Methods for Augmenting Semantic Models with Structural Information for Text Classification

Jonathan M. Fishbein[1] and Chris Eliasmith[1,2]

[1] Department of Systems Design Engineering
[2] Department of Philosophy
Centre for Theoretical Neuroscience
University of Waterloo
200 University Avenue West
Waterloo, Canada
{jfishbei,celiasmith}@uwaterloo.ca

Abstract. Current representation schemes for automatic text classification treat documents as syntactically unstructured collections of words or 'concepts'. Past attempts to encode syntactic structure have treated part-of-speech information as another word-like feature, but have been shown to be less effective than non-structural approaches. Here, we investigate three methods to augment semantic modelling with syntactic structure, which encode the structure across all features of the document vector while preserving text semantics. We present classification results for these methods versus the Bag-of-Concepts semantic modelling representation to determine which method best improves classification scores.

Keywords: Vector Space Model, Text Classification, Parts of Speech Tagging, Syntactic Structure, Semantics.

1 Introduction

Successful text classification is highly dependent on the representations used. Currently, most approaches to text classification adopt the 'bag-of-words' document representation approach, where the frequency of occurrence of each word is considered as the most important feature. This is largely because past approaches that have tried to include more complex structures or semantics have often been found lacking [1], [2].

However, these negative conclusions are premature. Recent work that employs automatically generated semantics using Latent Semantic Analysis and Random Indexing have been shown to be more effective than bag-of-words approaches in some circumstances [3]. As a result, it seems more a matter of determining how best to represent semantics, than of whether or not semantics is useful for classification.

Here we demonstrate that the same is true of including syntactic structure. A recent comprehensive survey suggests that including parse information will not help classification [2]. However, the standard method for including syntactic information is simply to add the syntactic information as a completely new,

C. Macdonald et al. (Eds.): ECIR 2008, LNCS 4956, pp. 575–579, 2008.

independent feature of the document. In contrast, the methods we investigate in this paper take a very different approach to feature generation by distributing syntactic information across the document representation, thus avoiding the limitations of past approaches.

2 Bag-of-Concepts and Context Vectors

The Bag-of-Concepts (BoC) [3] text representation is a recent text representation scheme meant to address the deficiencies of the Bag-of-Words (BoW) representations by implicitly representing synonymy relations between document terms. BoC representations are based on the intuition that the meaning of a document can be considered as the union of the meanings of the terms in that document. This is accomplished by generating term context vectors for each term within the document, and generating a document vector as the weighted sum of the term context vectors contained within that document.

Reducing the dimensionality of document term frequency count vectors is a key component of BoC context vector generation. We use the random indexing technique [4] to produce these context vectors in a more computationally efficient manner than using principal component analysis (PCA).

BoC representations still ignore the large amount of syntactic data in the documents not captured implicitly through word context co-occurrences. For instance, although BoC representations can successfully model some synonymy relations, since different words with similar meaning will occur in the same contexts, it can not model polysemy relations. For example, consider the word “can”. Even though the verb form (i.e., “I *can* perform that action.”) and the noun form (i.e., “The soup is in the *can*.”) of the word occur in different contexts, the generated term vector for “can” will be a combination of these two contexts in BoC. As a result, the representation will not be able to correctly model polysemy relations involving a word that can be used in different parts of speech.

3 Methods for Syntactic Binding

To solve the problem of modeling certain polysemy relations in natural language text, we need a representation scheme that can encode both the semantics of documents, as well as the *syntax* of documents. We will limit syntactic information to a collapsed parts-of-speech (PoS) data set (e.g.: nouns, verbs, pronouns, prepositions, adjective, adverbs, conjunctions, and interjections), and look at three methods to augment BoC semantic modelling with this information.

3.1 Multiplicative Binding

The simplest method that we investigate is multiplicative binding. For each PoS tag in our collapsed set, we generate a unique random vector for the tag of the same dimensionality as the term context vectors. For each term context vector, we perform element-wise multiplication between that term’s context vector and its identified PoS tag vector to obtain our combined representation for the term. Document vectors are then created by summing the the document's combined term vectors.

3.2 Circular Convolution

Combining vectors using circular convolution is motivated by Holographic Reduced Representations [5]. For each PoS tag in our collapsed set, we generate a unique random vector for the tag of the same dimensionality as the term context vectors. For each term context vector, we perform circular convolution, which binds two vectors $\underline{A} = (a_0, a_1, \ldots, a_{n-1})$ and $\underline{B} = (b_0, b_1, \ldots, b_{n-1})$ to give $\underline{C} = (c_0, c_1, \ldots, c_{n-1})$ where $\underline{C} = \underline{A} \otimes \underline{B}$ with $c_j = \sum_{k=0}^{n-1} a_k b_{j-k}$ for $j = 0, 1, \ldots, n-1$ (all subscripts are modulo-n). Document vectors are then created by summing the document's combined term vectors.

There are a number of properties of circular convolution that make it ideal to use as a binding operation. First, the expected similarity between a convolution and its constituents is zero, thus differentiating the same term acting as different parts of speech in similar contexts. As well, similar semantic concepts bound to the same part-of-speech will result in similar vectors; therefore, usefully preserving the original semantic model.

