Determining Smoker Status using Supervised and Unsupervised Learning with Lexical Features

Ted Pedersen
University of Minnesota, Department of Computer Science, Duluth, MN, USA

Abstract

This paper describes three University of Minnesota, Duluth systems that participated in the I2B2 NLP smoker–status challenge. The task was to identify if a patient was a smoker based the content of their medical record. We took both supervised and unsupervised learning approaches. The one supervised system learned a decision tree from 398 manually annotated training records provided by the task organizers. The two unsupervised methods used that same training data (minus the annotations) to construct feature vectors that were averaged together to create a second order representation of the contexts that were then clustered. Our supervised method resulted in an accuracy of 82% on the evaluation data, while the unsupervised methods attained 68% and 69%. Simply predicting the most frequent smoker–status class (unknown) results in accuracy of 61%.

INTRODUCTION

We cast the smoker–status challenge as a problem in text classification, where we assign medical records to one of five smoker–status categories. We also approach it as a problem in text clustering, where we attempt to group the records into some number of clusters, where each cluster is associated with a smoker–status. We view each medical record as a short document, and rely upon lexical features that are identified in the training data using frequency cutoffs or measures of association. There were three Duluth systems that participated in the challenge, one a supervised decision tree learner, and two unsupervised systems that rely on the use of second–order representations of context.

Our supervised learning techniques are based to some degree on our previous work in word sense disambiguation (e.g., [1], [2]). However, the smoker–status challenge is distinct and can not be approached identically. The goal of word sense disambiguation is to assign a meaning to a given word based on the surrounding context. The process of feature selection is somewhat simplified since the features that bear most directly on the target word's meaning will typically be in close proximity. However, in the medical records for the smoker–status challenge, there is no single target word, and in fact the smoking status is often of secondary concern or is simply not an issue in the record.

The unsupervised learning methods are based on our previous work in automatic discovery of word senses (e.g., [3]) and email clustering (e.g., [4]). The latter is of particular relevance, since an email message is a short document with no particular target word. The goal of email clustering is to categorize a message based on its overall topic, and as such there is no single focal point like a target word, and the entire text must be considered.

This paper continues with an overview of our supervised and unsupervised methods, and then describes the results of our system on both the training and evaluation data.

METHODS

We describe a baseline measure of performance that can be derived from the distribution of smoker–status categories in the data. Then we introduce the lexical features that our supervised and unsupervised systems utilize. Finally, we briefly summarize our supervised and unsupervised techniques.

Most Frequent Category Baseline

The training data consists of 398 records, and the evaluation (test) data consists of 104 records. The distribution of smoker–status categories in both samples of data is shown in Table 1.

The percentage of records associated with the most frequent category (UNKNOWN) serves as a

Table 1: Smoker–Status Distribution

	Training		Test	
status	%	freq	%	freq
UNKNOWN	63.3	252	60.6	63
NON-SMOKER	16.6	66	15.4	16
PAST-SMOKER	9.0	36	10.6	11
CURR-SMOKER	8.8	35	10.6	11
SMOKER	2.3	9	2.9	3
	100.0	398	100.1	104

baseline for both our supervised and unsupervised methods.

In supervised learning, a classifier can simply learn the most common status (UNKNOWN) and be correct 63% of the time when applied to the training data, and 61% of the time when applied to the evaluation data.

In unsupervised clustering, if all the records in the training data are assigned to a single cluster, then this too would attain an accuracy of 63% if that single cluster is mapped to the most frequent category (UNKNOWN). If all of the evaluation records are placed in a single cluster, then the resulting accuracy would be 61%.

When we refer to accuracy, we simply mean the number of records that are correctly classified divided by the total number of records.

Lexical Features

We only use lexical features in our supervised and unsupervised methods. These are words or Ngrams that occur in the training data, and can be easily identified via frequency cutoffs or measures of association. We made this choice since the fragmented and noisy content of the clinical records did not seem particularly suitable for deeper linguistic analysis. We also hypothesized that there would be certain cue words or Ngrams that would indicate smoker–status directly rather than attempting to make inferences based on indirect evidence.

The lexical features we experimented with for both the supervised and unsupervised methods included unigrams, bigrams, and trigrams. Unigrams are single words that exceed a given frequency threshold in the training data. Bigrams and trigrams are pairs or trios of words that occur in a particular order and above a certain frequency and/or measure of association threshold. For Ngram features, we may also specify some number of intervening words that may occur between the first and last word.

