

Semantic Clustering for a Functional Text Classification Task

Thomas Lippincott and Rebecca Passonneau

Columbia University
Department of Computer Science
Center for Computational Learning Systems
New York, NY USA
tom,becky@cs.columbia.edu

Abstract. We describe a semantic clustering method designed to address shortcomings in the common bag-of-words document representation for functional semantic classification tasks. The method uses WordNet-based distance metrics to construct a similarity matrix, and expectation maximization to find and represent clusters of semantically-related terms. Using these clusters as features for machine learning helps maintain performance across distinct, domain-specific vocabularies while reducing the size of the document representation. We present promising results along these lines, and evaluate several algorithms and parameters that influence machine learning performance. We discuss limitations of the study and future work for optimizing and evaluating the method.

1 Introduction

The bag-of-words document representation achieves excellent performance on many machine learning tasks. However, this performance can be sensitive to the changes in vocabulary that occur when the training data cannot be reasonably expected to be representative of all the potential testing data. In this situation, it may be possible to exploit higher-level relationships between the vocabularies by consolidating and generalizing the specific bag-of-words features into a smaller number of semantic clusters using an external semantic resource. This would have the dual benefits of retaining performance across domains and reducing the dimensionality of the document representation.

The approach presented here was developed in response to characteristics of the machine learning phase of the ANTA_rT(1), a component of the CLiMB research project (2). CLiMB developed a toolkit for image catalogers that facilitates harvesting descriptive meta-data from scholarly text for annotating digital image collections. ANTA_rT, collaborating with experts in art history and image cataloging, developed a set of functional semantic labels to characterize how art-historical texts function with respect to their associated images (see Table 4 for the complete list), drawing on texts from four art-historical periods.

Three data sets were prepared from art history texts covering two time periods: Near Eastern art (two data sets), and Medieval art (one data set). Each

data set consisted of 25 images of works of art and the paragraphs from the text describing them. The annotators were asked to tag each sentence with the relevant functional semantic labels using a special purpose browser interface that included a training session. These annotations were then used to train and evaluate binary classifiers for each label. In this study we focus on the three most frequently-used labels, *Image Content*, *Historical Context* and *Implementation*.

The rest of the paper is organized as follows: we describe the baseline results and the reasoning that led us to investigate the clustering method, and summarize previous work with similar aims. After outlining the steps of semantic clustering, we describe the algorithms and parameters that we evaluated. We present our results along several dimensions and discuss their significance, and conclude with a discussion of future work.

2 Motivation

Table 1 shows classifier performance by train and test data for the three functional semantic labels using bag-of-words features. We swap in each of the three data sets as train and test data. The performance is highest when both data sets focus on the same time period (Near East). When the Medieval data set is used for training or testing, performance decreases dramatically. This happens to a greater degree with the *Historical Context* and *Implementation* labels than with *Image Content*, which we attribute to greater sensitivity of those labels to period-specific terminology. This shows that the bag-of-words model does not maintain its performance across historical periods for important functional semantic labels. This will be a recurring problem as data sets from different historical periods are added.

Table 1: ROC Area of support vector machine classifiers for the three functional semantic labels using bag-of-words feature sets. This shows the performance hit taken when train and test data come from different domains.

Data Sets		Label		
Train	Test	Historical Context	Implementation	Image Content
Near East 1	Near East 2	0.630	0.607	0.602
Near East 1	Medieval	0.576	0.518	0.576
Near East 2	Near East 1	0.620	0.575	0.617
Near East 2	Medieval	0.514	0.521	0.578
Medieval	Near East 2	0.573	0.564	0.597
Medieval	Near East 1	0.541	0.538	0.603

Early in the ANTArT study, we noted (3) that certain semantic word classes are correlated with the functional semantic labels. The intuition was that, while the data sets may exhibit distinct vocabularies, the distribution of words with

the same hypernyms remains discriminative. Table 2 shows pairs of terms from texts on the two historical periods that are highly-correlated with the *Historical Context* label in their respective data set and belong to the same semantic class.

