold standard attribute annotations for all 300 posts. To illustrate, the sentence below would have the following labels:

$$\text{Daisy}_{name} \text{ is a } \text{10yr}_{age} \text{ old}_{age} \text{ lab}_{species}$$

We used these 300 annotated posts to train both a PI sentence classifier and a patient attribute tagger. The PI sentence classifier is a support vector machine (SVM) with a linear kernel (Keerthi and DeCoste, 2005), unigram and bigram features, and binary feature values. The PI sentences are the positive training instances, and the sentences in the general posts are negative training instances.

For the tagger, we trained a single conditional random fields (CRF) model to label all 7 types of patient attributes using the CRF++ package (Lafferty et al., 2001). We defined features for the word string and the part-of-speech tags of the targeted word, two words on its left, and two words on its right.

Given new texts to process, we first apply the PI sentence classifier to identify sentences that introduce a patient. These sentences are given to the patient attribute tagger, which labels the words in those sentences for the 7 patient attribute categories.

To evaluate the performance of the patient attribute tagger, we randomly sampled 200 of the 300 annotated documents to use as training data and used the remaining 100 documents for testing. For this experiment, we only applied the CRF tagger to the gold standard PI sentences, to eliminate any confounding factors from the PI sentence classifier. Table 1 shows the performance of the CRF tagger in terms of Recall (%), Precision (%), and F Score (%). Its precision is consistently high, averaging 91% across all seven attributes. But the average recall is only 47%, with only one attribute (AGE) achieving recall $\geq$ 80%. Nevertheless, the CRF's high precision justifies our plan to use the CRF tagger to harvest additional attribute terms from a large collection of unannotated texts. As we will see in Section 4, the additional terms harvested from the unannotated texts provide substantially more attribute information for the document classifier to use.

Attribute	**Rec**	**Prec**	**F**
SPECIES/BREED	59	93	72
NAME	62	100	76
POSSESSOR	12	100	21
AGE	80	91	85
GENDER	59	81	68
WEIGHT	19	100	32
DISEASE/SYMPTOM	35	73	47
Average	47	91	62

Table 1: Patient Attribute Tagger Evaluation

3.2 Creating a Veterinary Patient Attribute (VPA) Lexicon

The patient attribute tagger was trained with supervised learning, so its ability to recognize important words is limited by the scope of its training set. Since we had an additional 10,000 unannotated veterinary message board posts, we used the tagger to acquire a large lexicon of patient attribute terms.

We applied the PI sentence classifier to all 10,000 texts and then applied the patient attribute tagger to each PI sentence. The patient attribute tagger is not

perfect, so we assumed that words tagged with the same attribute value at least five times[1] are most likely to be correct and harvested them to create a veterinary patient attribute (VPA) lexicon. This produced a VPA lexicon of 592 words. Table 2 shows examples of learned terms for each attribute, with the total number of learned words in parentheses.

Species/Breed (177): DSH, Schnauzer, kitty, Bengal, pug, Labrador, siamese, Shep, miniature, golden, lab, Spaniel, Westie, springer, Chow, cat, Beagle, Mix, ...
Name (53): Lucky, Shadow, Toby, Ginger, Boo, Max, Baby, Buddy, Tucker, Gracie, Maggie, Willie, Tiger, Sasha, Rusty, Beau, Kiki, Oscar, Harley, Scooter, ...
Age (59): #-year, adult, young, YO, y/o, model, wk, y.o., yr-old, yrs, y, #-yr, #-month, #m, mo, mth, ...
Gender (39): F/s, speyed, neutered, spayed, N/M, FN, CM, F, mc, mn, SF, male, fs, M/N, Female, S, S/F, m/n, m/c, intact, M, NM, castrated, ...
Weight (5): lb, lbs, pound, pounds, kg
Possessor (7): my, owner, client, technician, ...
Disease/Symptom (252): abscess, fever, edema, hepatic, inappetance, sneezing, blindness, pain, persistent, mass, insufficiency, acute, poor, ...

Table 2: Examples from the Induced VPA Lexicon

3.3 Text Classification with Patient Attributes

Our ultimate goal is to incorporate patient attribute information into a text classifier to help it distinguish between patient-specific posts and general posts. We designed three sets of features:

Attribute Types: We create one feature for each attribute type, indicating whether a word of that attribute type appeared or not.

Attribute Types with Neighbor: For each word labeled as a patient attribute, we create two features by pairing its Attribute Type with a preceding or following word. For example, given the sentence: *"The tiny Siamese kitten was lethargic."*, if "Siamese" has attribute type SPECIES then we create two features: <*tiny*, SPECIES> and <SPECIES, *kitten*>.