We experimented with frequency thresholds of 2, 5, 10, and 20 for all of these features. We also used the log–likelihood ratio as a measure of association to identify bigrams and trigram features. We used a threshold value of 3.84, which is associated with a 95% chance that the words in the Ngram are not occurring together by chance (and are therefore associated). We also experimented with allowing 0, 5 and 10 intervening words within bigrams and trigrams.

The features that are identified via these thresholds are subject to a final filtering via a *stop–list* of non–content words. We used a list of 472 words, where 392 are mainly function words such as determiners and conjunctions. We also identified 80 words that occurred in more than half of the training records, and included those in the stop–list. Finally, we did not consider single characters or numeric values as features, and we converted the text to lower case prior to feature identification.

Supervised Learning

We experimented with a number of supervised learning algorithms that are supported in the Weka Machine Learning toolkit. These included a Support Vector Machine (SMO), a decision tree learner (J48) and a Naive Bayesian Classifier. These represent a broad cross–section of machine learning methodologies, and have been shown to perform well in text classification and related tasks such as word sense disambiguation.

These methods learn a model that attempts to cover or describe as many of the training examples as possible using the features we select, without over–fitting the data. We determined which model performs most accurately using 10–fold cross validation on the training data, and then applied that model to the evaluation data after refining the selected features.

In 10–fold cross validation, the training data is divided into ten equal sized pieces, where nine of these are used for training and one for evaluation. This process is repeated ten times so that each piece of training data serves as the evaluation data once and is assigned a smoking status category based on the model learned from the other nine pieces, and overall accuracy for the entire training data is computed.

Unsupervised Methods

We explored a variety of different clustering algorithms, but generally found that once the features were selected, there were no significant differences in the results obtained from k–means, the method of repeated bisections, or average–link agglomerative clustering.

The feature selection method used for unsupervised learning was the same as for supervised. After the features are selected, the records to be clustered are represented using two different schemes. The first is based on Latent Semantic Analysis. A feature by record matrix is constructed from the evaluation data, which shows how many times each of the features occurs in each of the records, where features can be unigrams, bigrams, or trigrams. This matrix is then reduced to 10% of its original number of column (records) via Singular Value Decomposition. The goal of this reduction is to group together records that are in some way related, to reduce the noise in the evaluation data. Then, each evaluation record is represented by replacing each of the features that occur in it with a vector that shows in which records that feature occurs. The centroid of these feature vectors then acts as the representation of the record. Thus, a record is represented by a vector that shows in which other records the features that occur within it occur.

The second method is the native SenseClusters second–order representation. This requires the use of bigram features in the training data from which a word by word co–occurrence matrix is constructed. The rows of the matrix represent the first word in the bigram, and the columns represent the second. The cells in the matrix contain the measure of association score for the bigram that led to its selection as a feature. This word by word matrix is reduced by Singular Value Decomposition as well (to 10% of its original number of columns) in an effort to limit the noise and the effect of polysemy that is present in word co–occurrence data. Each word in an evaluation record is replaced with its corresponding word vector, and these word vectors are averaged together to create the representation of the record.

After the record representations are created via either method, then clustering proceeds. The number of clusters is automatically determined using the PK2 measure [5], which compares the value of the clustering criterion function at successive numbers of clusters and stops when there is no significant improvement in the quality of the clustering solution. The clusters are assigned categories based on the distribution of categories each method successfully discovered in the training data.

We evaluated the efficacy of the different possible unsupervised formulations by clustering the training data using the features identified therein, and measuring the agreement of the discovered clusters to the actual smoker–status categories found in the data. It should be stressed that the manually annotated examples were only used for evaluation and were not used to determine features.

RESULTS AND DISCUSSION

We report results both on the training data and on the evaluation data.

Supervised Learning

The J48 decision tree learner was the most accurate method when evaluated using 10–fold cross validation on the training data. It achieved its highest levels of accuracy when using unigram features that occurred 5 or more times in the training data. The tree learned with these features had 24 leaves and 47 nodes. This can be viewed as specifying 24 different paths or rules through the tree, each of which indicates the smoker–status based on the combination of unigrams that occur (or not) in the evaluation record. The accuracy of the learned tree on the training data was 82.2%, meaning that 327 records were classified correctly, and 71 were not.

