Evaluating the Utility of Appraisal Hierarchies as a Method for Sentiment Classification

Jeremy FLETCHER and Jon PATRICK
Sydney Language Technology Research Group
The University of Sydney
Sydney, Australia, 2006
{jeremy, jonpat}@it.usyd.edu.au

Abstract

Recent studies of sentiment classification (determining whether a text is “positive” or “negative”) using Appraisal theory have provided mixed results. While some good results have been obtained, it is difficult to tell what aspects of Appraisal are particularly useful for this task. In this paper, we present a series of experiments to isolate features of Appraisal, in order to compare which parts aid the task of sentiment classification on movie reviews. We report results which on the surface challenge the utility of Appraisal Hierarchies for this task, when modelled using systemic features. However in the context of making a trade-off between coverage and scale of feature space, our results appear promising. We hence discuss the need for a balance between the size of a classifier’s structure and the overall accuracy.

1. Introduction

Sentiment classification is a field of growing interest in the computational linguistics world, as researchers see the need for what has been termed non-topical text analysis. Sentiment classification deals with the problem of determining whether a document is *positive* or *negative*. This task has wide-ranging applications, notably market research, and customer feedback. This paper sets about to determine the usefulness of the linguistic theory of *Appraisal* for Sentiment Classification.

Appraisal theory describes how opinion is expressed in text. Its description is in the form of system networks denoted by a taxonomy of expressions. In this work we rely on the description of these taxonomies in Martin and White’s *The Language of Evaluation: Appraisal in English* (2005), for both our linguistically guided hierarchies and realisations of features[1]. Figure 1 shows a visualisation of the system network we use from appraisal theory.

Intuitively, it would seem that appraisal, if it could be modelled effectively using computational methods, would be a useful tool for sentiment analysis. A linguistic theory which gives us insight into the underlying construction of the opinion of the author of a piece of text should, in theory, allow us to compute such opinion more effectively. Previous work on using Appraisal for Sentiment Analysis, however, has been unconvincing and somewhat inconclusive.

In this paper, we set about trying to isolate the areas of appraisal theory which are useful and applicable to sentiment analysis. Subsequently, we wish to determine where efforts in the automatic extraction of a document’s appraisal profile should be focussed.

2. Previous Work

There has been some work done on the use of Appraisal for sentiment analysis, including the work of Taboada and Grieve (2004) in which different categories of product reviews were analysed for different types of Attitude (the three sub-systems being Affect, Judgment and Appreciation), using adjectives which had been assigned particular proportions of each of these systems.

Perhaps the most relevant work though is that of Whitelaw, Argamon and Garg (2005), in which movie reviews (data set from (Pang and Lee, 2004)) are classified over a positive/negative dichotomy, using what they term *appraisal groups*. The frequencies of expressions within a text which bear opinion in the appraisal groups are counted. These counts are normalised against the total counts of appraisal groups within the

[1] In this study, we use only the ATTITUDE system from Martin and White’s Appraisal structure, and append a simultaneous ORIENTATION system (cf. Whitelaw, Argamon and Garg, 2005). We omit the ENGAGEMENT and GRADUATION systems as they are not suited to the computational methods used in this study.

Proceedings of the Australasian Language Technology Workshop 2005, pages 134–142,
Sydney, Australia, December 2005.

