Integrating Structure and Meaning: A New Method for Encoding Structure 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. We propose a new representation scheme using Holographic Reduced Representations (HRRs) as a technique to encode both semantic and syntactic structure. This method improves on previous attempts in the literature by encoding the structure across all features of the document vector while preserving text semantics. Our method does not increase the dimensionality of the document vectors, allowing for efficient computation and storage. We present classification results of our HRR text representations versus Bag-of-Concepts representations and show that our method of including structure improves text classification results.

Keywords: Holographic Reduced Representations, Vector Space Model, Text Classification, Part of Speech Tagging, Random Indexing, Syntax, Semantics.

1 Introduction

Successful text classification is highly dependent on the representations used. A representation of a dataset that leaves out information regarding prominent features of the dataset will result in poor performance, no matter how good the classification algorithm may be. In the case of natural language text, there are many choices which must be made in converting the raw data to high-dimensional vectors for these algorithms to process. Currently, most approaches to text classification adopt the 'bag-of-words' document representation approach, in which the grammatical structure, and semantic relationship between words in a document is largely ignored, and their frequency of occurrence is considered as most

C. Macdonald et al. (Eds.): ECIR 2008, LNCS 4956, pp. 514–521, 2008.
© Springer-Verlag Berlin Heidelberg 2008

important. This is largely because past approaches that have tried to include more complex structures or semantics have often been found lacking [1].

However, these negative conclusions are premature. More 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 [2]. 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 [1]. However, the standard method for including syntactic information is simply to add the syntactic information as a completely new, independent feature of the document. In contrast, our method takes a very different approach to feature generation by distributing syntactic information across the document representation. This avoids limitations of past approaches.

2 Bag-of-Words and Bag-of-Concepts

One of the simplest and most common text representation is the Bag-of-Words (BoW) scheme, where a document is represented as a vector of weighted (typically term frequency-inverse document frequency) word frequency counts. The dimensionality of these document vectors is typically very high; however, they are also typically very sparse.

The Bag-of-Concepts (BoC) text representation is a more recent representation scheme [2] meant to address the deficiencies of the 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. BoC representations is often significantly less than the dimensionality of BoW representation yielding better computational efficiency for classification tasks.

There have been two approaches taken to define a 'context' in BoC representations. The first is to use the Latent Semantic Indexing (LSI) model [3], which uses the entire document as a single context and each term context vector is a vector of the weighted counts in which it occurs in each document. The second is the Hyperspace Analogue to Language model [4], which uses individual words as contexts and each term context vector is a vector of the weighted counts in which it co-occurs with other words as determined by passing a fixed-size sliding window over the document. In this paper, we investigate both approaches for our new method.

3 Context Vectors and Dimensionality Reduction

Reducing the dimensionality of document term frequency count vectors is a key component of BoC context vector generation. Exploiting the Johnson-Lindenstrauss lemma [5], which states that if we project points into a random

subspace of sufficiently high dimensionality, we will approximately preserve the distances between the points, we can reduce the dimensionality of a large matrix in a more computationally efficient manner than using principal component analysis (PCA). Specifically for an $m \times n$ sparse matrix, the computational complexity of PCA using as singular value decomposition is $O(mnc)$ while the computational complexity of this random mapping is $O(nc \log m)$, where c is the number of non-zero entries per row (i.e., the average number of terms in a document). This random mapping dimensionality reduction is accomplished by multiplying a large $F_{m \times n}$ matrix by a random $R_{n \times k}$ matrix, with $k \ll n$ and where each row is constructed by randomly distributing a small number of +1s and -1s (usually around 1-2% of the matrix) and setting the rest of the elements to 0. The resulting context vector matrix FR is now $m \times k$, with the distance between every pair of rows approximately preserved from that in F.

However, performing this large matrix multiplication can be costly in terms of memory requirements, since the full $m \times n$ matrix F must be built. The random indexing technique [6], in contrast, assembles this lower dimensional matrix incrementally and avoids building this large matrix. In Random Indexing, we first create k-dimensional random index vectors for each dimension in our data, where k is significantly less than the total number of dimensions in the data. These random index vectors are created identically to the rows in the random projection matrix. Term context vectors are created by adding the context's random index vector to the term context vector every time a word occurs in a given context, The resulting term context vectors are equivalent to the ones created using the random mapping approach.

The advantage of random indexing is that it is an incremental approach, meaning that context vectors can start to be created without sampling all the data, while still avoiding the computationally costly singular value decomposition as utilized in LSI. But more importantly, random indexing avoids constructing the large context count matrix required in random mapping.

4 Limitations of Bag-of-Concepts

Sahlgren & Cöster [2] have shown that BoC has a classification advantage over BoW in certain situations. Nevertheless, the BoC scheme still ignores 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.