Table 2: Examples of discriminative bag-of-words features for the *Historical Context* label.

		Semantic Class
Near East	Medieval	
Sumer	England	<i>Geographic region</i>
priest	monk	<i>Religious person</i>
Egyptian	Irish	<i>Nationality</i>
ancient	medieval	<i>Time period</i>
ziggurat	cathedral	<i>Place of worship</i>

Several manually enumerated semantic classes were assembled, such as *body-parts* and *time-values*. While this method addresses feature set reduction from a semantic perspective, it could not scale or generalize, as it required manual compilation of the semantic classes, solely from the training data. The clustering method presented here is an attempt to automate and generalize this intuition.

3 Previous Work

Most approaches to word clustering are statistical, using distribution, collocation and mutual information to collapse features containing redundant information into a smaller number of equivalence classes. Pereira et al. (4) describe a method that uses distribution within different syntactic contexts as the basis for clustering. Tishby et al. (5) introduced the *information bottleneck* approach to compressing feature sets while maximally preserving information about the target variable. Bekkerman et al. (6) achieved excellent results on a common multi-label classification task (20 Newsgroup) by combining this approach with support vector machines, and discuss how relevant vocabulary size can affect the potential benefits of word clustering, in the context of different labeling methodologies. Slonim et al. (7) applied the approach to a document clustering task, and report performance by the number of word clusters generated.

Few studies have made use of semantic resources in word clustering. Termier et al. (8) attempted to combine the statistical approach (latent semantic indexing) with semantic properties derived from WordNet. The results showed the semantic information actually decreasing performance of simple LSI, and while several of the theoretical benefits are discussed, they are not achieved. However, the classification task that the hybrid method was evaluated on (Reuters, by topic) does not present the issues that would benefit from using an external semantic resource.

Several algorithms have been proposed for computing “semantic similarity” between two WordNet concepts based on hypo/hypernym relationships. In addition to the three purely WordNet-based metrics we used in this study, Resnik (9), Jiang (10) and Lin (11) have proposed measures that also consider information theoretic properties of the concepts in an auxiliary corpus. Budanitsky and Hirst (12) give a high-level overview of the potential for evaluating and comparing such metrics, and note the difficulty of designing simple measurements that would be generally useful for the nebulous concept “semantic similarity”.

4 Data Sets

Table 3 compares the size of the data sets in terms of word count and vocabulary size for the two parts of speech that we consider. It also shows the inter-annotator agreement as a weighted Kappa score (13). Since annotators could assign more than one label to a sentence (in practice, about 25% of the time) we use a set-distance metric (14) to count partial agreement.

Table 3: Vocabularies of the three data sets based on WordNet’s morphology function, and Kappa agreement score between the coders.

Data set	Tokens	Nouns	Verbs	Kappa score
Near East 1	3888	734	466	0.55
Near East 2	5524	976	614	0.50
Medieval	6856	1093	652	0.56

Two annotators working independently considered each sentence in the texts, and applied one or more functional semantic labels to characterize its function with respect to the image it describes. These labels, and their usage counts by the two coders, are shown in Table 4. Because of data sparseness for the other labels, our results only consider *Image Content*, *Historical Context* and *Implementation*.

5 Methodology

Figure 1 shows an overview of the experimental procedure. It begins with separate data sets for training and testing, a functional semantic label L to consider, a part-of-speech P to use for clustering, the number of clusters N to generate, and a similarity metric M . The similarity metric maps two WordNet senses to a real number between 0 and 1.

The training sentences are tokenized and lemmatised by WordNet’s *morph* function, and the set of unique lemmas of the specified part-of-speech is extracted. A matrix of the pairwise similarities is constructed using the specified

Table 4: Functional semantic labels with each annotator’s usage in all three texts.