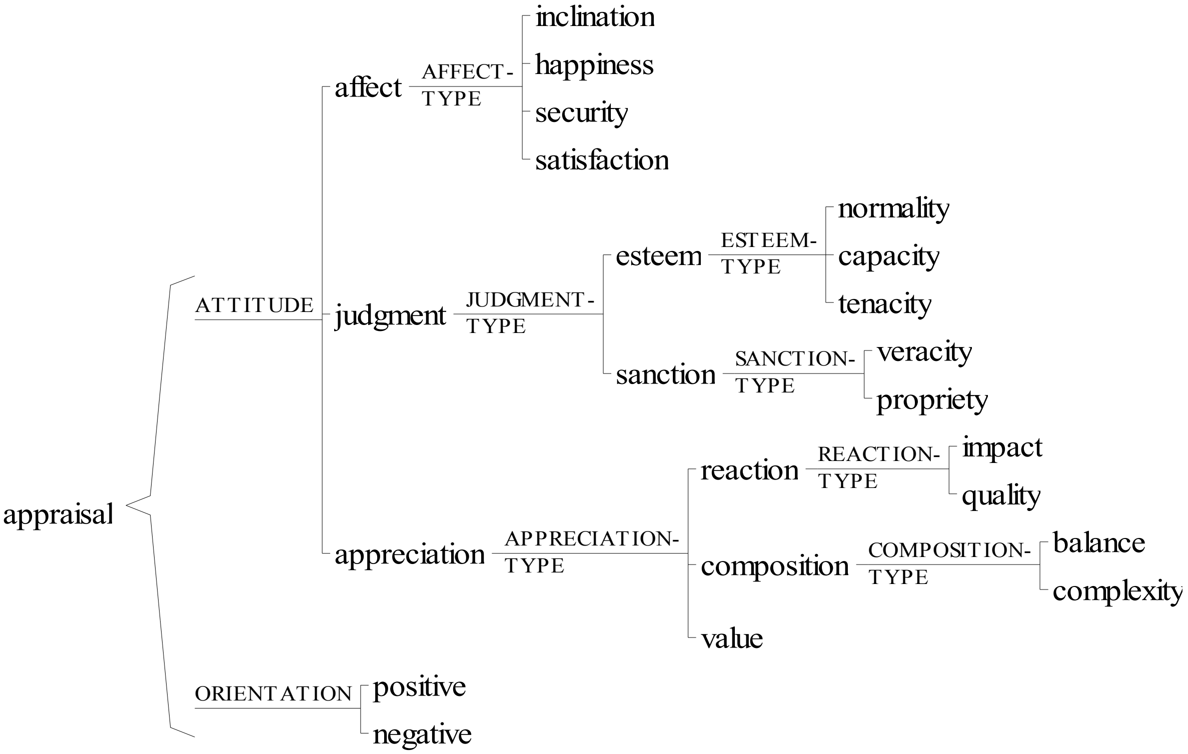

Figure 1: The Appraisal system network, comprising ATTITUDE and ORIENTATION sub-systems.

document.

They note, however, that work on similar taxonomies for other tasks such as the analysis of interpersonal distance (Whitelaw, Herke-Couchman and Patrick, 2004; Whitelaw and Patrick, 2004) and genre classification (Argamon and Dodick, 2004) use relative features within the hierarchies to model the “choice” made by the author about the way a particular structure is expressed.

Whitelaw and Patrick argue convincingly that this modelling of choice as it relates to meaning is an effective realisation of the tenets of the Systemic Functional Linguistic theory (see Halliday, 1994) which is the basis for the Appraisal model. Despite this, Whitelaw et al acknowledge that the use of this type of modelling of Appraisal for sentiment analysis gives inferior results to the simpler, non-theory conformant procedure they adopt.

3. Motivation

The results of Whitelaw et al using Appraisal for Sentiment Analysis were promising but unconvincing. Using their model of Appraisal theory, they were able to beat a baseline of simple bag-of-words analysis, and also improved on the then state of the art (Pang and Lee, 2004).

However, as we have already noted, to do this they removed the notion of modelling choice in the document, and used a simpler model of relative frequencies. This, then, raises questions about whether their Appraisal model does in fact match the true notion of Appraisal in Systemic Functional Linguistics.

Here, we adopt their methodology for populating the lexical realisations in the system network, in order to ascertain the areas of their use of Appraisal which are useful. However, in order to more closely model the linguistic phenomena of Appraisal, we revert to the use of relative systemic features (Whitelaw and Patrick, 2004).

We are hence attempting to approximate a computational method for linguistic analysis of Appraisal, in order to determine how useful such a method is for this task.

We distinguish between three key operations in the computational processing of system networks. Firstly, there is the system network design, in which the structure of the hierarchy is created (at this time, this process is a manual process, and the hierarchies used are those created by linguists). Secondly, there is the realisation of the system network, in which the concepts represented by nodes in the network are mapped to identifiable text features. And finally, there is the instantiation of a particular document as a representation of a system network. This final process involves some kind of abstraction of the text of a document using the realisations from the second process.