5 Holographic Reduced Representations

In order 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. Borrowing from a representation scheme introduced in cognitive science [7], Holographic Reduced Representations (HRRs), we can complement the BoC semantic modeling with parts of speech information to generate a more robust text representation. Eliasmith and Thagard [8] have previously shown that HRRs can be used to model both syntactic and semantic psychological data. As well, Eliasmith [9] has shown that HRRs can be successfully applied to language processing. The intuition behind this approach, is that we can "bind" part-of-speech information with a word's term context vector in order to encode both pieces of information in our representation.

HRRs use holographic transformations to encode and decode information in flat, constant dimension vectors. In order to encode the information contained within multiple vectors into a single vector, HRRs depend on circular convolution. Circular convolution 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 $j = 0, 1, \ldots, n-1$. Circular convolution is efficiently computed in time $O(n \log n)$.

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. So, the same term acting as different parts of speech in similar contexts, such as the word *can* that can act as both a noun and a verb, would not be similar in their bound HRR representation (e.g., "He kicked the *can*." would be distinct from "He *can* kick"). Second, the dimensionality of the vectors are constant under HRR operations, so the number of vectors encoded in the structure does not affect the complexity of the representation. Third, similar semantic concepts bound to the same part-of-speech result in similar vectors. So, since similarity reflects the structure of the semantic space, these binding operations usefully preserve the relevant geometric relations of the original semantic space.

HRRs also need to be combined in a manner that assembles the parts of the desired structure while preserving the similarity of the final structure to its components. For this, superposition (i.e. vector addition) is used. So if $\underline{C} = \underline{A} + \underline{B}$, $\underline{C}$ is most likely more similar to $\underline{A}$ or $\underline{B}$ than to any other vector.

6 HRR Document Representation

Our natural language representation takes advantage of the ability of HRRs to encode a document's structure in a way that is non-destructive to the document's semantics. Using the circular convolution and superposition operations of HRRs, our representation scheme can augment the semantic modeling of the BoC representations with part-of-speech information to better disambiguate document classes for classification.

We first determine the term context vectors for the data by adopting the random indexing method, described earlier. We then use a part-of-speech tagger

to extract some syntactic structure of the corpus documents. We collapse the set of possible part-of-speech tags returned by the tagger into the basic linguistic set (e.g.: nouns, verbs, pronouns, prepositions, adjective, adverbs, conjunctions, and interjections), and generate random HRR vectors of the same dimension as our term context vector for each possible tag.

To build the HRR document representation, we perform the following steps:

1. for each word in a document we take the term context vector of that word and bind it to the word's identified part-of-speech vector;
2. we take the $tf \times idf$-weighted sum of the resulting vectors in order to obtain a single HRR vector representing the document.

Like BoC document vectors, these HRR document vectors are normalized by dividing by the number of terms in the document in order to ensure that there is no classification bias to longer documents. But unlike BoC vectors, these HRR document vectors encode both semantic and syntactic information.

7 Experimental Setup

In the following sections we describe the setup for our text classification experiments. Specifically, we describe the text representations used for classification, and the classifiers and evaluation methodology used in the experiments.

7.1 Representations

We used the 20 Newsgroups corpus[1] as the natural language text data for our experiments. The purpose of these experiments was to compare the classification effectiveness of BoC and HRR text representations[2], not produce a top score for the 20 Newsgroups corpus.

The BoC representations were generated by first stemming all words in the corpus, using the Porter stemmer, to reduce the words to their root form. We then used Random Indexing to produce context vectors for the given text corpus. The dimensionality of the context vectors was fixed at 512 dimensions[3], which should be compared to the 118 673 unique stems within the corpus. We investigated the effects of both document-based context vectors and word-based context vectors. For word-based context vectors, we produced contexts using a sliding window extending 4 words in each direction from the focus word, where the term vector of the focus word was updated by adding to it the context vector of each word inside the sliding window weighted by $2^{(1-d)}$, where d is the

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

[2] We did not pursue comparison experiments with BoW representations as there are already published results (e.g. [2]) of BoW/BoC experiments in the literature.

[3] 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.

distance from the focus word. These context vectors were then $tf \times idf$-weighted and summed for each document.

The context vectors used in the HRR representations were generated in the exact same way as the BoC representations. The part-of-speech data was extracted using the Stanford Log-linear Part-of-Speech tagger[4] and random 512 dimensional HRR vectors were created for each tag in our collapsed tag set. This part-of-speech tag vector was then bound to its word's associated context vector by circular convolution, $tf \times idf$-weighted and summed for each document.

7.2 Classification and Evaluation

We performed Support Vector Machine (SVM) classification experiments[5] in order to investigate the classification effectiveness of the HRR and BoC representation. For the experiments in this paper, we used a linear SVM kernel function (with a slack parameter of 160.0). In these classification experiments, we used a one-against-all learning method employing 10-fold stratified cross validation[6]. The SVM classifier effectiveness was evaluated using the $\mathcal{F}_1$ measure.

8 Results

We only present the comparison results between the BoC text representations and HRR representations using document-based context vectors since the results for word-based context vectors showed the same comparison trends in the $\mathcal{F}_1$ scores, but produced lower total $\mathcal{F}_1$ scores.