Label	Coder A	Coder B
<i>Image Content</i>	220	215
<i>Historical Context</i>	123	156
<i>Implementation</i>	75	138
<i>Significance</i>	59	51
<i>Interpretation</i>	67	26
<i>Comparison</i>	26	32
<i>Biographic</i>	10	6

metric. An expectation maximization clusterer is then trained on the matrix, with the specified target cluster count. The output is a clustering model that can assign an unseen word to a cluster based on its similarities to each of the training lemmas.

Document representations are then built in the same manner for the training and testing data. Each lemma in the document is clustered according to its similarities to the training lemmas. Where the bag-of-words representation records lemma frequencies, the clustering representation records cluster frequencies. These are computed by applying the clustering model from the previous step to each lemma in the document. Training and testing of an SVM classifier for the label L is then performed using the two document representations.

1. Input
 - Labeled data sets $TRAIN$ and $TEST$
 - Functional semantic label L
 - Part of speech $P = noun|verb$
 - Cluster count N
 - Similarity metric $M : sense, sense- > real$
2. Build clustering model on training data
 - Find all U unique lemmas of type P in data using WordNet’s *morphy* function
 - Construct UxU matrix M where M_{xy} is the similarity between U_x and U_y
 - Use matrix to train expectation maximization clusterer with a target cluster count N
3. Build document representation of training and testing data
 - For each lemma l in the document, apply model the vector $[M(U_1, l), M(U_2, l) \dots M(U_{|U|}, l)]$
 - Features are the frequency of each cluster $1 \dots N$
4. Train SVM binary classifier for label L on training data
5. Evaluate classification performance on testing data

Fig. 1: Overview of semantic clustering

Semantic similarity metrics operate on unambiguous senses, and so we needed to address polysemy in the texts. We tested two simple policies when calculating the similarity between two tokens. The *first sense* policy simply takes the first WordNet sense for each token, which is its most frequent sense in WordNet’s sample texts (15). The *closest sense* policy chooses the pair of senses that maximized the similarity calculation for the pair of tokens. Both methods have drawbacks: our data comes from a specific academic field with its own vocabulary, and WordNet’s sample texts may be too general (e.g. the most common general sense of “profile” is not its art-historical meaning). The *closest sense* policy is more complicated, but an interesting experimental question in itself: it may use different senses for the same token depending on the token it is being compared to. Depending on the vocabulary, it might be that a preponderance of a semantic class will overcome this noise. This is discussed by Scott et al. (16), where the *closest sense* approach is also used.

The three similarity metrics we tested all base their values on positions within WordNet’s hyper/hyponym ontology of senses, taking two senses (S_1 and S_2) as input and returning a real number. The simplest metric, *path similarity*, uses the shortest *distance* between the two senses. *Leacock Chodorow similarity* (17) also takes into account the maximum *depth* of the ontology being used. *Wu-Palmer similarity* (18) is the most sophisticated metric we evaluated. It considers the most specific ancestor (least common subsumer or *LCS*) of the two senses, its depth in the ontology, and its shortest *distance* to each of the two senses.

WordNet has separate ontologies for verbs and nouns, and we tested the clustering method independently for both. The results indicate distinctive properties of the two ontologies. Miller et al. (15) discuss fundamental differences in the idea of “hyponymy” as applied to nouns versus verbs.

We varied the target number of clusters from 5 to 100 at intervals of 5. In principle, the maximum number of clusters that could be aimed for is the vocabulary size itself, in which each lemma would become its own cluster. Our results indicate that our 100-cluster upper bound may have been too conservative for at least one of the experiments (see figure 2, top right).

Expectation maximization (19) is an iterative algorithm often used for data clustering. It associates each data instance with a probability distribution describing potential cluster membership. Iteration ends when improvement falls below some small threshold (we use 1e-6) or the number of iterations passes some maximum value (we use 100). Our results here simply map each lemma (data instance) to its likeliest cluster. In future work we may use the full probability distribution. The computation is performed using the Weka (20) implementation of expectation maximization clustering.

The support vector machine is a machine learning algorithm that has achieved widespread success on text classification tasks. It divides the training data by an N-1-dimensional plane, where N is the dimensionality of the feature representations. This division is optimized to achieve the maximum separation between the training instances. We use an extended version that handles the typical situation where the training data is not linearly separable, by penalizing misclassified

instances and optimizing the separation achieved. Its explicit design for achieving maximum generality makes it particularly attractive for our study. We use the Weka implementation of sequential minimal optimization (21), which is a method for efficiently solving the maximization problem, and a linear kernel.

Our classification results are presented as the area under the ROC (receiver operating characteristics) curve. This is particularly well-suited for evaluating classification tasks with sparse data and large skew in the label distribution (22) such as ours.

6 Results

6.1 Clustering Parameters

Figure 2 shows the average performance by number of clusters used, broken down by the part of speech used for the clusters and the functional semantic label targeted by the classifier. The most striking feature is the superior performance of the verb clusters.

While the *Image Content* label shows the highest performance, it also shows the least regularity with respect to the cluster count parameter. Its performance is likely due to it being the easiest of the labels to learn, which has been noted in earlier work (1). Its irregularity may also support the intuition that physical descriptors such as colors and dimensions are less tied to historical period.

The shaded areas in the noun cluster graphs (left side) each correspond to one of the three data sets. On the X-axis they highlight the interval from the smallest effective (performed above-random) cluster count to the largest. Their height represents the average performance across that interval. The labels all show the counter-intuitive property that the effective ranges for data sets with a larger vocabulary are always a subset of those with a smaller vocabulary. In other words, in choosing how many clusters to induce from a data set, there appears to be a narrower range of good choices for a larger vocabulary.

The black rectangles in the verb cluster graphs (right side) mark the highest performance for each data set. For the two labels that show the clearest trends in the verb clusters, *Historical Context* and *Implementation*, the data set with the smallest vocabulary peaks before the two larger data sets, supporting the intuition that a smaller vocabulary will perform best with fewer clusters. The regularity of the verb clusters, with steadily increasing performance compared to the erratic behavior of the noun clusters, lends more credibility to this observation. This distinction between verb and noun behavior must be confirmed on larger data sets before looking for an explanation in linguistic theory or resource-specific biases.

Several of the performance curves (particularly for the Near East 1 and Medieval data sets, *Historical Context* label, verb clusters) appear to be increasing at the upper limit cluster count (100). This indicates that the optimal cluster count is higher than we expected, and the testing range should be increased.

The three functional semantic labels show the same relationship when varying the sense choice iteration methods and similarity metrics (Figure 3). Better

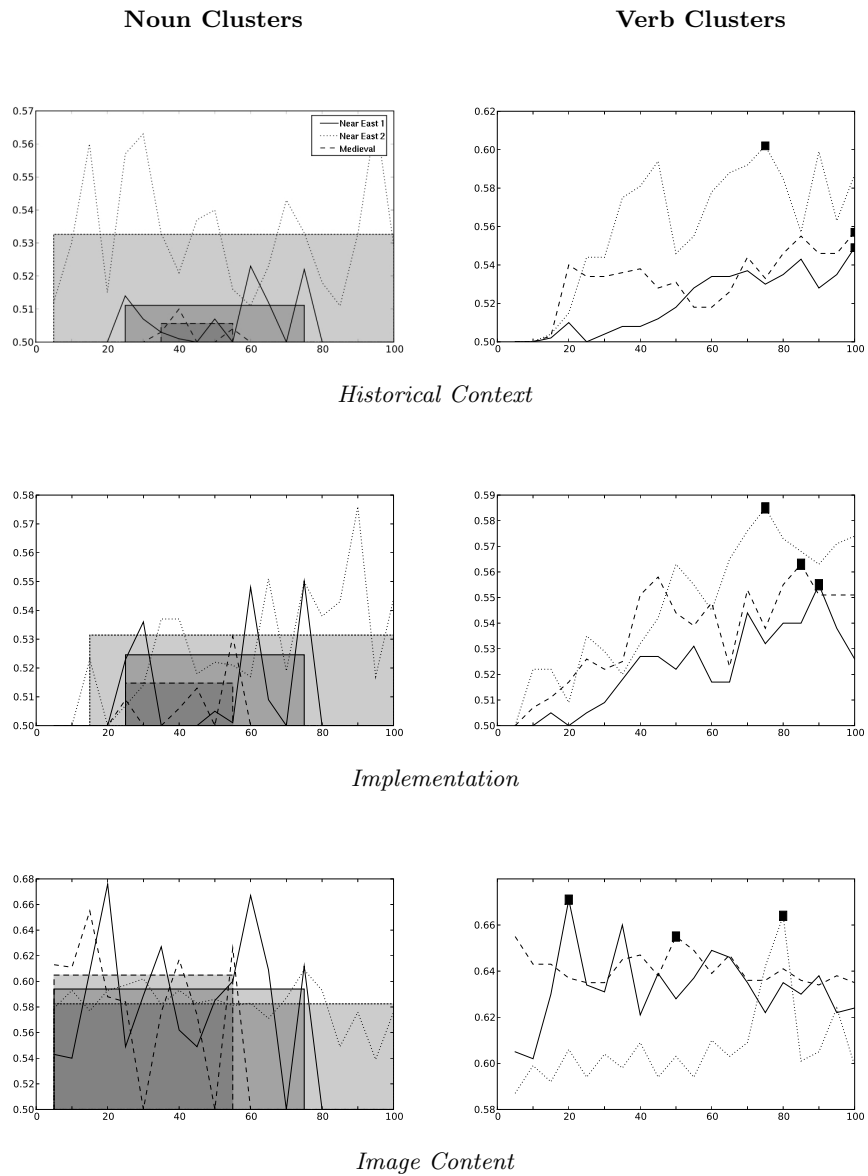


Fig. 2: Classifier performance (ROC average on Y-axis), by cluster count (X-axis), by part-of-speech for the three most common labels. The shaded regions in the left-hand figures cover the interval of cluster counts that had positive performance for each data set, and their height is the average performance over that interval. The black boxes in the right-hand figures indicate the best performance on the data set.

performance is achieved by simply using the first WordNet sense, rather than maximizing the similarity on a pairwise basis. The Wu-Palmer similarity metric outperforms the Path and Leacock Chodorow metrics. The differences are least pronounced on the *Image Content* label, which is the most domain-independent label and similar to a traditional “topical” classification. The standard deviations are massive compared to these differences, on the order of several times the magnitude: this may be due to the broad range of variables we tested, and requires further investigation.

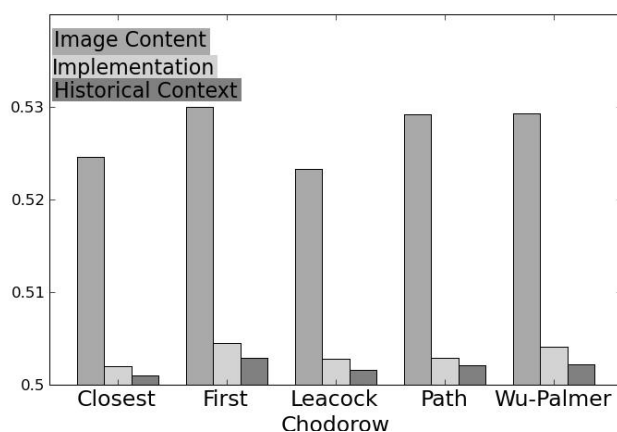


Fig. 3: Comparison of average performance above random (for ROC area, 0.5) by sense choice iteration method (“closest sense” and “first sense”) and similarity metric. Each triplet shows the average performance for all runs with that parameter, broken down by the three labels.

6.2 Cluster Quality

The original motivation for the study was the observation that there were obvious word-groups recurring above the lexical level. It is natural, therefore, to want to examine a cluster with respect to some simple characterization of its members. It is not guaranteed that a cluster will have a simple concept that describes its members, except in the cases where the clustering process mirrors this human intuition.

The example in Figure 4 demonstrates an effective cluster (roughly identified as “human body parts”) while also illustrating some shortcomings. All words in the figure are placed in the same cluster by the model, which was trained on the first Ancient Near East data set. Examining the training data manually, every recognizable body-part term is clustered appropriately, for perfect recall.

The benefit comes from the words that occur solely in the training or testing data: in a bag-of-words or statistical approach, these would have been useless features. But there are problems with the cluster’s precision: words like “vessel”, “ass” are examples of polysemy creating false positives (as in “blood vessel” and “posterior”, when they are actually used as “drinking vessel” and “donkey”). “quick” is an example of clustering on an extremely rare usage of the word (“an area of the body highly sensitive to pain”), although as a reviewer pointed out this particular example would be addressed by initial part-of-speech tagging (but consider “orb”). There is also no strict guarantee of a simple characterization of the resulting clusters. “Human physiology”, while very accurate in this case, is not precise: most people lack literal claws and beaks. This is because the clustering does not try to reconcile its results explicitly with the WordNet hierarchy. The cluster is in fact generally looking for “body parts”, and the focus on human physiology is due to the texts in question. But this is exactly the point: classifiers trained on veterinary texts could use the same cluster on texts in human medicine.

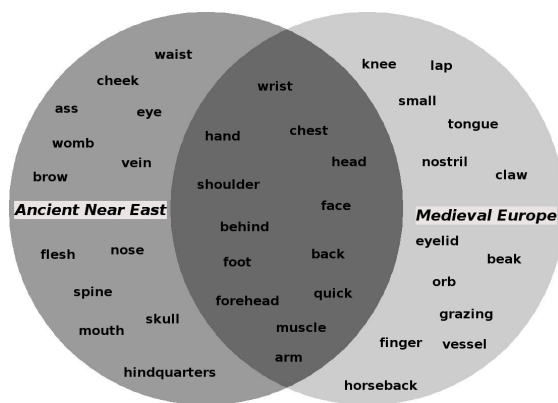


Fig. 4: Cluster example that roughly captures “human physiology”.

The same experimental run also automatically induced high-quality clusters that roughly correspond to “quantities”, “occupations”, “building materials”, “geographic regions” and so forth.

6.3 Performance

Table 5 compares the ROC areas of the bag-of-words-based classifiers with that of the *best*-performing cluster-based classifier. The clustering method outperforms bag-of-words 66% of the time when the train and test data come from different domains. This comparison assumes that we are able to choose the optimal parameters for building our clusters. Further investigation and optimization

of the parameters discussed above with respect to the vocabulary of the training data is the key to realizing this.

Table 5: Comparison of ROC Area for the three functional semantic labels, when the train and test data come from different domains. Rows with a white background use bag-of-words features, rows with a grey background use the clustering features.

Data Sets		Label		
Train	Test	Historical Context	Implementation	Image Content
Near East 1	Medieval	0.576	0.518	0.576
		0.549	0.530	0.676
Near East 2	Medieval	0.514	0.521	0.578
		0.568	0.609	0.576
Medieval	Near East 2	0.573	0.564	0.597
		0.555	0.563	0.655
Medieval	Near East 1	0.541	0.538	0.603
		0.557	0.558	0.656

6.4 Improved Generality and Dimensionality

The semantic clusters show less sensitivity to training and testing domains than the bag-of-words features. Table 6 compares the standard deviation of classifier performance for the three labels for the basic bag-of-words model to the cluster model. The results also show the relationship between the domain-sensitivity of the functional semantic label and the improvements from the clustering method.

Table 6: Standard deviation across different data set combinations

Label	Bag-of-Words	Clusters
<i>Image Content</i>	0.03279	0.02359
<i>Historical Context</i>	0.06657	0.02116
<i>Implementation</i>	0.03636	0.01903

Comparing the original bag-of-words vocabularies of the data sets with the range of cluster counts we generated shows a reduction in document representation size between 10 and 100 times. The computational benefits of this reduction are purchased at the cost of the off-line clustering process, which is resource-intensive but highly parallelizable. If it results in greater generality, it need only be done for a single domain in the given task.

7 Future Work

While space limitations prevented discussing it here, when considering any single configuration of the clustering parameters, the performance-cluster count graphs exhibit an unexpected periodic behavior. This is masked by the averages and maximums presented above, and it will require further investigation to determine what clustering parameters or characteristics of the data are responsible for this behavior.

There is an enormous amount of semantic and lexical information in WordNet that could be incorporated into the clustering process. At present, senses are associated via metrics that depend solely on directly mapping lemmas into the noun and verb hierarchies, which limits the resulting clusters in several ways (e.g. biased towards reproducing this structure, dropping information provided by adjectives, etc.). Evaluating and refining potential additions to the process (e.g. noun-adjective lexical relationships, metonymy, etc.) is a major future task.

The clustering model generated by the expectation maximization algorithm may be used more fully: rather than the simple approach of membership-frequency, the probability distribution could be used in generating the document representations. For example, in the current method, the sense “orange (the fruit)” might be counted part of a “fruit” cluster or an “edible object” cluster, but not both, even if it has near-equal probabilities of membership in either cluster. Crediting all clusters according to the lemma’s probability of membership would capture this. At the other extreme, a simpler, discrete clustering algorithm like K-means, might be more appropriate (and less computationally intensive) for our current approach. Finally, it may be possible to use fully non-parametric clustering (i.e. without a specified number of clusters) to determine the optimal cluster size, but this is complicated by the fact that optimal cluster size in this case is not simply determined by the inter/intra-cluster coherence. It also depends on the utility of the resulting clusters for generalizing in the particular domains. For example, separate clusters for “grain”, “legumes” and so forth might maximize intra-cluster coherence on some training data, but if the ideal semantic class is “agricultural products” these smaller clusters will be sub-optimal.

To address the affects of word sense ambiguity we have begun manual disambiguation on the three data sets, which will give us an upper limit on the improvements to expect from automatic approaches. Another possibility is using disambiguated corpora such as those used for SENSEVAL, which would also test the method on larger and more familiar data. This would require creating a functional semantic task, similar to the ANTA_rT project, using the new data set.

The reviewers suggested several methods and labeling tasks to compare our method with. A baseline usage of WordNet might be to use hypernym-frequency as the feature set. Sentiment analysis could be a useful source of well-explored labeling tasks that can be partitioned into distinct domains (e.g. by “product type”) for training and testing. The size of the ANTA_rT data set was a limitation, and large product reviews databases (Amazon, IMDB, etc.) could help us understand the significance of the irregular behavior of noun clusters.

Finally, while our method presented here makes no use of distributional statistics or correlations of the words and labels, combining the approaches could raise clustering performance dramatically while maintaining generality. This could be particularly useful for deciding how many clusters to generate, e.g. if and how to subdivide broad clusters. For this study we performed exhaustive tests of cluster count to find trends related to the vocabulary of the data sets. It would be useful, when choosing to partition a set of words into one large cluster or two smaller clusters, to determine how the two potential choices would relate to label distribution in the training data.

8 Conclusion

We have presented a text classification task that responds poorly to the typical bag-of-words feature space, due to the nature of the data sets and labels involved. We described a method that builds a more general and compact document representation using measures of semantic similarity, and presented results testing several options for its component algorithms and parameters. We argued that the results show several of the desirable properties of the approach, and outlined future work in constructing an implementation that optimizes performance while maintaining these properties.

This approach has potential applications for any task that uses bag-of-words for document representation. Information retrieval could use clustering to expand open class terms or handle queries of a functional semantic nature. New multilingual WordNet implementations with pointers between senses could be used to automatically extend these benefits across languages. This document representation could also prove useful for investigating more abstract cognitive processes, such as analogy and inference, and for drawing comparisons between lexical resources and the cognitive structures they try to represent.