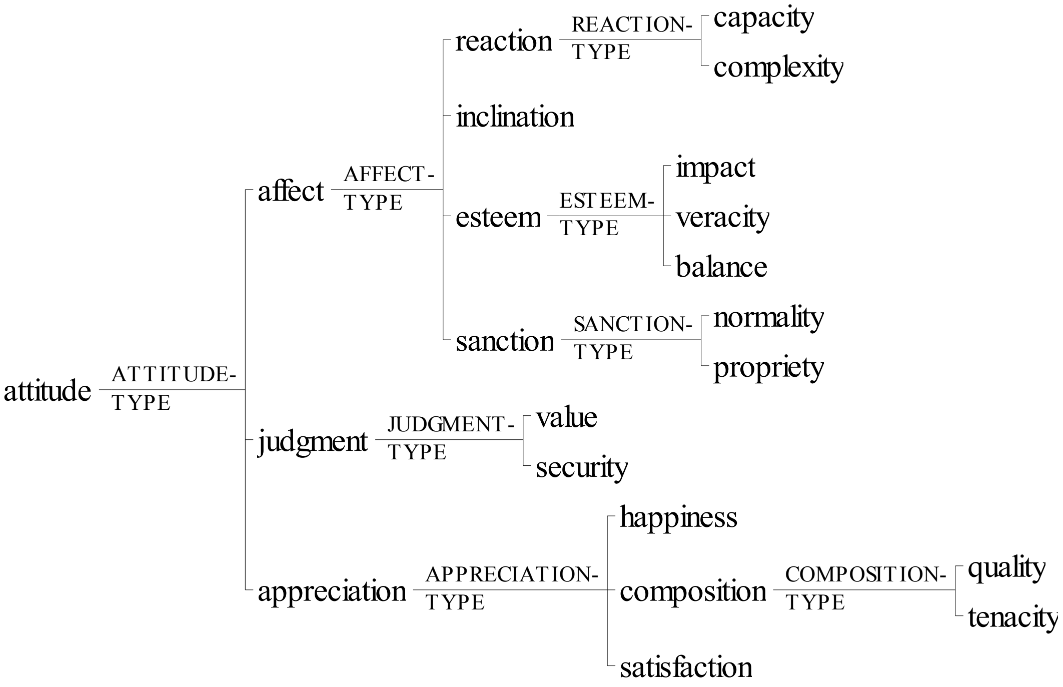

Figure 2: The ATTITUDE system from the *RelationshipsShuffled* system network. Nodes appear at the same depth as they did in the original tree (Figure 1), but their relationships to the next level are modified. (see note)

