Weakly-supervised appraisal analysis

Article (Published Version)

Linguistic **I**ssues in **L**anguage **T**echnology – **LiLT**
Volume 8, Issue 2 October 2012

Weakly-supervised Appraisal Analysis

Jonathon Read

John Carroll

Published by CSLI Publications

LiLT volume 8, issue 2 October 2012

Weakly-supervised Appraisal Analysis

Jonathon Read, *University of Oslo*
John Carroll, *University of Sussex*

Abstract

This article is concerned with the computational treatment of Appraisal, a Systemic Functional Linguistic theory of the types of language employed to communicate opinion in English. The theory considers aspects such as Attitude (how writers communicate their point of view), Engagement (how writers align themselves with respect to the opinions of others) and Graduation (how writers amplify or diminish their attitudes and engagements). To analyse text according to the theory we employ a weakly-supervised approach to text classification, which involves comparing the similarity of words with prototypical examples of classes. We evaluate the method's performance using a collection of book reviews annotated according to the Appraisal theory.

1 Introduction

The proliferation of electronically-published text presents a wealth of content of interest to organisations seeking to distill public opinion. This article considers how such text might be analysed according to Appraisal (Martin and White, 2005), a theory of evaluation in text. Appraisal was developed in the tradition of Systemic Functional Linguistics (Halliday, 1994), in which the choices writers make in deciding

LiLT Volume 8, Issue 2, October 2012.
Weakly-supervised Appraisal Analysis.

how to convey meaning are represented in a typology. The full Appraisal typology is depicted in Figure 1. Read and Carroll (2012) present an overview; for a complete description see Martin and White (2005).

Appraisal consists of three subsystems that operative interactively. ATTITUDE describes three types of private state: emotion, ethics and aesthetics. It is qualified with a polarity—either positive (+) or negative (−). The example below contains instances of QUALITY (appreciation of the writing under review) and TENACITY (judgments of an author's reliability):

> This is clever$^{+\text{QUALITY}}$ stuff that bears the hallmarks of careful$^{+\text{TENACITY}}$ thought but$^{\text{COUNTER}}$ rather$^{\uparrow\text{DEGREE}}$ less$^{\downarrow\text{DEGREE}}$ careful$^{+\text{TENACITY}}$ execution.

The example also contains an instance of COUNTER, a class of the second subsystem, ENGAGEMENT, which considers the lexical choices used by authors to convey their point of view and to agree/disagree with the opinions of others. This particular instance serves to counter the initial statement. The final subsystem, GRADUATION, serves to amplify or diminish the attitude or engagement conveyed by an expression. Graduating items are qualified with a direction—either up-scaling (↑) or down-scaling (↓). In the example, *rather* serves to amplify *less*, which results in a strong diminishment of the expressed attitude.

Appraisal is of interest to the sentiment analysis and opinion mining communities due to its comprehensive typology of evaluation-bearing words, its consideration of how writers report the opinions of other people, and its analysis of the types of language employed to modify the strength of evaluation. Such information can inform existing approaches to sentiment analysis (Whitelaw et al., 2005), and augment opinion mining with an informative classification of the type of opinions expressed (Bloom et al., 2007).

The remainder of this article is structured as follows. Section 2 presents our general method for weakly-supervised text classification, which involves estimating the similarity of a word with respect to prototypical examples of classes, using either lexical association or semantic spaces. Section 3 then presents several experiments that evaluate the method's performance in discriminating between various aspects of the Appraisal theory. The shortcomings and successes of the approach are presented in Section 4. We then review other computational treatments of Appraisal in Section 5, before offering conclusions and directions for future work in Section 6.