The macro-averaged $\mathcal{F}_1$ showed that the HRR representations produced the best results, with a score of 58.19. The BoC representations produced a macro-averaged $\mathcal{F}_1$ score of 56.55. These results were calculated to be statistically significant under a 93.7% confidence interval.

Figure 1 shows the correlation between the macro-averaged SVM $\mathcal{F}_1$ scores of BoC and HRR text representations for each category in the 20 Newsgroups corpus. The graph shows that the HRR representations produce similar classification scores for some classes and significantly higher scores for other classes. This may be explained by noticing that the classes that the HRR representations outperform BoC representations are the classes in the corpus that are highly related to other classes in the corpus.

The learning curves for the representations are included in Figure 2. The graph shows that the HRR representations consistently produce better SVM classification when compared to BoC representation no matter how much of the class data is used for training. This result indicates that in situations where there is limited class data from which to learn a classification rule, the extra

[4] Available at http://nlp.stanford.edu/software/tagger.shtml.

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

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

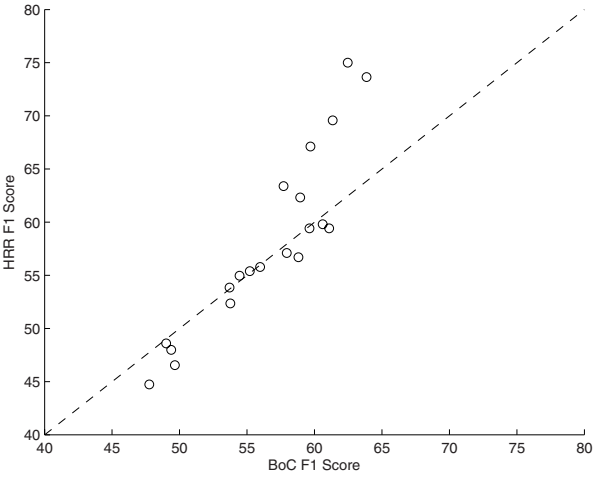

Fig. 1. Correlation between SVM $\mathcal{F}_1$ scores of BoC and HRR text representations for each corpus category

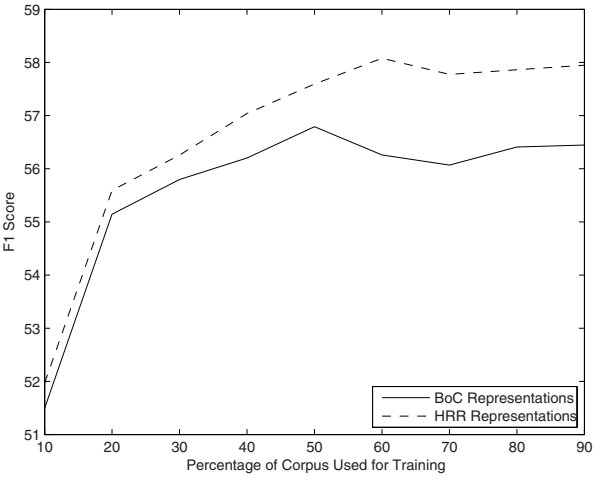

Fig. 2. Learning curves of SVM $\mathcal{F}_1$ scores of BoC and HRR text representations

part-of-speech information contained within the HRR representation assists in better classifying documents.

9 Conclusions and Future Research

Using HRRs, we have created a novel document representation scheme that encodes both the structure and the semantics of natural language text documents.

Our results show that including both the structure and semantics of natural language text in our HRR text representations can improve the text classification $\mathcal{F}_1$ score of SVM classifiers when compared to the BoC approach. We have also demonstrated the sustained superiority of the HRR representations when using various amounts of data to train the classifiers.

Our results suggest many areas of further research. We have only investigated a single natural language corpus in this paper and further investigations using different corpora should be undertaken to examine the effectiveness of HRR representations under different text domains. As well, the document vectors were fixed to 512 dimensions in the experiments, but it would be interesting to analyze the effects of the vector dimensionality on the classification results.

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. 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)
2. 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)
3. Deerwester, S.C., et al.: Indexing by latent semantic analysis. Journal of the American Society of Information Science 41(6), 391–407 (1990)
4. Lund, K., Burgess, C.: Producing high-dimensional semantic spaces from lexical co-occurrence. Behaviour Resrach Methods, Instrumentation and Computers 28(2), 203–208 (1996)
5. Johnson, W.B., Lindenstrauss, J.: Extensions to Lipshitz mapping into Hilbert space. Contemporary Mathematics 26 (1984)
6. 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)
7. Plate, T.A.: Holographic Reduced Representation: Distributed representation for cognitive structures. CSLI Publications (2003)
8. Eliasmith, C., Thagard, P.: Integrating structure and meaning: A distributed model of analogical mapping. Cognitive Science 25(2), 245–286 (2001)
9. Eliasmith, C.: Cognition with neurons: A large-scale, bilogically realistic model of the Wason task. In: Proceedings of the XXVII Annual Conference of the Cognitive Science Society (2005)