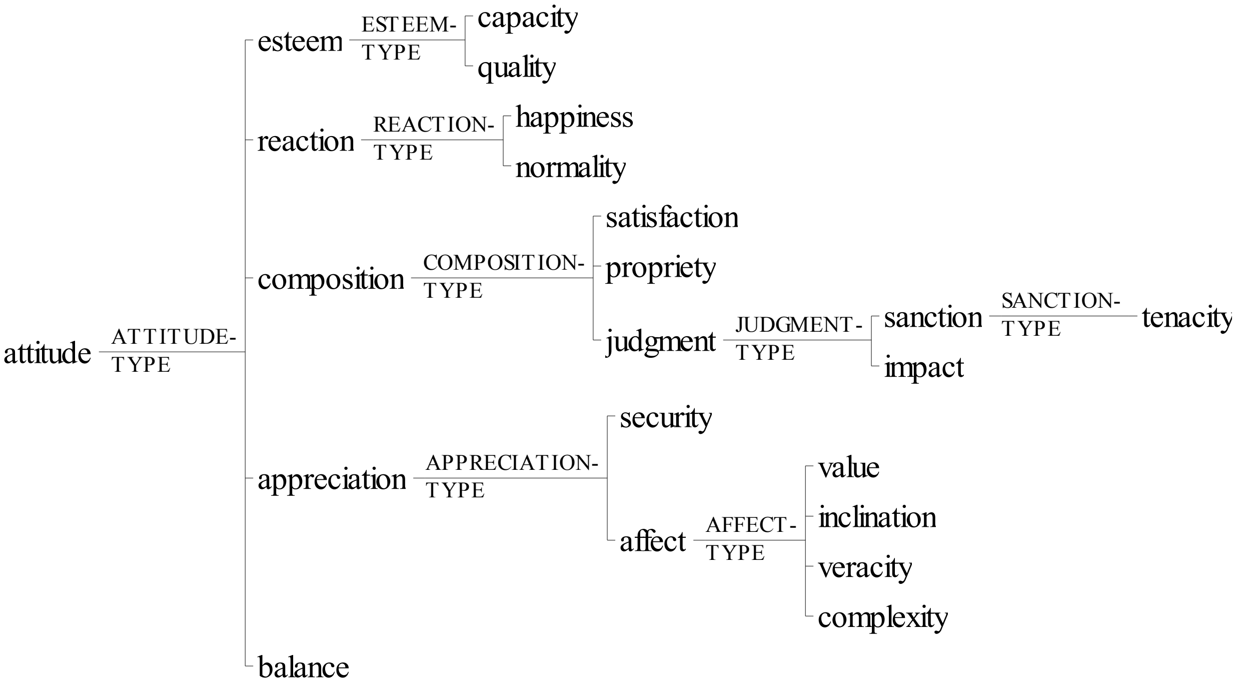

Figure 3: The ATTITUDE system from the *Hierarchy Shuffled* system network. Leaf nodes stay as leaf nodes, but other nodes are randomly assigned to places in the hierarchy. (see note)

NOTE: These systems each occur simultaneously with the ORIENTATION system within APPRAISAL. However, because of the nature of the shuffling, the ORIENTATION system remains the same in both cases.

Experiment	Raw Counts	Systemic Features	Original System Network	Relationship Shuffled (see Figure 2)	Hierarchy Shuffled (see Figure 3)	Attitude Realisations Shuffled	Orientation Realisations Shuffled
Baseline – ABOW	X						
Baseline – System Counts	X		X				
1		X	X				
2		X		X			
3		X			X		
4		X		X		X	X
5		X			X	X	X
6		X		X		X	
7		X			X	X	

Table 1: Feature types and networks used in the experiments

Our hypothesis is that there are particular elements of value in having the Appraisal in a document according to its conformance to the systemic network structure.

4. Realising the Appraisal System Network

In order to compute the appraisal profile of a document, we must be able to relate the content words in the document to the Appraisal system network of the theory.

The most common way of doing this is to attach to appropriate concept nodes in the tree a set of unigram features which are leaf-level "realisations". The system network is then instantiated for each document by counting all the realisation features within a document, and aggregating these counts up the tree.

However, one of the problems of this method is how to create a set of these realisations. For Appraisal, there are some small example texts from the Systemic Functional Linguistics literature, but not enough to allow for reasonable coverage within a computational framework.

To circumvent this lack of coverage of realisation, we took the example text from Martin and White (2005) as seed terms, and using the method of Whitelaw, Argamon and Garg (2005), expand the lexicon by generating synonyms from WordNet and two online thesauri[2]. From this, we also get a measure of the "confidence" of each expanded term, by counting the number of times a particular term is encountered from thesaural expansion in a particular node in the system network. For example, we may encounter "joyous" as a synonym of two different realisations of HAPPINESS ("happy" and "jubilant"), indicating that it is perhaps a stronger indication of that node than something which only occurs as a synonym once.

Note also that a particular unigram realisation may occur at numerous places within the system network. A particular unigram does not necessarily have a unique location within the system network. For example, the adjective "good" may be used in different contexts to realise SATISFACTION, PROPRIETY, QUALITY or VALUE, to name a few. Thus, each instance of "good" in a document increases the counts at each of these positions in the network.

While this method in no way guarantees complete coverage for the corpus, it does increase the coverage significantly, while still assuring Appraisal items can be identified computationally.

