Integrating Background Knowledge Into Text Classification

Sarah Zelikovitz
College of Staten Island – CUNY
2800 Victory Blvd
Staten Island, NY 11694
zelikovitz@postbox.csi.cuny.edu

Haym Hirsh
Rutgers University
110 Frelinghuysen Road
Piscataway, NJ 08854
hirsh@cs.rutgers.edu

Abstract

We present a description of three different algorithms that use background knowledge to improve text classifiers. One uses the background knowledge as an index into the set of training examples. The second method uses background knowledge to reexpress the training examples. The last method treats pieces of background knowledge as unlabeled examples, and actually classifies them. The choice of background knowledge affects each method's performance and we discuss which type of background knowledge is most useful for each specific method.

1 Using Background Knowledge

Supervised learning algorithms rely on a corpus of labeled training examples to produce accurate automatic text classifiers. An insufficient number of training examples often results in learned models that are suboptimal when classifying previously unseen examples. Numerous different approaches have been taken to compensate for the lack of training examples. These include the use of unlabeled examples [Bennet and Demiriz, 1998; Blum and Mitchell, 1998; Nigam *et al.*, 2000; Goldman and Zhou, 2000], the use of test examples [Joachims, 1999], and choosing a small set of specific unlabeled examples to be manually classified [Lewis and Gale, 1994].

Our approach does not assume the availability of either unlabeled examples or test examples. As a result of the explosion of the amount of data that is available, it is often the case that text, databases and other sources of knowledge that are related to the text classification task are readily available from the World Wide Web. We incorporate such "background knowledge" into different learners to improve classification of unknown instances. The use of external readily available textual resources allows learning systems to model the domain in a way that would be impossible by simply using a small set of training instances. For example, if a text classification task is to determine the sub-discipline of physics that a paper *title* should belong to, background knowledge such as abstracts, physics newsgroups, and perhaps even book reviews of physics books can be used by learners to create more accurate classifiers.

We present three methods of incorporating background knowledge into the text classification task. Each of these methods uses the corpus of background knowledge in a different way, yet empirically, on a wide variety of text classification tasks we can show that accuracy on test sets can be improved when incorporating background knowledge into these systems. We ran all three methods incorporating background knowledge on a range of problems from nine different text classification tasks. Details on the data sets can be found at (www.cs.csi.cuny.edu/~zelikovi/datasets; each varied on the size of each example, the size of each piece of background knowledge, the number of examples and number of items of background knowledge, and the relationship of the background knowledge to the classification task.

2 Methods

In our first approach we use Naive Bayes and EM as in [Nigam *et al.*, 2000]. We can substitute more general background knowledge for unlabeled examples, and obtain improvements in accuracy on text classifiers that are created using both the training set and the set of background knowledge. Naive Bayes classifiers make the assumption that examples (both labeled and unlabeled) have been generated by a mixture model that has a one-to-one correspondence with classes. Even if this assumption is true for the labeled data and the test data, by its very nature, background knowledge should not fit this assumption at all. However, the interesting observation that we make is that to gain leverage out of unlabeled examples, the unlabeled data that we have need not be specifically and accurately unlabeled examples. As long as the vocabulary and classification structure closely resembles the training/test data, background knowledge can improve classification accuracy in textual data using the EM algorithm.

A second approach that we take is based upon a nearest neighbor text classifier using WHIRL [Cohen, 1998; Cohen and Hirsh, 1998]. Instead of simply comparing a test example to the corpus of training examples, we use the items of background knowledge as "bridges" to connect each new example with labeled training examples. A labeled training example is useful in classifying an unknown test instance if there exists some set of unlabeled background knowledge that is similar to both the test example and the training example. We call this a "second-order" approach to classification [Zelikovitz and Hirsh, 2000; 2002], in that data are no longer di-

rectly compared but rather, are compared one step removed, through an intermediary.