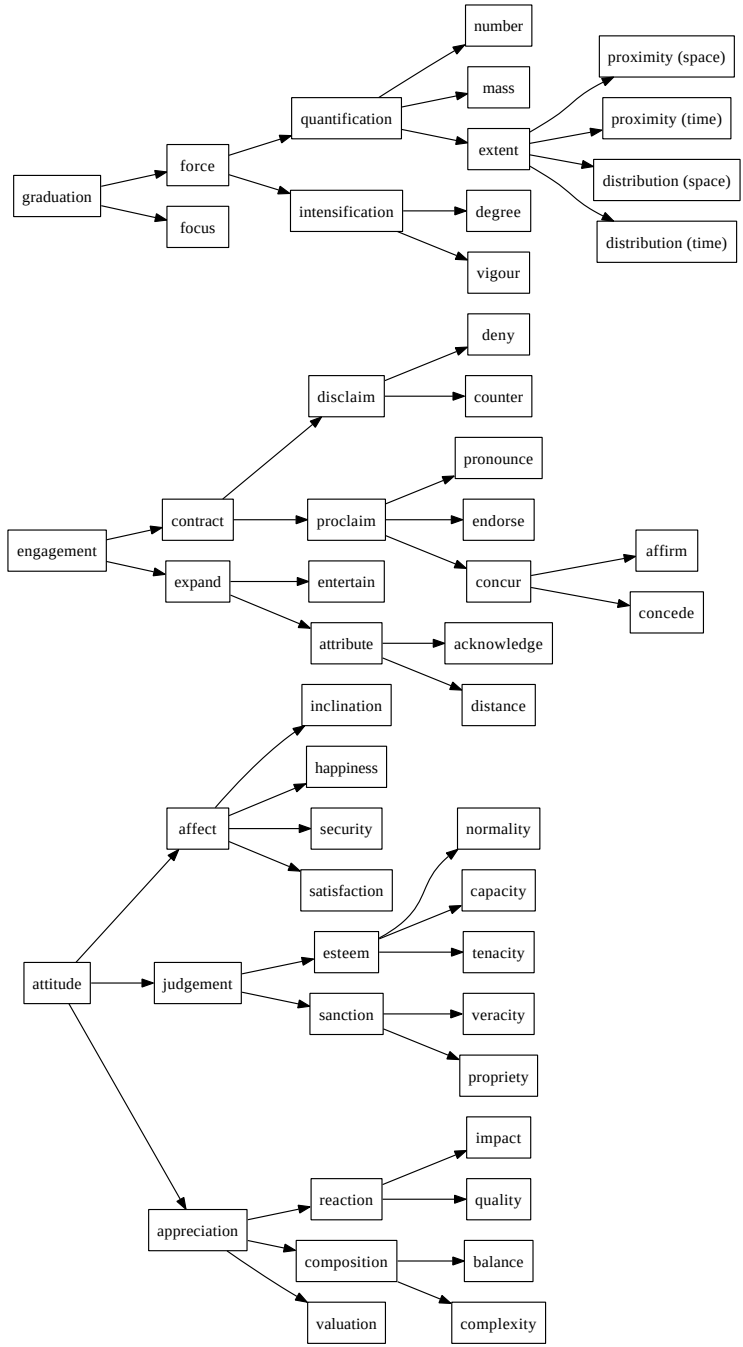

FIGURE 1 The hierarchy of types described by Appraisal Theory.

2 Weakly-supervised Text Classification

Our approach to classifying text according to Appraisal is based on Turney's (2002) notion that the sentiment of text can be determined by estimating the similarity of its words with prototypical examples of positive and negative sentiment. However, Turney's method (SO-PMI-IR) determined the sentiment of phrases using pointwise mutual information, whereas the method we employ is generalised to use other techniques for determining similarity, and is applicable to tasks with several classes.

The method involves choosing the maximal scoring class, c^*, from a set of classes, C, for a vector of words, W:

$$c^* = \arg\max_{c \in \mathbf{C}} score\left(\mathbf{W}, c\right) \tag{1.1}$$

where the score is calculated as the sum of each word's mean similarity with a set of prototypical words of a class, c_P:

$$score\left(\mathbf{W}, c\right) = \gamma_c \sum_{w \in \mathbf{W}} \frac{\sum_{p \in c_P} similarity\left(w, P\right)}{|c_P|} \tag{1.2}$$

where γ is a vector of weights indicating the preference for each class, and $similarity\left(w_1, w_2\right)$ is a function that estimates the semantic similarity of two words. We consider two methods for estimating the similarity of words: lexical association and semantic spaces, which are effective in dealing with subjective language (Read and Carroll, 2009).

Measures of lexical association examine first-order similarity between words (Grefenstette, 1994). That is, they determine the similarity of a pair of words by considering how likely they are to occur near each other. There are many measures of lexical association, including various likelihood measures and hypothesis tests (Evert, 2004). Following Turney (2002), we employ pointwise mutual information to measure lexical association. It is defined as:

$$I\left(w_1, w_2\right) = \log_2 \frac{P\left(w_1, w_2\right)}{P\left(w_1\right) P\left(w_2\right)} \tag{1.3}$$

While Lexical Association measures first-order similarity, Semantic Spaces measure second-order similarity. Words that are similar in the second order may not necessarily co-occur, but rather occur in similar contexts (Grefenstette, 1994).

Semantic spaces represent concepts as a set of points in a large number of dimensions. The location of each point along each dimension is measurement of the strength of association with that axis. The development of semantic spaces began in the field of cognitive science, where they were constructed by defining scales of interest and having several

human subjects specify the position of each concept on each scale (Osgood et al., 1957). A semantic space can therefore be thought of as a cuboid of data with k concepts $\times$ m dimensions $\times$ n subjects. Osgood et al. proposed that the meaning of a concept can be represented by collapsing the cuboid along the subject dimension. It is then possible to assess the similarity of concepts by applying a distance metric on the vectors extracted from the resulting matrix.

However, constructing a semantic space using human subjects is clearly a labourious task, and the allocation of dimensions relies on the intuitions of the researchers. As an alternative, Lund and Burgess (1996) describe Hyperspace Analogue to Language (HAL), which selects dimensions from the vocabulary found in a corpus, and populates the semantic space matrix with cooccurrence counts (where cooccurrence is defined with respect to some window of n words). Lund and Burgess found that performing distance measurements on semantic spaces built in this manner can: (1) determine nearest neighbours of words, (2) classify words according to a shared hypernym and (3) produce similarity scores that correlate with human reaction times in a lexical priming study.

Lowe (2001) formalised the notion of semantic space as a quadruple, $\langle \mathtt{A,B,S,M} \rangle$. $\mathtt{B}$ is a set of $b_{1..D}$ basis elements (where D is the number of dimensions in the space). $\mathtt{B}$ can be a set of word types, stems, n-grams, or indeed any reasonable feature of text. $\mathtt{A}$ is a lexical association function that maps co-occurrence frequencies of a target word, t, with basis elements so that t is represented as a vector $v = [\mathtt{A}(b_1, t), \mathtt{A}(b_2, t), ..., \mathtt{A}(b_D, t)]$. $\mathtt{A}$ is often defined simply as the identify function. Lowe (2001), however, asserted that this is unsatisfactory as raw co-occurrence counts can result in a frequency bias. Instead, lexical association measures such as the odds-ratio, pointwise mutual information and the log-likelihood ratio are preferable. $\mathtt{S}$ is a similarity measure that maps pairs of vectors v and w onto a value that represents their contextual similarity. Suitable measures include euclidean, city block, and cosine. In our application of the semantic space technique we used cosine, as it conveniently maps to a value between -1 and 1, and accounts for any random scaling effects that might be caused by the range of the lexical association function or the number of basis elements (Lowe, 2001). It is defined (Levy et al., 1998) as:

$$cosine(v, w) = 1 - \frac{\sum_{b \in \mathtt{B}} v_b w_b}{\sqrt{\sum_{b \in \mathtt{B}} {v_b}^2} \sqrt{\sum_{b \in \mathtt{B}} {w_b}^2}} \tag{1.4}$$