Bibliography

- [1] Rebecca Passonneau and Tae Yano and Tom Lippincott and Judith Klavans: Functional Semantic Categories for Art History Text: Human Labeling and Preliminary Machine Learning. International Conference on Computer Vision Theory and Applications, Workshop 3: Metadata Mining for Image Understanding (2008)
- [2] Judith Klavans and Carolyn Sheffield and Eileen Abels and Joan Bedouin and Laura Jenemann and Tom Lippincott and Jimmy Lin and Rebecca Passonneau and Tandeep Sidhu and Dagobert Soergel and Tae Yano: Computational Linguistics for Metadata Building: Aggregating Text Processing Technologies for Enhanced Image Access. In: *OntoImage 2008: 2nd International "Language Resources for Content-Based Image Retrieval" Workshop*. (2008)
- [3] Tae Yano: Experiments on Non-Topical Paragraph Classification of the Art History Textbook. unpublished (2007)
- [4] Fernando Pereira and Naftali Tishby and Lillian Lee: Distributional clustering of English words. In: *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*. (1993) 183–190
- [5] Naftali Tishby and Fernando C. Pereira and William Bialek: The information bottleneck method. In: *Proceedings of the 37-th Annual Allerton Conference on Communication, Control and Computing*. (1999) 368–377
- [6] Ron Bekkerman and Ran El-Yaniv and Naftali Tishby and Yoad Winter: On feature distributional clustering for text categorization. In: *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, ACM (2001) 146–153
- [7] Noam Slonim and Naftali Tishby and Yoad Winter: Document clustering using word clusters via the information bottleneck method. In: *ACM SIGIR 2000*, ACM press (2000) 208–215
- [8] Re Termier and Marie-christine Rousset and Michle Sebag : Combining statistics and semantics for word and document clustering. In: *Ontology Learning Workshop, IJCAI01*. (2001) 49–54
- [9] Philip Resnik: Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence*. (1995) 448–453
- [10] J Jiang and D Conrath: Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In: *the 10th International Conference on Research in Computational Linguistics*. (1997)
- [11] Dekang Lin: An Information-Theoretic Definition of Similarity. In: *Proceedings of the 15th International Conference on Machine Learning*, Morgan Kaufmann (1998) 296–304
- [12] Alexander Budanitsky and Graeme Hirst: Evaluating WordNet-Based measures of semantic distance. *Computational Linguistics* **32**(1) (2006) 13–47

- [13] Artstein, Ron and Poesio, Massimo: Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics* **34**(4) (2008) 555–596
- [14] Rebecca J. Passonneau: Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In: *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy (May 2006)
- [15] George A. Miller and Richard Beckwith and Christiane Fellbaum and Derek Gross and Katherine Miller: *Introduction to WordNet: An On-line Lexical Database* (1993)
- [16] Sam Scott and Stan Matwin: Text Classification using WordNet Hypernyms. In: *Use of WordNet in Natural Language Processing Systems: Proceedings of the Conference*, pages 3844. Association for Computational Linguistics. (1998) 45–52
- [17] Leacock and Chodorow: Filling in a sparse training space for word sense identification (1994)
- [18] Z. Wu and M. Palmer: Verb semantics and lexical selection. In: *32nd Annual Meeting of the Association for Computational Linguistics*. (1994)
- [19] A. P. Dempster and N. M. Laird and M. Diftler and C. Lovchik and D. Magruder and F. Rehnmark: Maximum likelihood from incomplete data via the EM algorithm (1977)
- [20] Ian H. Witten and Eibe Frank: *Data Mining: Practical machine learning tools and techniques*. 2nd Edition edn. Morgan Kaufmann, San Francisco (2005)
- [21] John C. Platt: Sequential minimal optimization: A fast algorithm for training support vector machines. Technical report, *Advances in Kernel Methods - Support Vector Learning* (1998)
- [22] Tom Fawcett: ROC graphs: Notes and practical considerations for data mining researchers. Technical Report Tech report HPL-2003-4, HP Laboratories, Palo Alto, CA, USA (2003)