3.3 Text-Based Binding

Text-based binding combines a word with its PoS identifier before the semantic modelling is performed. This is accomplished by concatenating each term's identified PoS tag name with the term's text. Then, the concatenated text is used as the input for Random Indexing to determine the term's context vector. Document vectors are then created by summing the the document's term vectors.

4 Experimental Setup

We performed Support Vector Machine (SVM) classification experiments[1] in order to investigate the classification effectiveness of our syntactic binding methods compared against the standard BoC representation. For the experiments in this paper, we used a linear SVM kernel function (with a slack parameter of 160.0) and fix the dimensionality of all context vectors to 512 dimensions[2]. We used the 20 Newsgroups corpus[3] as the natural language text data for our experiments. In these classification experiments, we used a one-against-all learning method

[1] We used the SVM^{perf} implementation, which optimizes for $\mathcal{F}_1$ classification score, available at http://svmlight.joachims.org/svm_perf.html.

[2] The dimensionality of the vectors has been chosen to be consistent with other work. There is as yet no systematic characterization of the effect of dimensionality on performance.

[3] Available at http://people.csail.mit.edu/jrennie/20Newsgroups/.

employing 10-fold stratified cross validation[4]. The SVM classifier effectiveness was evaluated using the $\mathcal{F}_1$ measure. We present our aggregate results for the corpus as macro-averages[5] over each document category for each classifier.

5 Results

Table 1 shows the macro-averaged $\mathcal{F}_1$ scores for all our syntactic binding methods and the baseline BoC representation under SVM classification. All of the syntactic binding methods produced higher $\mathcal{F}_1$ scores than the BoC representation, thus showing that integrating PoS data with a text representation method is beneficial for classification. The circular convolution method produced the best score, with a macro-averaged $\mathcal{F}_1$ score of 58.19, and was calculated to be statistically significant under a 93.7% confidence interval.

Table 1. Macro-Averaged SVM $\mathcal{F}_1$ scores of all methods

Syntactic Binding Method	$\mathcal{F}_1$ **Score**
BoC (No Binding)	56.55
Multiplicative Binding	57.48
Circular Convolution	**58.19**
Text-based Binding	57.41

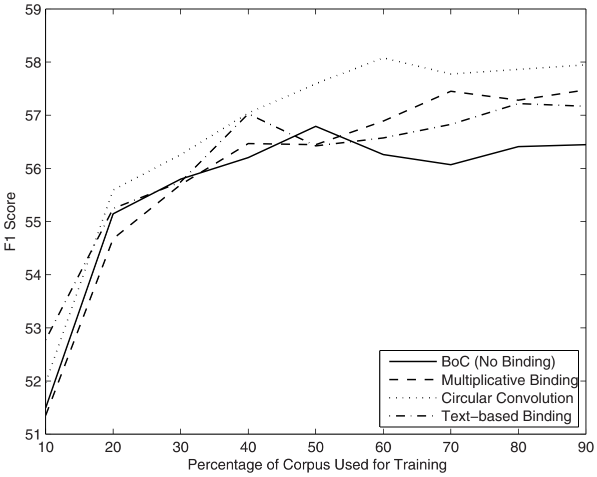

Fig. 1. Learning curves of SVM $\mathcal{F}_1$ scores of all methods

[4] This cross-validation scheme was chosen as it better reflects the statistical distribution of the documents, although produces lower $\mathcal{F}_1$ scores.

[5] Since the sizes of document categories are roughly the, same micro-averaging yields similar results and have been omitted for brevity.

The learning curves for the methods are included in Figure 2. The graph shows that circular convolution consistently produces better SVM classification results when compared to the other methods after 20% of the data is used for training. This result indicates that in situations where there is limited class data from which to learn a classification rule, combining the PoS data using circular convolution leads to the most efficient method to assist the classifier in better distinguishing the classes.

6 Conclusions and Future Research

Of all the methods investigated, the circular convolution method of binding a document's PoS information to its semantics was found to be the best. The circular convolution method had the best SVM $\mathcal{F}_1$ score and was superior using various amounts of data to train the classifiers.

Our results suggest areas of further research. One area of is to further investigate alternative binding schemes to augment text semantics, since all of the methods can bind more information than just PoS data. As well, further investigations using different corpora, such as the larger Reuters corpus, should be undertaken to examine the effectiveness of the syntactic binding methods under different text domains.

Acknowledgment. This research is supported in part by the Canada Foundation for Innovation, the Ontario Innovation Trust, the Natural Science and Engineering Research Council and the Open Text Corporation.

References

1. Kehagias, A., et al.: A comparison of word- and sense-based text categorization using several classification algorithms. Journal of Intelligent Information Systems 21(3), 227–247 (2003)
2. Moschitti, A., Basili, R.: Complex Linguistic Features for Text Classification: A Comprehensive Study. In: McDonald, S., Tait, J.I. (eds.) ECIR 2004. LNCS, vol. 2997, pp. 181–196. Springer, Heidelberg (2004)
3. Sahlgren, M., Cöster, R.: Using Bag-of-Concepts to Improve the Performance of Support Vector Machines in Text Categorization. In: Proceedings of the 20th International Conference on Computational Linguistics, pp. 487–493 (2004)
4. Sahlgren, M.: An Introduction to Random Indexing. In: Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering (2005)
5. Plate, T.A.: Holographic Reduced Representation: Distributed representation for cognitive structures. CSLI Publications, Stanford (2003)