Finally, $\mathtt{M}$ is a mapping of one semantic space onto another. A semantic space is functional without $\mathtt{M}$, but transformations such as Latent Se-

mantic Analysis (Landauer and Dumais, 1997) can build a more structured model.

3 Experiments

This section presents various experiments that evaluate the performance of the weakly-supervised method for text classification in classifying words according to aspects of Appraisal. Section 3.1 describes the data employed for evaluation and our experimental setup is presented in Section 3.2. The results of a number of tasks are then presented: classifying expressions of Appraisal (Section 3.3); extracting expressions of Appraisal (Section 3.4); determining the polarity of ATTITUDE (Section 3.5); and determining the direction of GRADUATION (Section 3.6).

3.1 Data

We use the corpus created by Read and Carroll (2012) during the course of an annotation study of Appraisal, which assessed the inter-annotator agreement exhibited by two coders when labelling a corpus of thirty-eight book reviews (approximately 37,000 words) with instances of the types of Appraisal Theory. The degree of agreement observed varied greatly depending on the level of abstraction in the Appraisal hierarchy (a mean F-score of 0.698 at the most abstract level, and 0.395 at the most concrete level).

Low agreement is perhaps to be expected—as noted by Wiebe et al. (2004), interpretation of subjective language is itself subjective. This is well demonstrated by the following example of disagreement between the annotators:

(a) Like him, Vermeer—or so he chose to believe—was an artist neglected$^{-\text{SATISFACTION}}$ and wronged$^{-\text{SATISFACTION}}$ by critics ...

(b) Like him, Vermeer—or so he chose to believe—was an artist neglected and wronged$^{-\text{PROPRIETY}}$ by critics ...

Annotator (a) interprets the sentence as representing the artist's emotion response to the critics, whereas (b) reads the sentence as a negative judgment of the actions of the critics. Such examples are typical of the disagreement between the annotators. However, satisfactorily resolving these would require the decisions of further annotators, and is thus left to future work. Despite this low agreement, the intersection of both annotator's decisions contains 2,223 annotations that may be employed to assess computational methods for the identification of Appraisal—though we acknowledge that this perhaps represents a subset of Appraisal annotations that are inherently less ambiguous.

We created two sets of labelled expressions from the Appraisal cor-

pus: test data from the intersection of the two annotator's selection, and development data from the remainder of the annotators' selections (the symmetric difference). Each set contained several single- and multi-word expressions, a label describing its Appraisal type and its location in the source text.

3.2 Experimental Setup

The weakly-supervised method for word labelling described above requires a set of prototypical examples for each class. Prototypes for the Appraisal types come directly from Martin and White's (2005) examples and are listed in the appendix. However, it was not possible to obtain prototypes for the graduation class of VIGOUR. This class is reserved for instances where lexical choice intensifies appraisal (e.g. *this worried me* versus *this terrified me*). In order to consider the effects of vigour it is necessary to rank semantically-related lexical items according to their strength. This is a particularly challenging task—e.g. see Wilson et al. (2006)—which would require a distinct approach and is thus not addressed in this article.

A second prerequisite of both techniques for word similarity described above is a source of word frequencies for probability estimates. Lexical association methods are generally computationally inexpensive, and so have been applied on very large corpora (Turney and Littman, 2003). The semantic spaces method is significantly more computationally intensive, and so mostly has only been used with medium-sized corpora. However, experimental results indicate that using as large a training corpus as possible will yield better results, so in these experiments we sampled word occurrence and cooccurrence frequencies from the largest corpus available at the time the work was carried out, the English Gigaword corpus (Graff, 2003), a collection of newswire articles published by four international news agencies which contains around 1.7 billion words.

We consider two versions of our approach for both similarity methods. In the first, all values in γ (the set of class weights in Equation 1.2) are set to 1 (i.e. unweighted). In the second version, the values are set to the relative frequency of the class (as observed in the development data), effectively applying a class prior. Henceforth this version is indicated with a '(w)' suffix.

We investigated the following feature types (or basis elements, under Lowe's formalisation): lemmatised words, lemmatised words with part-of-speech tags, and adjectives/adverbs only. We also varied the number of features used by the techniques. While previous studies have examined this aspect of semantic spaces (Lowe and McDonald, 2000), the

size of the feature set greatly affects the runtime of the semantic spaces technique, and thus deserves evaluation in the context of this application. We used a logarithmic scale of feature set sizes with the n most frequently occurring features selected (n = 5000, 10000, 20000, 40000, 80000, 160000 and 320000). As PMI is unreliable when applied to low-frequency words (Church and Hanks, 1990), it was not calculated for words occurring less than five times in the corpus. This constraint was carried over to the semantic spaces method in order to maintain comparability between the techniques.

Levy et al. (1998) found that the optimal cooccurrence window size varied depending on the application, but tended to peak at 10 words either side of a target word. Importantly though, performance dropped off very quickly after the peak. It is therefore crucial not to use too large a window. The experiments here used a context window of m words on each side of the target word, with m varying from 1 to 10.

Performance is measured in terms of Precision (P, the proportion of correct identifications relative to the total number of identifications), Recall (R, the proportion of correct identifications relative to the total number of possible identifications) and F-measure (F_1, the harmonic mean of precision and recall). Our optimisation procedure involved assessing all parameter combinations on the development data, and selecting the configuration that yielded the greatest F_1. Statistical significance was estimated using a paired t-test over sets that represent the outcome of a test for a particular system with a value of 1 if the method chose the correct classification and 0 otherwise.

3.3 Classifying Expressions of Appraisal

The first task assessed is that of classifying expressions of Appraisal taken out of context, in order to determine whether using the weakly-supervised techniques for the discrimination of Appraisal types is practical. We optimised the parameters using the development data set, by finding the best F_1 when classifying expressions according to most concrete level of the typology. The lexical association method achieved its highest performance using 20,000 plain lemmas with a window of 10 features. The semantic space method performed best with 320,000 plain lemmas with a window of 2 features.

The evaluation utilised these optimal parameters to determine performance on the test data set. We evaluate performance at each level of the Appraisal hierarchy (see Figure 1, in which the members of levels are aligned in columns). When classifying at abstract levels of the hierarchy we obtain prototypes by unifying the prototypes of all child types. Figure 2 shows the F_1 of each method at each level of the hi-

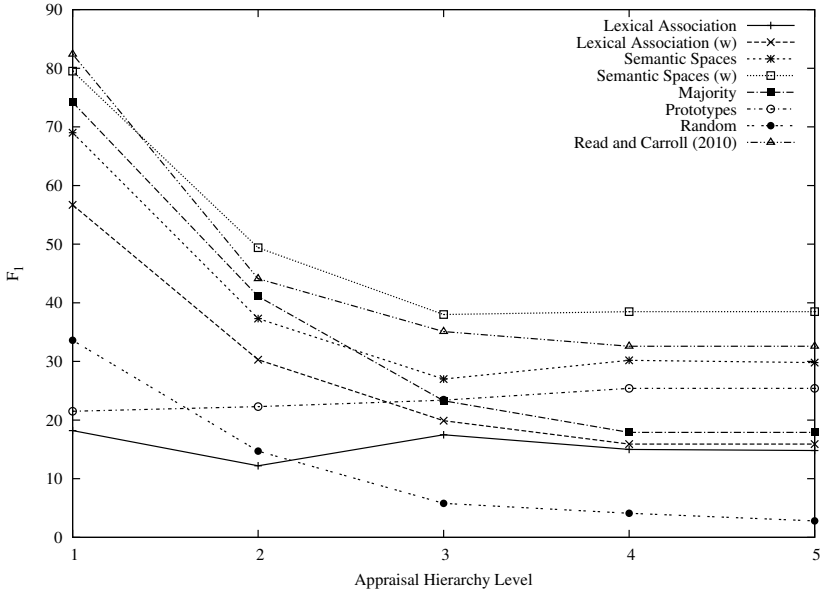

FIGURE 2 The classification performance at the various levels of the Appraisal hierarchy. (w) indicates weighted versions of the methods.

erarchy, along with three baselines: choosing the majority class in the development set, counting the presences of prototypes, and selecting a class at random. The figure also plots the results of a support vector machine classifier at each level, as reported by Read and Carroll (2012). Table 1 lists the precision, recall and F_1 of each method as means taken from all levels of the hierarchy. The results indicate that the weighted semantic space method performed best across all levels of the hierarchy, having a mean F_1 of 48.79. This result compares favourably to supervised classification; the mean F_1 of a support vector machine classifier was 45.40 (Read and Carroll, 2012).

The differences between the lexical association and prototypes methods are not significant at levels 4 and 5 of the hierarchy. Furthermore, while overall the prototype method performs better, at levels 1 and 2 it is outperformed by the lexical association method. These observations suggest that the lexical association method is not effective when the task involves the discrimination of a large number of Appraisal types.

Method	**Prec**	**Rec**	F_1
Semantic Spaces (w)	49.75	47.86	48.79
Semantic Spaces	39.43	37.94	38.67
Majority Baseline	34.88	34.88	34.88
Lexical Association (w)	30.55	25.38	27.72
Prototypes Baseline	87.05	13.65	23.58
Lexical Association	17.14	14.24	15.55

TABLE 1 Classifying expressions of Appraisal

3.4 Extracting Appraisal words

In order to investigate how well the methods can identify Appraisal in free text we extended the basic method with an additional parameter, a threshold t. Each single word is only labelled with an Appraisal type if the score for that word is above the threshold. The test data for this experiment again comes from the intersection of the annotators' selections in the Appraisal corpus. However, in this experiment it is not viable to create development data from the symmetric difference of the sets. The experiment involves extracting from free text, so all words in the corpus are included in the experimental data (be they annotated or otherwise). Optimising on a set comprised of the symmetric difference would result in the consideration of words that, while unlabelled in the development set, would be labelled in the test set (and vice-versa). Thus, any optimisation procedure performed on the symmetric difference of the annotators' selections would result in erroneous parameters.