5. Experiments

We ran a set of experiments to classify Pang and Lee's (2004) movie review corpus as containing positive or negative sentiment[3].

In order to test our hypothesis, we developed a set of experiments to isolate particular attributes of the structure of the system network. In order to make the results comparable, we performed a process of randomising or shuffling the nodes in the network, thereby eliminating some of the linguistic information contained within the

[2] `http://m-w.com and http://thesaurus.reference.com`

[3] This dataset is freely available at `http://www.cs.cornell.edu/people/pabo/movie-review-data/`

particular representation of the tree.

There are two ways in which we randomise the network, and two levels of intensity with which we do it.

The first method of randomising the network involves keeping each node at the same depth as it was in the original, and simply randomly assigning a parent node to each, then continuing this process up the tree. The (shuffled) system network which results from this process is given in Figure 2 (*RelationshipsShuffled Network*).

The second method involves complete random assignment of nodes within the hierarchy. This means that any node can appear at any point within the hierarchy, with only the leaf nodes (which contain the realisations) staying as leaf nodes within the system. This system is shown in Figure 3 (*HierarchyShuffled Network*).

Of course, once this process has been executed, there is no longer a relationship between the labelling of particular nodes in the original and shuffled networks. For example, in Figure 2 AFFECT no longer encompasses HAPPINESS, SECURITY or SATISFACTION, and thus bears little resemblance to its function in the original network.

Running experiments using these shuffled networks and comparing the results to those we get on the original tree gives us some measure of the utility of the arrangement of systems within the original tree. If there is something of particular use to sentiment analysis that can be gleaned from the structure of the original tree, it should be reflected in the results on the different networks.

The second level of intensity involves shuffling the realisations between leaf nodes. Hence, word-level realisations are no longer grouped together as they were discovered in the process of thesaural expansion.

5.1 Methodology

We use the confidence measure we attain from the thesaural expansion to weight each realisation placed in the tree. That is, if a particular unigram has a high confidence measure for a particular node, the count value for a document will be increased more than if the confidence was low. This has two implications: firstly, words which are included as realisations from thesaural expansions of peripheral unigrams (and thus are less likely to be accurate realisations of appraisal) have little impact when found in a document, while increasing the impact of those unigrams which we are confident have some semantic similarity to our hand-crafted seed terms. Secondly, it means that individual unigrams are weighted for each node they realise. That is, a unigram may be a strong indicator of a particular node, but a weak indicator for another (due perhaps to having another, less common sense). This weighting then accounts for this case, rather than assigning the same weight for each realisation of each node.

We acknowledge that this confidence measure is a heuristic, and lacks manual crafting, but it increases the confidence about the decisions which have to be made in a computational process.

Once system network instance counts have been accumulated for a particular document, we calculate proportions of systems to their parents and siblings, using System Percentage (SYSPERC) and System Contribution (SYSCON) (Whitelaw, Herke-Couchman and Patrick, 2004; Whitelaw and Patrick, 2004).

SYSPERC: The proportion of the total system usage made up by this particular sub-system.

SYSCON: The proportion of a super-system's usage made up by a particular sub-system.

These features, once calculated were used as data for WEKA's (Witten and Frank, 1999) implementation of the SMO (Platt, 1998) support vector machine learning algorithm. We used a linear kernel and default parameters. Evaluation was done using 10-fold cross validation.

5.2 Experimental Results

We ran a series of experiments to evaluate the accuracy of classification of movie reviews created using the features of the linguistically modelled system network, and the same features throughoutt the shuffled system networks.

The results from the shuffled system networks were compared to the results on the hand-crafted hierarchies, as well as two baselines. Our experiments (summarised in Table 1) are as follows:

Baseline 1: *Appraisal-Bag-of-Words* (ABOW) – relative frequencies of all words which appear as realisations of systems in the Appraisal system network, normalised by document length. Omitted from the experiment are the realisations which do not appear in any document in the corpus, leaving 4,381 features.

Baseline 2: *Bag of Nodes* – the relative frequencies of the raw counts of each node in the system hierarchy, normalised by the total number of appraisal counts in the document (i.e. the aggregated count at the root of the hierarchy)

Experiment 1: SYSPERC and SYSCON measures at all levels in the hierarchy, using the original linguistically created system network.