Attribute Pairs: We create features for all pairs of patient attribute words that occur in the same sentence. For each pair, we create one feature representing the words themselves and one feature representing the attribute types of the words.

[1] After our text classification experiments were done, we reran the experiments with the unigrams+lexicon classifier using thresholds ranging from 1 to 10 for lexicon creation, just to see how much difference this threshold made. We found that values ≥ 5 produced nearly identical classification results.

4 Evaluation

To create a blind test set for evaluation, our annotators labeled an additional 500 posts as *patient-specific* or *general*. Specifically, they labeled those 500 posts with PI sentences. The absence of a PI sentence meant that the post was general. Of the 500 texts, 48 (9.6%) were labeled as general posts. We evaluated the performance of the PI sentence classifier on this test set and found that it achieved 88% accuracy at identifying patient introductory sentences.

We then conducted a series of experiments for the document classification task: distinguishing patient-specific message board posts from general posts. All of our experiments used support vector machine (SVM) classifiers with a linear kernel, and ran 10-fold cross validation on our blind test set of 500 posts. We report Recall (%), Precision (%), and F score (%) results for the patient-specific posts and general posts separately, and for the macro-averaged score across both classes. For the sake of completeness, we also show overall Accuracy (%) results. However, we will focus attention on the results for the general posts, since our main goal is to improve performance at recognizing this minority class.

As a baseline, we created SVM classifiers using unigram features.[2] We tried binary, frequency, and tf-idf feature values. The first three rows of Table 3 show that binary feature values performed the best, yielding a macro-averaged F score of 81% but identifying only 54% of the general posts.

The middle section of Table 3 shows the performance of SVM classifiers using our patient attribute features. We conducted three experiments: applying the CRF tagger to PI sentences (per its design), and labeling words with the VPA lexicon either on all sentences or only on PI sentences (as identified by the PI sentence classifier). The CRF features produced extremely low recall and precision on the general posts. The VPA lexicon performed best when applied only to PI sentences and produced much higher recall than all of the other classifiers, although with lower precision than the two

[2] We also tried unigrams + bigrams, but they did not perform better.

Method	Patient-Specific Posts			General Posts			Macro Avg			
	Rec	Prec	F	Rec	Prec	F	Rec	Prec	F	Acc
Unigram Features										
Unigrams (freq)	96	96	96	58	60	59	77	76	77	92
Unigrams (tf-idf)	99	93	96	33	84	48	66	89	76	93
Unigrams (binary)	98	95	97	54	79	64	76	87	81	94
Patient Attribute Features										
CRF Features (PI Sents)	99	91	95	02	25	04	51	58	54	90
VPA Lexicon Features (All Sents)	96	96	96	60	63	62	78	79	79	93
VPA Lexicon Features (PI Sents)	96	98	97	81	66	73	88	82	85	94
Unigram & Patient Attribute Features										
CRF Features (PI Sents)	97	96	97	60	71	65	79	83	81	94
VPA Lexicon Features (PI Sents)	98	98	**98**	79	78	**78**	88	88	**88**	**96**

Table 3: Experimental Results

best unigram-based SVMs.

The bottom section of Table 3 shows results for classifiers with both unigrams (binary) and patient attribute features. Using the CRF features increases recall on the general posts from 54 $\rightarrow$ 60, but decreases precision from 79 $\rightarrow$ 71. Using the patient attribute features from the VPA lexicon yields a substantial improvement. Recall improves from 54 $\rightarrow$ 79 and precision is just one point lower. Overall, the macro-averaged F score across the two categories jumps from 81% to 88%.

We performed paired bootstrap testing (Berg-Kirkpatrick et al., 2012)) to determine whether the SVM with unigrams and VPA lexicon features is statistically significantly better than the best SVM with only unigram features (binary). The SVM with unigrams and VPA lexicon features produces significantly better F scores at the $p < 0.05$ level for general post classification as well as the macro average. The F score for patient-specific classification and overall accuracy are statistically significant at the $p < 0.10$ level.

Attribute	CRF Tagger	VPA Lexicon
SPECIES/BREED	270	1045
NAME	36	43
POSSESSOR	12	233
AGE	545	1773
GENDER	153	338
WEIGHT	27	83
DISEASE/SYMPTOM	220	2673

Table 4: Number of Attributes Labeled in Test Set

Finally, we did an analysis to understand why the VPA lexicon was so much more effective than the CRF tagger when used to create features for text classification. Table 4 shows the number of words in PI sentences (identified by the classifier) of the test set that were labeled as patient attributes by the CRF tagger or the VPA lexicon. The VPA lexicon clearly labeled many more terms, and the additional coverage made a big difference for the text classifier.

5 Conclusions

This work demonstrated how annotated data can be leveraged to automatically harvest a domain-specific lexicon from a large collection of unannotated texts. Our induced VPA lexicon was then used to create patient attribute features that improved the ability of a document classifier to distinguish between patient-specific message board posts and general posts. We believe that this approach could also be used to create specialized lexicons for many other domains and applications. A key benefit of inducing lexicons from unannotated texts is that they provide additional vocabulary coverage beyond the terms found in annotated data sets, which are usually small.

6 Acknowledgements

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 us with samples of their data.

References

D. Baker and A. McCallum. 1998. Distributional clustering of words for text classification. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*.

T. Berg-Kirkpatrick, D. Burkett, and D. Klein. 2012. An Empirical Investigation of Statistical Significance in NLP. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing*.

S. Bloehdorn and A. Hotho. 2006. Boosting for text classification with semantic features. In *Advances in Web mining and Web usage Analysis*.

H. Borko and M. Bernick. 1963. Automatic Document Classification. *J. ACM*, 10(2):151–162.

I. Br, J. Szab, and A. Benczr. 2008. Latent dirichlet allocation in web spam filtering. In *Proceedings of the 4th international workshop on Adversarial information retrieval on the web*.

A. Carlson, J. Betteridge, B. Kisiel, B. Settles, R. Estevam, J. Hruschka, and T. Mitchell. 2010. Toward an Architecture for Never-Ending Language Learning. In *Proceedings of the Twenty-Fourth National Conference on Artificial Intelligence*.

O. Etzioni, M. Cafarella, A. Popescu, T. Shaked, S. Soderland, D. Weld, and A. Yates. 2005. Unsupervised Named-Entity Extraction from the Web: An Experimental Study. *Artificial Intelligence*, 165(1):91–134.

J. Furnkranz, T. Mitchell, and E. Riloff. 1998. A Case Study in Using Linguistic Phrases for Text Categorization from the WWW. In *Working Notes of the AAAI/ICML Workshop on Learning for Text Categorization*.

W. Hoyle. 1973. Automatic Indexing and Generation of Classification Systems by Algorithm. *Information Storage and Retrieval*, 9(4):233–242.

T. Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the European Conference on Machine Learning (ECML)*.

S. Keerthi and D. DeCoste. 2005. A Modified Finite Newton Method for Fast Solution of Large Scale Linear SVMs. *Journal of Machine Learning Research*.

Z. Kozareva, E. Riloff, and E. Hovy. 2008. Semantic Class Learning from the Web with Hyponym Pattern Linkage Graphs. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-08)*.

J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning*.

H. Lodhi, J. Shawe-Taylor, N. Christianini, and C. Watkins. 2001. Text classification using string kernels. In *Advances in Neural Information Processing Systems (NIPS)*.

T. McIntosh and J. Curran. 2009. Reducing Semantic Drift with Bagging and Distributional Similarity. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics*.

G. Miller. 1990. Wordnet: An On-line Lexical Database. *International Journal of Lexicography*, 3(4).

K. Nigam, A. McCallum, S. Thrun, and T. Mitchell. 2000. Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning*, 39(2-3):103–134, May.

E. Riloff and R. Jones. 1999. Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence*.

E. Riloff and W. Lehnert. 1994. Information Extraction as a Basis for High-Precision Text Classification. *ACM Transactions on Information Systems*, 12(3):296–333, July.

B. Roark and E. Charniak. 1998. Noun-phrase Co-occurrence Statistics for Semi-automatic Semantic Lexicon Construction. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*, pages 1110–1116.

S. Scott and M. Stan. 1998. Text classification using WordNet hypernyms. In *In Use of WordNet in Natural Language Processing Systems: Proceedings of the Conference*.

F. Sebastiani. 2002. Machine learning in automated text categorization. In *ACM computing surveys (CSUR)*.

V. Vyas and P. Pantel. 2009. Semi-automatic entity set refinement. In *Proceedings of North American Association for Computational Linguistics / Human Language Technology (NAACL/HLT-09)*.

P. Wang, J. Hu, H. Zeng, and Z. Chen. 2009. Using Wikipedia knowledge to improve text classification. In *Knowledge and Information Systems*.

S. Zelikovitz and H. Hirsh. 2001. Using LSI for text classication in the presence of background text. In *Proceedings of the 10th International Conference on Information and Knowledge Management (CIKM)*.