Instead, we conducted a cross-validation evaluation procedure, dividing the test data set into twelve folds, each of three texts. The threshold, t, started at zero and increased in increments of 0.0005 until the resultant recall was zero. Taking the average precision, recall and F_1 across all folds provided an evaluation for each level. The best performing configurations for Appraisal extraction across all folds were: 320,000 lemmas and a context window of 2 features for the semantic space method, and 80,000 lemmas and a context window of 1 feature for the lexical association method. An analysis of variance indicated that the optimised threshold was very consistent across all folds.

Table 2 lists the mean extraction performance across levels with the majority, prototype and random baselines, while Figure 3 shows the change in performance at each hierarchical level. The weighted semantic space method is the best performer (F_1= 13.3). However, the

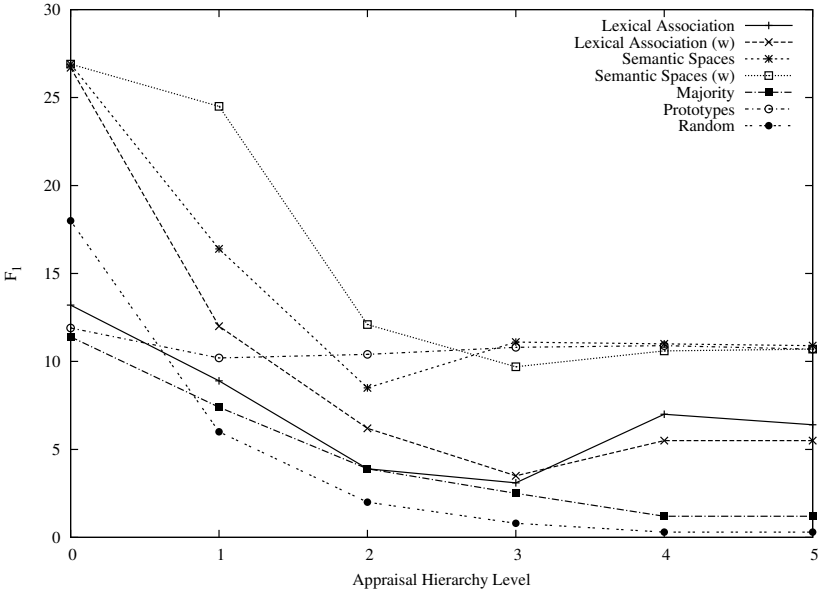

FIGURE 3 Extracting and labelling performance at the various levels of the Appraisal hierarchy. (w) indicates weighted versions of the methods.

difference between prototypes, lexical association, weighted lexical association and semantic space methods are not significant at levels 2, 4 and 5 of the hierarchy. This indicates that only the weighted semantic space method outperforms prototype extraction when discriminating between many types of Appraisal.

3.5 Determining the Polarity of Attitude

Our next experiment involved an evaluation of the weakly-supervised methods' abilities to label instances of attitude according to the polarity, either positive or negative (using the polarity prototypes listed in the appendix). The optimal lexical association configuration used 160,000 lemmas and a window of 10 features, while the semantic space method worked best with 80,000 lemmas and a window of 4 features.

Table 3 lists the results of the methods using these parameters, as well as baselines of choosing polarity based on (1) the majority class in the development data and (2) the occurrence of prototypes. Surprisingly the weighted methods perform worse in this experiment. The reason for this is indicated by the majority baseline result, which is below than random choice because the majority class is different in the

Method	**Prec**	**Rec**	**F_1**
Semantic Spaces (w)	13.42	19.92	13.55
Semantic Spaces	12.46	21.12	12.42
Lexical Association (w)	7.37	15.58	9.90
Prototypes Baseline	20.00	6.12	9.30
Lexical Association	6.09	13.68	8.34
Majority Baseline	0.62	20.52	1.20

TABLE 2 Extracting Appraisal words

Method	**Prec**	**Rec**	**F_1**
Semantic Spaces	75.04	70.09	72.48
Semantic Spaces (w)	73.88	69.23	71.54
Lexical Association	66.02	59.18	62.41
Lexical Association (w)	65.78	58.71	62.04
Majority Baseline	46.74	46.74	46.74
Prototype Baseline	80.98	1.44	2.83

TABLE 3 Classifying the polarity of attitude

development and test set. Thus, the prior probability estimated from the development set is not reliable. However, the results show that the unweighted semantic space method is reasonably reliable in determining the polarity of attitudinal expressions (with an F_1 of 72.48).

3.6 Determining the Direction of Graduation

Our final experiment involved determining the weakly-supervised methods' ability to label instances of graduation according to their direction, either UP or DOWN. Test data was collected from the Appraisal corpus, using all expressions where the annotators agreed that some type of graduation was present, and on the direction of the graduation. Development data was compiled from instances of graduation in the symmetric difference of the sets of annotations. The optimal parameters for this task were 5,000 lemmas and a window of 10 features for lexical association and 320,000 lemmas and a window of 10 features for semantic spaces.

Table 4 lists the results of the methods using these parameters, with the majority and prototype baselines. The best performing weakly-supervised technique (weighted semantic space, $F_1 = 78.21$) gives a modest but significant improvement over the majority baseline, which

Method	Prec	Rec	$\mathbf{F}_1$
Semantic Spaces (w)	81.22	75.41	78.21
Majority Baseline	77.10	77.10	77.10
Semantic Spaces	76.34	69.29	72.64
Lexical Association (w)	81.85	62.23	70.96
Lexical Association	77.50	56.17	65.13
Prototype Baseline	84.24	10.50	18.67

TABLE 4 Classifying the direction of graduation

is particularly strong with 77.10% of graduations being up-scaling.