Experiment 2: SysPerc and SysCon measures at all levels in the hierarchy, using the *RelationshipsShuffled* system network.

Experiment 3: SysPerc and SysCon measures at all levels in the hierarchy, using the *HierarchyShuffled* system network.

Experiment 4: SysPerc and SysCon measures at all levels in the hierarchy, using the *RelationshipsShuffled* system network, and realisations randomly assigned in both the Attitude and Orientation networks.

Experiment 5: SysPerc and SysCon measures at all levels in the hierarchy, using the *HierarchyShuffled* system network, and realisations randomly assigned in both the Attitude and Orientation networks.

Experiment 6: SysPerc and SysCon measures at all levels in the hierarchy, using the *RelationshipsShuffled* system network, and realisations randomly assigned in just the Attitude network.

Experiment 7: SysPerc and SysCon measures at all levels in the hierarchy, using the *HierarchyShuffled* system network, and realisations randomly assigned in just the Attitude network.

The results of these experiments are shown in Table 2.

Experiment	**Acc. (%)**
Baseline (ABOW)	83.7
Baseline (System counts)	71.8
1	72.4
2	72.6
3	72.8
4	67.8
5	69.5
6	68.5
7	70.4

Table 2: 10-fold cross validation results for different system networks and feature set configurations. (See Table 1 for details of experiments)

6. Analysis of Results

Immediately apparent from these results is the degradation of accuracy when you move from the Appraisal-bag-of-words features to systemic features. This mimics the results of Whitelaw et al who report that the use of these systemic features produces inferior results to their simpler measures. The most likely cause for this discrepancy is the fact that the Appraisal tree is reasonably shallow, so the aggregative properties of these features do not have the scope of previous experiments on these networks.

Occam's Razor tells us to "not multiply entities without necessity" and BOW classifiers rampantly ignore this economy argument. Our real objective should be to produce the classifier that attains the highest accuracy with the least model complexity, and to this end we need to devise new metrics of performance that balance performance against classifier size. In this light, the "efficiency" of the Appraisal model can be seen as superior to BOW. Of course, the complexity of a classification system relies on more than feature set size.

Our objective with this set of experiments, however, is to draw comparisons between the results of the linguistically created network and our shuffled hierarchies.

What we note about these results, is that there is very little difference between whether the hierarchy used is the linguistically created network, or one of those which was randomised to some degree. We can see that the accuracy on our original tree is 72.4%, whereas the results of the same feature set using our *RelationshipsShuffled* and *HierarchyShuffled* trees were 72.6% and 72.8% respectively. This leads us to believe that there is no advantage for sentiment analysis in the use of the structure of the original Appraisal network when modelled computationally in the manner we have described[4].

However, what we do notice is the distinct drop in accuracy once the realisations are randomly assigned to the leaf nodes in our hierarchy. Our accuracies drop by approximately 5% once this shuffling of realisations has taken place.

Given that the results above show no benefit in the structure of the hierarchy, we can deduce that the benefit comes from having our unigram realisations grouped together in some semantic categories.

Experiments 6 and 7 attempted to isolate the shuffling of realisations within the Orientation network, as we felt that this decrease in accuracy may be due simply to the fact that each realisation (in experiments 1-5) had been assigned either a "positive" or "negative" Orientation value. This type of processing of Semantic Orientation (SO) has been exploited for sentiment classification previously (Hatzivassiloglou and McKeown, 1997; Turney, 2002).

[4] The small *increases* in accuracy over experiments 2 and 3 are most probably not statistically significant.

However, leaving the ORIENTATION realisations unshuffled produces only a minor increase in accuracy (Exp 4/5 v Exp 6/7), and the results of these experiments are still well below the results on those where the ATTITUDE realisations are also unshuffled (Exp 2/3 v Exp 6/7). This indicates that it is not only the Semantic Orientation of our realisations which aid classification, but also the categories of ATTITUDE.

Despite this, when we compare the results achieved on this type of analysis to the simple Appraisal-bag-of-words classification, there is a very marked decrease in accuracy.