Finally we use the background knowledge to redescribe both the training and the test examples. To do this, we add the background knowledge documents to the training set, to create a large, sparse term-by-document ($t \times d$) matrix. We then use Latent Semantic Indexing (LSI) [Deerwester *et al.*, 1990] to automatically redescribe textual data in a new smaller semantic space using singular value decomposition. The original space is decomposed into linearly independent dimensions or "factors", and the terms and documents of the training and test examples are then represented in this new vector space [Zelikovitz and Hirsh, 2001; 2002]. Documents with high similarity no longer simply share words with each other, but instead are located near each other in the new semantic space. Since this semantic space was created by incorporating the background knowledge, the model of the domain that it creates reflects both the training set and the background knowledge.

3 Comparison of Approaches

Different types of background knowledge are most useful for each of these three systems. The system based upon WHIRL performs best on the problems where the form and size of the background knowledge is substantially different than the training and test data. For example, we classify names of companies by area using Yahoo! pages as background knowledge. These background pieces of data are not really classifiable, in the sense that they do not necessarily belong to any specific class. Since this WHIRL-based method does not attempt to classify the background knowledge, but merely uses it to index into the training corpus, it makes the best use of this background knowledge.

For the data sets where the background knowledge fits very closely to the training and test classification task, EM outperforms the other systems. For example, EM performed best when classifying physics papers by subdiscipline using abstracts as background knowledge. This is consistent with the way EM makes use of background knowledge. Since EM actually classifies the background knowledge, and uses the background knowledge to decide on the parameters of its generative model, the closer the background knowledge is to the training and test sets, the better EM will perform. Ideally, for EM, we wish the background knowledge to be generated from the same model as the training and test sets.

Reexpressing the data and background with LSI seems to be most effective when there is very limited training data. On the smallest data sets, it outperforms all the other methods in many domains. When very few training examples exist, this method can still build a space that correctly models the domain by using the available background knowledge.

We are currently looking at methods to evaluate sets of background knowledge to determine the amount of background knowledge as well as the measure of relevance that it must have to the training set to be useful for each of these learners.

References

[Bennet and Demiriz, 1998] K. Bennet and A. Demiriz. Semi-supervised support vector machines. *Advances in Neural Information Processing Systems*, 12:368–374, 1998.

[Blum and Mitchell, 1998] A. Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pages 92–100, 1998.

[Cohen and Hirsh, 1998] William Cohen and Haym Hirsh. Joins that generalize: Text categorization using WHIRL. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, pages 169–173, 1998.

[Cohen, 1998] William Cohen. Integration on heterogeneous databases without common domains using queries based on textual similarity. In *Proceedings of ACM-SIGMOD 98*, 1998.

[Deerwester *et al.*, 1990] S. Deerwester, S. Dumais, G. Furnas, and T. Landauer. Indexing by latent semantic analysis. *Journal for the American Society for Information Science*, 41(6):391–407, 1990.

[Goldman and Zhou, 2000] S. Goldman and Y. Zhou. Enhancing supervised learning with unlabeled data. In *Proceedings of the Seventeenth International Conference on Machine Learning*, 2000.

[Joachims, 1999] T. Joachims. Transductive inference for text classification using support vector machines. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pages 200–209, 1999.

[Lewis and Gale, 1994] David D. Lewis and William A. Gale. A sequential algorithm for training text classifiers. In *SIGIR94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 3–12, 1994.

[Nigam *et al.*, 2000] Kamal Nigam, Andre Kachites Mccallum, Sebastian Thrun, and Tom Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2/3):103–134, 2000.

[Zelikovitz and Hirsh, 2000] S. Zelikovitz and H. Hirsh. Improving short text classification using unlabeled background knowledge to assess document similarity. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 1183–1190, 2000.

[Zelikovitz and Hirsh, 2001] S. Zelikovitz and H. Hirsh. Using LSI for text classification in the presence of background text. In *Proceedings of the Tenth Conference for Information and Knowledge Management*, 2001.

[Zelikovitz and Hirsh, 2002] S. Zelikovitz and H. Hirsh. Integrating background knowledge into nearest-Neighbor text classification. In *Proceedings of the 6th European Conference on Case Based Reasoning*, 2002.