4 Discussion

This section discusses the outcomes of the weakly-supervised methods in performing an automatic analysis of Appraisal, considering their systematic shortcomings and successes.

Extraction Algorithms Depend on Prototype Frequency From raw frequency-based contingency tables it seems that the methods tended to select classes with high-frequency prototypes. We investigated this by calculating Pearson's correlation coefficient (r) between the prototype distribution and the distribution of Appraisal types output by the methods. In the extraction task, most of the distributions were correlated to various degrees (semantic space $r = 0.72$, weighted semantic space $r = 0.34$, lexical association $r = 0.14$, weighted lexical association $r = 0.07$), but this was not the case in the classification task. This suggests that the component of the method introducing the dependency on prototype frequencies (at least in the case of the semantic space method) is the threshold used for extraction, as it is the only distinction between the classification and extraction algorithms.

Words Must Co-occur with all Prototypes to Score Highly An analysis of the Lexical Association contingency tables indicated a surprising number of misclassifications of the type COUNTER as DENY. All of these instances included the word *but* in the expression. This is very surprising, as *but* is a prototype for COUNTER and should be highly indicative of that type. A detailed analysis of the concurrences indicated that *but* infrequently co-occurs with another prototype of COUNTER, *however*, such that they share a low pointwise mutual information score. In fact, this score is so low that when the mean score for COUNTER prototypes is calculated, the result is lower than that for DENY, even though the problem word *but* is a prototype of COUNTER.

		happy	**unhappy**
HAPPINESS	love	0.120	0.037
	laugh	0.082	0.035
	hate	0.078	0.081
	mean	0.093	0.051
SECURITY	confident	0.169	0.121
	anxious	0.138	0.272
	uneasy	0.060	0.188
	mean	0.122	0.193

TABLE 5 Similarity scores for the prototypes of HAPPINESS and SECURITY from the semantic space method for the words *happy* and *unhappy*.

This kind of error accounts for a number of misclassifications in the lexical association method's output, such as instances of DENY labelled as CONCUR, and a similar situation also occurs when the semantic space method misclassifies instances of COUNTER as DISTANCE. A potential solution to this problem is to use the maximum score for a prototype rather than the mean.

Prototypes are Insufficient to Discriminate Between some Classes Many of the misclassifications made by the methods involve pairs of classes that are closely related in the Appraisal hierarchy. For example, instances of HAPPINESS are frequently mistaken for SECURITY by the semantic space method. Table 5 demonstrates a particular example, comparing the semantic space method's scores for the words *happy* and *unhappy* for the classes of HAPPINESS and SECURITY. While one would expect the words to be indicative of HAPPINESS, they both score higher for SECURITY. When one considers the SECURITY prototypes, however, it is easy to concede that the words might be strongly associated. It therefore appears that the prototypes are insufficient for fine-grained discrimination between certain Appraisal classes.

Unweighted Semantic Space Output Distribution Correlates with the Test Distribution As one would expect, the classification decisions of all weighted methods strongly correlate with the distribution of Appraisal types found in the development and test data. Interestingly, however, the output of the unweighted Semantic Space method also correlates with the test distribution to a weak degree ($r = 0.11$). This is surprising as the unweighted versions of the algorithms are not augmented with any estimate about the actual distribution of Appraisal

types. This correlation suggests that the scores produced by the Semantic Space method implicitly represent the real distribution of Appraisal types, to a weak degree. It is not clear why this is the case, so this remains an interesting topic for further investigation.

5 Related Work

Taboada and Grieve (2004) examined word relatedness to the high-level classes of the attitude subsystem, AFFECT, JUDGEMENT and APPRECIATION. The work is similar to this study in that relatedness to a class was also estimated using an extension of Turney's (2002) technique, except that rather than using class prototypes Taboada and Grieve examined cooccurrence with three pronoun-copular pairs: *I was* for AFFECT, *he was* for JUDGEMENT and *it was* for APPRECIATION. They do not provide a quantitative evaluation of the method, but note that the variation of relatedness across different types of reviews (e.g. books, computers, hotels) appear to make sense intuitively—reviews of consumer products tend to contain more appreciation whereas reviews of creative works such as books and movies have higher values of judgement).

Whitelaw et al. (2005) investigated how insights from the Appraisal theory could be informative for the more general task of sentiment classification (i.e. classifying a unit of text according to whether it is generally positive or generally negative). They defined a lexicon of frames of sentiment that included slots for the high-level attitude type (i.e. AFFECT, JUDGEMENT or APPRECIATION), and whether the FORCE and FOCUS was low, neutral or high. These frames supplemented bag-of-words-based machine learning techniques for sentiment classification, resulting in some gain in accuracy.

Argamon et al. (2009) later considered how such lexicons might be automatically constructed, focusing on all eleven leaf types of ATTITUDE and FORCE. To build the lexicon they employ a set of seed terms (taken from Martin and White (2005)), which are expanded by following synonymy relations in WordNet. Then, training data for supervised machine learning classifiers is collated from the glosses of all expanded terms. They found that a Naïve Bayes classifier performed best with an F_1 of 37.1 for attitude types (baseline 15.8) and 75.7 for force (baseline 33.4).

Bloom et al. (2007) describe an approach for the extraction of Appraisal expressions. Constraining their approach to adjectival expressions, they identify both potential attitudes (limited to the high-level types of AFFECT, JUDGEMENT and APPRECIATION) and targets of Appraisal using manually-constructed lexicons. Attitudes are then linked

to targets using the analysis from a dependency parser and a manually-defined set of linking patterns. Finally a probabilistic model is used to disambiguate the attitude type in the case of multiple hits in the lexicon. Bloom et al. (2007) show that the extracted expressions can be used as input find generalised association rules to find interesting patterns in product reviews.