When we manually inspected this tree, we noticed a few features that were clearly spurious (e.g., *curvature* and *undergone*). This is not surprising, since there were over 3,600 unigram features identified in the training data, and even a few wrong choices in building the decision tree could result in the inclusion of features that were not strictly necessary. Thus, we removed those features that we did not believe were associated with smoking–status. This left us with a set of of nine unigram features:

> cigarette, drinks, quit, smoke, smoked, smoker, smokes, smoking, tobacco

These are binary features that indicate if the given word occurs in a record or not. We then learned the decision tree again using just these features to represent the training data. The resulting tree included all of these features, and is shown in Figure . This tree has 10 leaves and 19 nodes, so the number of paths through the tree (i.e., rules) has dropped from 24 to 10.

The numbers in parenthesis in Figure indicate the number of examples in the training data that are covered by that rule, and how many of those are classified incorrectly by that rule. Thus, there were 253 training records that were assigned the status UNKNOWN based on the rule that quit,

```
quit = 0
| smoking = 0
| | smoker = 0
| | | tobacco = 0
| | | | smoke = 0
| | | | | drinks = 0
| | | | | | cigarette = 0
| | | | | | | smoked = 0: UNKNOWN (253/3)
| | | | | | | smoked = 1: PAST-SMOKER (2/1)
| | | | | | cigarette = 1: NON-SMOKER (3/1)
| | | | | drinks = 1: NON-SMOKER (6/3)
| | | | smoke = 1: NON-SMOKER (16)
| | | tobacco = 1
| | | | smokes = 0: NON-SMOKER (39/7)
| | | | smokes = 1: CURRENT-SMOKER (2)
| | smoker = 1: CURRENT-SMOKER (11/5)
| smoking = 1: CURRENT-SMOKER (42/22)
quit = 1: PAST-SMOKER (24/4)
```

Figure 1: J48 Tree from Training Data (9 features)

Table 2: J48 on Training Data (9 features)

a	b	c	d	e	classified as
20	5	1	10	0	a = PAST-SMOKER
0	51	2	13	0	b = NON-SMOKER
0	1	250	1	0	c = UNKNOWN
5	4	2	24	0	d = CURR-SMOKER
0	3	1	5	0	e = SMOKER

smoking, smoker, tobacco, smoke, drink, cigarette, and smoke did not occur. Of those, only 3 were incorrectly assigned. This indicates that there were only 3 training records where the status was something other than UNKNOWN and none of those words occurred.

The accuracy of this smaller tree on the training data increased to 86.7%, meaning that 345 patients were classified correctly, and 53 were not. For the evaluation data, the decision tree learned from the training data with the reduced set of features attained an accuracy of 82%, meaning that 85 of 104 records were classified correctly.

Confusion matrices for the training and evaluation data are shown in Tables 2 and 3. These show the true distribution of smoker–status along the rows, and the predicted status on the columns. The values in the diagonals of the matrix indicate when the predicted and actual status agreed, and the off diagonals show the errors that were made.

Table 3: J48 on Evaluation Data (9 features)

a	b	c	d	e	classified as
62	0	1	0	0	a = UNKNOWN
1	10	1	0	4	b = NON-SMOKER
0	2	4	0	5	c = PAST-SMOKER
0	0	0	0	3	d = SMOKER
0	1	1	0	9	e = CURR-SMOKER

Unsupervised Learning

For unsupervised learning we found that bigrams that occur 2 or more times in the training data with up to 5 intervening words were the most effective features. We then followed an intuition similar to that used in supervised learning, and limited the bigram features to pairs where one of the words began the string *smok*. This would include *smoker*, *smoking*, *smoked*, etc. This resulted in a total of 96 bigram features, of which the most frequent 15 are shown below as examples:

> social smoking, pack smoking, smoking alcohol, smoking family, smoke drink, cigarette smoking, allergies smoking, allergies smoked, smoking quit, quit smoking, smoker drinks, former smoker, social smoke, denies smoking, habits smoking, ...

These bigrams are identified as features in an evaluation record if both words occur in the given order within 5 words (or less) of each other.

We found that the results on the training data from Latent Semantic Analysis and the native SenseClusters second–order method are very similar. For the training data, LSA attains an accuracy of 68.1%, getting 271 of 398 correct. Second order SenseClusters attains an accuracy of 68.3%, getting 272 of 298 correct.

When applied to the the evaluation data, the two methods again performed at nearly identical levels of accuracy. LSA attains 68%, getting 71 of 104 correct, while second–order SenseClusters achieves 69%, getting 72 correct. This is essentially a tie, as it was in the case of the training data.