Most probably this is due to the additional granularity which can be achieved by looking at words on an individual level. What is important to note is that although in the ABOW experiment there is no preordained measure of sentiment attached to the words, the machine learner distinguishes words which have an intrinsic positive or negative connotation some of which are "bad", "mess", "waste", "worst", "stupid", along with "fun", "great", "terrific", "memorable" and "hilarious". Perhaps more interesting is that some word features which do not intrinsically contain a semantic orientation become strong word features in the ABOW experiments. Words such as "very", "also", "nowadays", "many" and "leave" are indicators of positive sentiment, and words such as "only", "have", "work", "plot" and "intended" seem to indicate negative sentiment.

This indicates that there is perhaps some value to analysing the structure of the text, and how rhetorical structure is realised differently in positive and negative reviews. Another reason for these strong word features is perhaps their collocation with other sentiment-bearing expressions. In this case, a process for identifying frequent collocations in the text may also be a useful tool for identifying better sentiment-bearing expressions, as well as increasing the number of realisations. This acknowledges the need for more complex realisations of the system network.

The peculiarities of particular words being indicators of a particular orientation of sentiment are worth exploring. For example, the fact that "plot" tends to be indicative of negative sentiment suggests that those movie reviews which make specific reference to the plot are more likely to be negative. This raises questions about different styles of reviewing; are there ways to extract information about how hard or leniently a reviewer gives his or her opinion? When dealing with a style of text which is opinion heavy, especially when resolving the opinion into a positive/negative dichotomy, issues of review style come into effect.

In fact, one of the largest problems with the use of these hierarchies, and perhaps the reason why accuracy using them is less than with ABOW features, is simply the lack of coverage. While we have isolated potential sites for sentiment within a text by collecting and expanding lists of Appraisal realisations, it is reasonable to expect that there are many more which are not captured. Moreover, those Appraisal expressions which we do have form only small proportion of the text as a whole[5].

Experiment	**# *features***	**Acc. (%)**
W:A[6]	1047	77.6%
ABOW	4318	83.7%
BOW[6]	48,314	87.0%

Table 3: Comparison of percentage accuracy and size of feature set.

Although the Appraisal results generally and hierarchical use in particular do not appear to be competitive this is not the whole story. An investigation of the size of the classifying indicates a strong efficiency in the Appraisal classifier. Using full bag-of-words features on the same set of documents, Whitelaw et al attain accuracy of 87.0%. However, to attain this accuracy, they used 48314 features, whereas our results on the Appraisal-bag-of-words (83.7%) are attained through the use of only 4381 features, less than ten percent of the size (see Table 3). To draw direct comparisons between the accuracies of these feature sets, however, it would be necessary to compare the different types of selected features over similar set sizes. How effective is the use of the 50 best appraisal features compared to the 50 best unigram BOW features? These questions are open for discussion.

Whitelaw et al's result on a similar feature set to our Appraisal-bag-of-words, using 1047 features is 77.6% accuracy. This continuum relationship between size of feature set and accuracy once again emphasises the need for some balance in a real working system. One of the problems which must be addressed within any type of text classification system is our ability to reach the accuracy of feature indiscriminate classifiers using classifiers with a much smaller internal structure.

[5] Even with the thesaural expansion, the Appraisal realisations only make up ~22% of the unique words within the text.

[6] These experiments are documented in Whitelaw, Argamon and Garg (2005)

7. Discussion and Future Work

We have presented here a set of experiments designed to test the utility of particular parts of Appraisal theory for sentiment analysis. While we are still far from a definitive answer as to whether this type of processing is useful in this domain, the results here show that there is little benefit to be gained from structure of the Appraisal network. Perhaps one reason for this is that the network itself is quite shallow. The deepest node in the hierarchy is only four links from the root, indicating that the type of aggregative statistics gained from System Contribution and System Percentage are ill-suited to this particular network.

Furthermore, it is also possible that we are not yet able to effectively approximate a model of Appraisal theory using computational methods. Linguists in particular would argue that our Appraisal-as-realisations methodology does not do justice to the complexities of the theory.