Bloom and Argamon (2009) discuss how the linking patterns can be learned automatically, by enumerating all links of potential attitudes and potential targets, using frequencies of the pattern seen as linking an attitude and target, and of the pattern with an attitude but no target. The automatically-learned patterns are slightly inferior to the manually-defined patterns, but require much less developmental effort. Bloom and Argamon (2010) discussed replacing their manually-constructed attitude lexicon with Turney and Littman's (2003) automatically-generated lexicon of sentiment-bearing words, showing that there was little difference between the two methods.

6 Conclusion and Future Directions

This article reported the application of a weakly-supervised approach for text classification to the analysis of aspects of Appraisal, a systemic functional linguistic theory of evaluation. This weakly-supervised method is appealing because it does not require labelled training data, but simply a large corpus and a manually-specified set of prototypical examples for each class. Of the two versions of the method, the semantic space version, while being more computationally-expensive, achieved better results with an average F_1 of 45.0 over all levels of the Appraisal hierarchy, compared to lexical association's F_1 of 22.1. The results also indicated that the semantic space method is more robust when dealing with many classes.

There are a number of possibilities that could be explored for improving the performance of the methods presented in this article. Firstly, Pang and Lee (2004) demonstrated that disregarding objective sentences can benefit supervised sentiment classification. The same procedures could be employed in the sentiment and appraisal analysis tasks, so that the weakly-supervised methods are only applied to subjective expressions. This may also be of benefit when training the weakly-supervised methods, assuming that objective text has no bearing on the sentiment of nearby subjective text.

Similarly, the performance of the weakly-supervised methods might be improved through procedures that tailor the training corpus for the relevant domain. This could be done automatically by determining

which words represent the domain, and using these words to constrain the methods' concept of cooccurrence, such that basis elements only count as cooccurrences if they also occur in proximity to the domain words.

Further gains in the performance of the weakly-supervised methods might be obtained by experimenting with feature selection. In this article the experimental set-up was simply to use the most frequent features, but it may be more productive to select features whose cooccurrences are the most variable (i.e. they provide more information by virtue of occurring with more features), or to employ a feature subsumption analysis (Riloff et al., 2006). Also, bootstrapping methods could also be applied, since they have been found to be effective for other sentiment analysis techniques (Riloff and Wiebe, 2003, Zagibalov and Carroll, 2008).

As noted in Section 3.1, the annotators of the book review corpus often formulated contradictory—but equally valid—interpretations of sentences. In the current work, this has the effect of reducing the quantity of reliable testing data. However, in future work we will consider a fuzzy interpretation of the annotations, wherein both annotator's decisions are considered as valid. The existing approach may be readily applied to this new task, though we anticipate making refinements in the threshold-selection process.

Our method, being focused on labelling words essentially out of context, is the first step in developing techniques for automatic analysis of Appraisal. Ongoing research in the broader field of sentiment analysis can be used to enhance our approach. For example, Wilson et al. (2009) present range of features for machine learning algorithms which are tailored to determine the polarity of specific instances of sentiment-bearing words. Such features might also be evaluated using the Appraisal corpus, not only in determining context-sensitive polarity but also in predicting a shift away from the attitude type determined by our weakly-supervised method. A further issue is that of determining the appraisal and polarity connoted by text (i.e. where opinion is implicit); this remains an open challenge for the sentiment analysis community at large.

We note that the work presented in this article is complementary to that of Bloom et al. (2007, 2009), as their work utilises a manually-defined lexicon. The present work might replace this lexicon, and could extend their system by automatically-regenerating the lexicon for new domains, and enabling analysis of appraisal expressions with respect to engagement and graduation. We also note that the fine-grained discrimination of attitude types provided by our approach could add further

detail to their opinion mining system.

Combining computational analysis of Appraisal with existing techniques in subjectivity analysis, sentiment analysis and opinion mining would open up new applications for opinion processing. For instance, being able to discriminate between the JUDGEMENT and APPRECIATION would enable companies to distinguish between opinions regarding their corporate image and commentary on the qualities of their products. It could also lead to new data collection techniques for social scientists, as instead of conventional time-consuming questionnaires, these researchers might utilise Appraisal-aware text mining software to collate expressions of affect and judgement relating to an issue of interest. Applications such as these need to be able to process text from a wide variety of domains and topics, and so future work should investigate ways of improving the performance of the weakly-supervised methods to a sufficiently robust level.

Acknowledgments

The work of the first author was supported by a UK EPSRC studentship and was carried out at the University of Sussex. We are very grateful to David Hope, for his help in annotating the book review corpus.

References

Argamon, Shlomo, Kenneth Bloom, Andrea Esuli, and Fabrizio Sebastiani. 2009. Automatically determining attitude type and force for sentiment analysis. In Z. Vetulani and H. Uszkoreit, eds., *Human Language Technologies as a Challenge for Computer Science and Linguistics*. Springer.

Bloom, Kenneth and Shlomo Argamon. 2009. Automated learning of appraisal extraction patterns. In *Corpus-linguistic applications: Current studies, new directions*, pages 249–260. Rodopi.

Bloom, Kenneth and Shlomo Argamon. 2010. Unsupervised extraction of appraisal expressions. In *Advances in Artificial Intelligence*, no. 6085 in Lecture Notes in Computer Science, pages 290–294. Springer.

Bloom, Kenneth, Navendu Garg, and Shlomo Argamon. 2007. Extracting appraisal expressions. In S. T. Gries, S. Wulff, and M. Davies, eds., *Proceedings of NAACL HTL 2007*, pages 308–315. Rochester, New York.

Church, Kenneth W. and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16:22–29.

Evert, Stefan. 2004. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, University of Stuttgart.

Graff, David. 2003. English gigaword. Linguistic Data Consortium, Philadelphia.

Grefenstette, Gregory. 1994. Corpus-derived first-, second- and third-order word affinities. In *Proceedings of Euralex*, pages 279–290. Amsterdam.

Halliday, Michael A. K. 1994. *An Introduction to Functional Grammar*. London: Edward Arnold.

Landauer, Thomas K. and Susan Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review* 104(2):211–240.

Levy, Joseph P., John A. Bullinaria, and Malti Patel. 1998. Explorations in the derivation of word co-occurrence statistics. *South Pacific Journal of Psychology* 10(1):99–111.

Lowe, Will. 2001. Towards a theory of semantic space. In *Proceedings of the 21st Annual Meeting of the Cognitive Science Society*, pages 576–581. Edinburgh, UK: Springer Verlag.

Lowe, Will and Scott McDonald. 2000. The direct route: Mediated priming in semantic space. In *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society*. Philadelphia, PA.

Lund, Kevin and Curt Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments & Computers* 28(2):203–208.

Martin, James R. and Peter R. R. White. 2005. *The Language of Evaluation: Appraisal in English*. Palgrave Macmillan.

Osgood, C.E., G.J. Suci, and P.H Tannenbaum. 1957. *The measurement of meaning*. Urbana, U.S.A.: University of Illinois Press.

Pang, Bo and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the Forty-Second Annual Meeting of the Association for Computational Linguistics*, pages 271–278. Barcelona, Spain.

Read, Jonathon and John Carroll. 2009. Weakly supervised techniques for domain-independent sentiment classification. In *Proceedings of the 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion Measurement*, pages 45–52. Hong Kong.

Read, Jonathon and John Carroll. 2012. Annotating expressions of Appraisal in English. *Language Resources and Evaluation* 46(3):421–447. 10.1007/s10579-010-9135-7.

Riloff, Ellen, Siddharth Patwardhan, and Janyce Wiebe. 2006. Feature subsumption for opinion analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 440–448. Sydney, Australia.

Riloff, Ellen and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of the Eighth Conference on Empirical Methods in Natural Language Processing*, pages 105–112. Sapporo, Japan.

Taboada, Maite and Jack Grieve. 2004. Analyzing Appraisal automatically. In *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text*, pages 158–161.

Turney, Peter D. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 417–424. Philadelphia, PA, USA.

Turney, Peter D. and Michael L. Littman. 2003. Measuring praise and criticisms: Inference of semantic orientation from association. *ACM Transactions on Information Systems* 21(4):315–346.

Whitelaw, Casey, Navendu Garg, and Shlomo Argamon. 2005. Using appraisal groups for sentiment analysis. In *Proceedings of the Fourteenth ACM International Conference on Information and Knowledge Management*. Bremen, Germany.

Wiebe, Janyce, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Computational Linguistics* 30(3):277–308.

Wilson, Theresa, Janyce Wiebe, and Paul Hoffman. 2009. Recognizing contextual polarity: An exploration of phrase-level sentiment analysis. *Computational Linguistics* 35(3):399–433.

Wilson, Theresa, Janyce Wiebe, and Rebecca Hwa. 2006. Recognizing strong and weak opinion clauses. *Computational Intelligence* 2(22):73–99.

Zagibalov, Taras and John Carroll. 2008. Unsupervised classification of sentiment and objectivity in Chinese text. In *Proceedings of the Third Joint International Conference on Natural Language Processing*, pages 304–311. Hyderabad, India.

Appendix: Prototypes

Appraisal

INCLINATION	demand, fear, request
HAPPINESS	love, laugh, hate
SECURITY	confident, anxious, uneasy
SATISFACTION	angry, pleasure, satisfaction
NORMALITY	normal, familiar, lucky
CAPACITY	expert, powerful, successful
TENACITY	brave, careful, loyal
VERACITY	honest, credible, deceptive
PROPRIETY	moral, evil, unfair
IMPACT	dramatic, intense, remarkable
QUALITY	beautiful, ugly, lovely
BALANCE	harmonious, logical, unified
COMPLEXITY	simple, precise, elegant
VALUATION	effective, appropriate, valuable
DENY	never, no, not
COUNTER	amazingly, but, however
PRONOUNCE	fact, indeed
ENDORSE	prove
AFFIRM	obviously
CONCEDE	admittedly
ENTERTAIN	apparently, perhaps, seem
ACKNOWLEDGE	argue, believe, say
DISTANCE	claim
NUMBER	few, many
MASS	large, small
PROXIMITY-SPACE	far, near
PROXIMITY-TIME	ancient, recent
DISTRIBUTION-SPACE	sparse, wide
DISTRIBUTION-TIME	long, short
DEGREE	extremely, slightly, very
VIGOUR	(None, see Section 3.2)
FOCUS	kind, true, sort

Polarity

POSITIVE	benefit, best, excellent, good, nice, perfect, supreme
NEGATIVE	abuse, bad, disastrous, evil, outrage, sad, wrong

Graduation

UP-SCALING	many, large, far, ancient, wide, long, extremely, very
DOWN-SCALING	few, small, near, recent, sparse, short, slightly