Tables 4 and 5 show the confusion matrices for LSA and SenseClusters with the evaluation data. These show that the two unsupervised methods agreed in nearly all cases, only differing with respect to six records. It is also clear that both unsupervised methods identified three clusters rather than five, since there are two columns in each confusion matrix made up of zeros. This seems quite reasonable, since the distinction between SMOKER, CURRENT-SMOKER, and PAST-SMOKER is rather subtle. If these three categories are collapsed into a single category, then SenseClusters reaches accuracy of 79%, and LSA attains 77%.

Table 4: LSA on Evaluation Data

a	b	c	d	e	classified as
63	0	0	0	0	a = UNKNOWN
10	0	0	0	6	b = NON-SMOKER
1	3	0	0	7	c = PAST-SMOKER
1	0	0	0	2	d = SMOKER
2	1	0	0	8	e = CURR-SMOKER

Table 5: SenseClusters on Evaluation Data

a	b	c	d	e	classified as
63	0	0	0	0	a = UNKNOWN
10	0	0	0	6	b = NON-SMOKER
2	1	0	0	8	c = PAST-SMOKER
1	0	0	0	2	d = SMOKER
2	0	0	0	9	e = CURR-SMOKER

The results in both the supervised and unsupervised experiments are characterized by the fact that the UNKNOWN category dominates the distribution, and can be determined based on the absence of a few lexical features (such as we employed in the supervised experiments).

Supervised learning with our very simple decision tree was quite effective, except in the case of CURRENT-SMOKER, where many records were categorized as current-smoker when they were something else (false positives). These errors are spread out fairly evenly amongst the different actual categories of records, as shown in column d in Table 2.

The unsupervised methods tended to assign most of the records to a single cluster that was associated with the UNKNOWN category. Two smaller clusters were created that were associated with CURRENT-SMOKER and NON-SMOKER, as shown in Tables 4 and 5.

CONCLUSION

We described three systems from the University of Minnesota, Duluth that participated in the smoker–status task in I2B2 NLP challenge. There was one supervised system that learned a simple decision tree from a manually refined feature set. This attained 10–fold cross validation accuracy of 87% on the training data, and 82% on the evaluation data. Results of two unsupervised clustering systems attained accuracy of 68% on the training data, and 68% and 69% on the evaluation data. All of these results exceed the baseline established by the most frequent status, which is 63% in the training data and 61% in the test data.

If manually annotated training data is available, then supervised methods are tremendously effective. However, if no such data is available then reasonable results can be obtained via unsupervised methods, especially if there is a reasonably large sample of records available.

The supervised experiments were carried out with the SenseTools[1] package, which integrates the Ngram Statistics Package and the Weka Machine Learning Toolkit in order to extract features and learn classification models from the data.

The unsupervised experiments were done using the SenseClusters[2] package, which integrates the Ngram Statistics Package, SVDPACKC, and the Cluto Clustering Toolkit in order to extract features, reduce dimensionality and cluster data.

Both SenseTools and SenseClusters are freely available open source projects developed at the University of Minnesota, Duluth.

Acknowledgments

This research was partially supported by a National Science Foundation Faculty Early CAREER Development Award (#0092784).

Address for Correspondence

Ted Pedersen, University of Minnesota, Duluth, Department of Computer Science, 1114 Kirby Drive, Duluth, MN 55812, USA
tpederse@d.umn.edu

References

[1] T. Pedersen. A decision tree of bigrams is an accurate predictor of word sense. In *Proceedings of the Second Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 79–86, Pittsburgh, July 2001.

[2] T. Pedersen. Machine learning with lexical features: The Duluth approach to senseval-2. In *Proceedings of the Senseval-2 Workshop*, pages 139–142, Toulouse, July 2001.

[3] A. Purandare and T. Pedersen. Word sense discrimination by clustering contexts in vector and similarity spaces. In *Proceedings of the Conference on Computational Natural Language Learning*, pages 41–48, Boston, MA, 2004.

[4] A. Kulkarni and T. Pedersen. Name discrimination and email clustering using unsupervised clustering and labeling of similar contexts. In *Proceedings of the Second Indian International Conference on Artificial Intelligence*, pages 703–722, Pune, India, December 2005.

[5] T. Pedersen and A. Kulkarni. Selecting the right number of senses based on clustering criterion functions. In *Proceedings of the Posters and Demo Program of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics*, pages 111–114, Trento, Italy, April 2006.

[1] http://www.umn.edu/home/tpederse/sensetools.html

[2] http://senseclusters.sourceforge.net