As we discussed, there is some merit in the level of distinction gained by the machine learner on the Appraisal-bag-of-words features. In actuality, in the process, the learner discovers patterns between the word-level realisations and the sentiment of the document as a whole.

A potentially useful extrapolation of this principle is having a machine learner craft an "optimal" hierarchy for Appraisal, for a particular task. While the Appraisal hierarchy we see in the linguistics literature is useful for general descriptions of linguistic phenomena, it is probably true that modifications to suit a particular task could amplify the delineation of some aspect of the text (for example sentiment analysis), thereby increasing the accuracy of computational processing.

Another method of testing the applicability of this theory to the computational process of sentiment analysis is to use the Appraisal network for a different type of text classification (for example, topic classification). If, in fact, there is a notable decrease in the accuracy of the Appraisal model on non-emotive texts, then we can see that there is a particular relationship between Appraisal theory and the computational process of sentiment analysis. However, if the only real utility provided by the network is some kind of smoothing process, or a benefit from the aggregative properties of the hierarchy, which would perhaps be attested by similar results on emotive and non-emotive texts, then we can no doubt create (or construct using automatic methods) better hierarchies for doing this.

Overall, there is still some question over whether Appraisal theory is useful for the computational process of sentiment classification. The results here suggest that there is some value to be gained from the grouping of words into Appraisal clusters, but it is also true that only doing this decreases the accuracy over using the words themselves.

However, before a definitive answer is given to this question, we would need to assess our computational model of Appraisal in mode detail. There is obviously some level of uncertainty in the model, due to the method of realising nodes in the system network. It is also likely that the instantiations of the networks need to be modelled in richer ways.

There is ongoing research into the way realisations of linguistic phenomena are modelled computationally, and is important that other methods of such realisation are explored before the use of Appraisal for sentiment classification is discarded.

On the criteria of efficiency, however, the Appraisal model appears to work very well, although a true comparison has not been achieved in this paper. That would require comparison to a result from the top 4318 features in a BOW experiment excluding the Appraisal terms.

References

S. Argamon and J. T. Dodick. 2004. "Conjunction and modal assessment in genre classification". In *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text*. AAAI.

Michael A. K. Halliday. 1994. *Introduction to Functional Grammar*. Edward Arnold.

V. Hatzivassiloglou and K. R. McKeown. 1997. *Predicting the semantic orientation of adjectives*. In "Proceedings of the 35th ACL and 8th EACL", pages 174-181, Somerset, New Jersey. ACL.

J. R. Martin and P. R. R. White. 2005. *The Language of Evaluation: Appraisal in English*. Palgrave, London.

Bo Pang and Lillian Lee. 2004. "A Sentimental education: Sentiment analysis using subjectivity summarization base on minimum cuts". In *Proc. 42nd ACL*,. Pages 271-278. Barcelona, Spain.

M. Taboada and J. Grieve. 2004. "Analysing appraisal automatically". In *AAAI Sprint Symposium on Exploring Attitude and Affect in Text*. AAAI.

Peter D. Turney. 2002. "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews". In *Proc. 40th ACL,* pages 417-424, Philadelphia, Pennsylvania.

Casey Whitelaw, Shlomo Argamon and Navendu Garg. 2005. Using Appraisal Taxonomies for Sentiment Analysis. In *Proceedings of the First Computational Systemic Functional Grammar Conference*, University of Sydney, Sydney, Australia.

Casey Whitelaw, Maria Herke-Couchman and Jon Patrick. 2004. “Identifying Interpersonal distance using systemic features”. In *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text.*

Casey Whitelaw and Jon Patrick. 2004. “Selecting Systemic Features for Text Classification”. In *Proc. Australasian Language Technology Workshop, 2004*. Macquarie University, Sydney Australia.

Janyce M. Wiebe, Theresa Wilson, Rebecca F. Bruce, Matthew Bell and Melanie Martin. 2004. “Learning Subjective Language”. In *Computational Linguistics 30(3),* pages 277-308. MIT Press.

Ian H. Witten and Eibe Frank. 1999. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann.