Automatically Determining Attitude Type and Force for Sentiment Analysis

Shlomo Argamon*, Kenneth Bloom*, Andrea Esuli†, Fabrizio Sebastiani†

*Linguistic Cognition Laboratory – Department of Computer Science
Illinois Institute of Technology – 10 W. 31st Street – Chicago, IL 60616, USA
{argamon,kbloom1}@iit.edu

†Istituto di Scienza e Tecnologie dell'Informazione – Consiglio Nazionale delle Ricerche
Via G Moruzzi, 1 – 56124 Pisa, Italy
{andrea.esuli,fabrizio.sebastiani}@isti.cnr.it

Abstract

Recent work in sentiment analysis has begun to apply fine-grained semantic distinctions between expressions of attitude as features for textual analysis. Such methods, however, require the construction of large and complex lexicons, giving values for multiple sentiment-related attributes to many different lexical items. For example, a key attribute is what type of *attitude* is expressed by a lexical item; e.g., `beautiful` expresses appreciation of an object's quality, while `evil` expresses a negative judgement of social behavior. In this paper we describe a method for the automatic determination of complex sentiment-related attributes such as *attitude type* and *force*, by applying supervised learning to WordNet glosses. Experimental results show that the method achieves good effectiveness, and is therefore well-suited to contexts in which these lexicons need to be generated from scratch.

1. Introduction

Recent years have seen a growing interest in *non-topical text analysis*, in which characterizations are sought of the opinions, feelings, and attitudes expressed in a text, rather than just of the topics the text is about. A key type of non-topical text analysis is *sentiment analysis*, which includes several important applications such as *sentiment classification*, in which a document is labelled as a positive ("thumbs up") or negative ("thumbs down") evaluation of a target object (film, book, product, etc.), and *opinion mining*, in which text mining methods are used to find interesting and insightful correlations between writers' opinions. Immediate applications include market research, customer relationship management, and intelligence analysis.

Critical to sentiment analysis is identifying useful features for the semantic characterization of the text. At the lexical level, most work on sentiment analysis has relied on either raw "bag-of-words" features from which standard text classifiers can be learned, or "semantic orientation" lexicons (Turney and Littman, 2003), which classify words as positive or negative (possibly with a weight), and on the use of those categories as a basis for analysis. Recent work, however, has started to apply more complex semantic taxonomies to sentiment analysis, either by developing more complex lexicons (Taboada and Grieve, 2004; Whitelaw et al., 2005) or by applying multiple text classifiers (Wilson et al., 2004) using supervised learning.

Both approaches present practical difficulties—supervised learning requires extensive text annotation, while developing lexicons by hand is also very time-consuming. The purpose of this paper is to explore the use of (semi-)supervised learning techniques to "bootstrap" semantically complex lexicons of terms with sentimental valence. Previous applications of such lexicons to sentiment analysis (Taboada and Grieve, 2004; Whitelaw et al., 2005) have used the framework of Martin and White's (2005) Appraisal Theory, developed for the manual analysis of evaluative language. This framework assigns several sentiment-related features to relevant lexical items, including *orientation* (Positive or Negative), *attitude type* (whether Affect, Appreciation of inherent qualities, or Judgement of social interactions), and *force* of opinion expressed (Low, Median, High, or Max). Such challenging multi-dimensional analysis can allow more subtle distinctions to be drawn than can just classifying terms as Positive or Negative.

We examine here the extent to which such a lexicon can be learned automatically, starting from a core (manually-constructed) lexicon of adjectives and adverbs. We apply a variant of a technique (Esuli and Sebastiani, 2005) originally developed for classifying words as Positive or Negative based on dictionary glosses. Experiments show that this variant works well for detecting *attitude type* and *force* as defined in Appraisal Theory.

After a brief overview of relevant aspects of Appraisal Theory (Sec. 2.), we describe our method for the automatic classification of lexical items by attitude type (Sec. 3.). Section 4. presents our experimental setup and results, followed by a brief overview of related work (Sec. 5.), and by concluding remarks (Sec. 6.).

2. Appraisal Theory

Appraisal Theory is a systemic-functional approach to analyzing how subjective language is used to express an attitude of some kind towards some target (Martin and White, 2005). Appraisal theory models appraisal as comprising three main linguistic systems: "Attitude", which distinguishes different kinds of attitudes that can be expressed (including Attitude Type and Orientation); "Amplification", which enables strengthening or weakening such expression (including Force and Focus); and "Engagement", which conveys different possible degrees of commitment to the opinion expressed (including identification and relation of the speaker/writer to the source of an attributed evaluation). Previous application of Appraisal Theory to sentiment analysis (Taboada and Grieve, 2004; Whitelaw et al., 2005) has focused on three key components:

Attitude Type specifies the type of appraisal being expressed as one of Affect, Appreciation, or Judgement (with further sub-typing possible). Affect refers to a personal emotional state (e.g., `happy`, `angry`), and is the most explicitly subjective type of appraisal. The other two options differentiate between the Appreciation of 'intrinsic' object properties (e.g., `slender`, `ugly`) and social Judgement (e.g., `heroic`, `idiotic`). Figure 1 gives a detailed view of the Attitude Type taxonomy, together with illustrative adjectives.

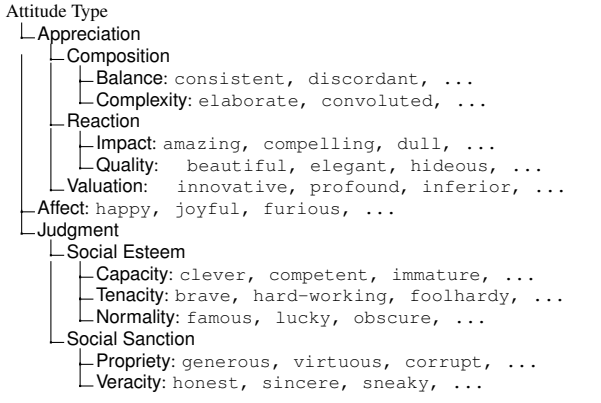

Figure 1: Options in the "Attitude Type" taxonomy, with examples of appraisal adjectives from the base lexicon described in Section 4.1..

Orientation determines whether the appraisal is Positive or Negative (this has also been termed "semantic orientation" or "polarity" in the sentiment analysis literature).

Force describes the intensity of the appraisal being expressed. Force may be realized via modifiers such as `very` (increased force) or `slightly` (decreased force), or may be realized lexically in a head word, e.g., `wonderful` vs. `great` vs. `good`.

Little research to date has applied such schemes in a computational context. Taboada and Grieve (2004) used a small lexicon of adjectives manually classified for top-level attitude type, expanded by a technique based on pointwise mutual information (PMI) (Turney and Littman, 2003). Their analysis showed that different types of review texts contain different amounts of each attitude type. Whitelaw et al. (2005) further showed how using attitude type, force and orientation, together with shallow parsing of evaluative adjective groups, can improve sentiment-based text classification. The current work explores how a lexicon such as that used in that work can be learned in a fully automatic fashion, concentrating on assigning the correct attitude type and force to lexical items.

These semantic features are also related to other analyses of term "value" or "sentiment" in the literature. Osgood's (1957) Theory of Semantic Differentiation delineated three dimensions of affective meaning: "evaluative", i.e., Orientation; "potency", referring to the strength of feeling expressed; and "activity", referring to how active or passive an evaluation is. This was the basis for Kamps and Marx's (2002) analyses of affective meaning in WordNet. Mullen and Collier (2004) estimated values for Osgood's three dimensions for adjectives in WordNet, by comparing path lengths to appropriate pairs of anchor words (such as `good` and `bad`) in WordNet's synonymy graph, using document-level averages of these values as input to SVMs for sentiment classification.

Also relevant is the Lasswell Value Dictionary, as applied in the General Inquirer (Stone et al., 1966). The purpose there is to classify words as relating to various basic "values", such as wealth, power, respect, rectitude, skill, enlightenment, affection, and wellbeing. Some of these have parallels in Appraisal Theory (for example "rectitude", which is similar to the attitude type of Social Sanction), while other Lasswell categories, such as "wealth" or "enlightenment" appear unrelated to any Attitude Type.

3. Methodology

3.1. Semi-supervised learning of orientation

The method we use in this paper for determining the attitude type and force of terms is inspired to the method proposed by Esuli and Sebastiani (2005) for determining orientation (called there "PN-polarity"). That method relies on training, in a semi-supervised way, a binary classifier that labels terms as either Positive or Negative. A *semi-supervised* method is a learning process whereby only a small subset $L \subset Tr$ of the training data Tr are manually labelled. In origin the training data in $U = Tr - L$ are instead unlabelled; it is the process itself that labels them, automatically, by using L (with the possible addition of other publicly available resources) as input. The method starts from two small seed (i.e. training) sets L_p and L_n of known Positive and Negative terms, respectively, and expands them into the two final training sets $Tr_p \supset L_p$ and $Tr_n \supset L_n$ by adding them new sets of terms U_p and U_n found by navigating the WordNet (2.0) graph along the synonymy and antonymy relations.

Perhaps more significant is the idea that terms are given vectorial representations based on their WordNet *glosses*. For each term t_i in $Tr \cup Te$ (Te being the test set, i.e. the set of terms to be classified), a textual representation of t_i is generated by collating all the glosses of t_i as found in WordNet[1]. Each such representation is converted into vectorial form by standard text indexing techniques.

The idea is that terms of similar semantic types should tend to have "similar" glosses: for instance, the glosses of `honest` and `intrepid` will both contain positive expressions, while the glosses of `disturbing` and `superfluous` will both contain negative expressions.

Once the vectorial representations for all terms in $Tr \cup Te$ have been generated, those for the terms in Tr are fed to a supervised learner, which thus generates a binary classifier. This latter, once fed with the vectorial representations of the terms in Te, classifies each of them as either Positive or Negative. Note that this method allows to classify *any* term, independently of its POS, provided there is a gloss for it in the lexical resource.

In this paper we adopt this gloss-based representation method using the above described vectorial representations to represent the terms of our lexicon.

3.2. Learning attitude type and force

Force is the simpler case here—we are faced with four categories, with each term belonging to exactly one of the four. Since the categories (Low, Median[2], High, and Max) are ordered along a scale of value, deciding which one applies to a given term is an *ordinal regression* problem. However, for the time being we (suboptimally) assume the problem is a 1-of-n *classification* problem (thereby disregarding the order among the categories), with n=4. We defer the use of ordinal regression for this problem to future work.

In determining attitude type, on the other hand, we are essentially faced with eleven binary distinctions, each consisting in determining whether the term belongs not to any of the eleven fine-grained attitude types of Figure 1. Note that in Appraisal Theory a term can have more than one such attitude type (e.g. `fair` is labeled, in the base lexicon described in Section 4.1., with attitude types Quality, Propriety, and Veracity)[3]. This means this is an *at-least-*

[1] In general a term t_i may have more than one gloss, since it may have more than one sense.

[2] Though Medium would be a more correct term in this scale, Median is the term used by Martin and White (2005).

[3] Out of a total of 1855 terms in our lexicon, 192 have more than one attitude type assigned.

1-of-n task, for $n = 11$, since we only work on terms that carry appraisal, and which thus belong to at least one of the attitude type classes. Note also that the eleven attitude types are leaves in a hierarchy.

This also allows us, if desired, to apply a hierarchical classification method, whereby the structure of the hierarchy is taken into account. Thus, in determining attitude type we consider two alternative classification methods. The *flat* method simply ignores the fact that the categories are organized into a hierarchy and plainly generates eleven independent binary classifiers $\hat{\Phi}_1, \ldots, \hat{\Phi}_{11}$; each such classifier $\hat{\Phi}_i$ is generated by using all the terms in Tr_i as positive examples and all terms not belonging to Tr_i as negative examples.

The *hierarchical* method is similar, but generates binary classifiers $\hat{\Phi}_j$ for each leaf *and* for each internal node. For an internal node c_j, as the set of positive training examples, the union of the sets of positive training examples of its descendant categories is used. For each node c_j (be it internal or leaf), as the set of negative examples we use the union of the positive training examples of its sibling categories (minus possible positive training examples of c_j). Both choices follow consolidated practice in the field of hierarchical categorization (Esuli et al., 2006). At classification time, test terms are classified by the binary classifiers at internal nodes, and only the ones that are classified as belonging to the node percolate down to the lower levels of the tree. The hierarchical method has the potential advantage of using more specifically relevant negative examples for training.

Regarding the vectorial representations used for terms, we collate all glosses for a given term into a single document; note that only glosses of synsets having the correct POS (adjective or adverb) are considered (see Section 4.3.). From the resulting documents we then remove stop words, stem terms, and compute term weights by cosine-normalized $tfidf$, a standard text indexing function from the IR tradition.

4. Experiments

We examined the use of two base learners for this task: (i) multinomial Naive Bayes, using Andrew McCallum's Bow implementation[4], and (ii) (linear kernel) Support Vector Machines, using Thorsten Joachims' SVMlight implementation [5]. We also compared three possible classification modes for combining binary classifiers for a multiple labeling problem: (i) m-of-n, which may assign zero, one, or several classes to the same test term; (ii) *at-least-1-of-n*, a variant of m-of-n which always assigns one class when m-of-n would assign no class; (iii) 1-of-n, which always assigns exactly one class. Note that, from what we have said in Section 3.2., the a priori optimal approaches for classifying according to attitude type and force are (ii) and (iii), respectively. However, we have run experiments in which we test each of (i)-(iii) on both attitude and force. There are several justifications for this; for instance, trying (i) on attitude type is justified by the fact that forcing at least one category assignment, as at-least-1-of-n does, promises to bring about higher recall but lower precision, and nothing guarantees that the balance will be favourable. Suboptimal as some of these attempts may be a priori, they are legitimate provided that we use the correct evaluation measure for the task.

All experiments reported in this paper were evaluated by running 10-fold cross validation on the eleven seed sets $Tr = \{Tr_1, \ldots, Tr_{11}\}$. To guarantee that, for each of the 10 experiments, each category c_i is adequately represented both in the training and in the validation set, we split *each* set Tr_i in 10 roughly equal parts, each of which is used in turn as the validation set (*stratified* cross-validation).

[4] `http://www-2.cs.cmu.edu/~mccallum/bow/`

[5] `http://svmlight.joachims.org/`

4.1. The lexicon

The lexicon[6] Tr has been constructed manually to give appraisal attribute values for a large number of evaluative adjectives and adverbs. Values for attitude type, orientation, and force are stored for each term. The lexicon was built starting with words and phrases given as examples for the different appraisal options in (Martin and White, 2005), finding more candidate terms and phrases using WordNet and two online thesauri[7]. Candidates were then manually checked and assigned attribute values. Very uncommon terms were automatically discarded, thus reducing the amount of manual work required.

The attitude type dimension of the corpus is defined by eleven different leaf categories, described in Section 2., each one containing 189 terms on the average (the maximum is 284 for Affect, the minimum is 78 for Balance); every term is labelled by at least one and at most three categories (the average being 1.12). The hierarchy of the attitude taxonomy is displayed in Figure 1. Force comprises four values in the corpus: Low (e.g., `adequate`), Median (e.g., `good`), High (e.g., `awesome`), and Max (e.g., `best`). Most (1464) entries in the corpus have Median force, with 30 Low, 323 High, and 57 Max.

4.2. Evaluation measure

For evaluation we use the well-known F_1 measure, defined as the harmonic mean of *precision* (π) and *recall* (ρ):

$$\pi = \frac{TP}{TP + FP} \quad (1)$$

$$\rho = \frac{TP}{TP + FN} \quad (2)$$

$$F_1 = \frac{2\pi\rho}{\pi + \rho} = \frac{2TP}{2TP + FP + FN} \quad (3)$$

where TP stands for true positives, FP for false positives, and FN for false negatives. Note that F_1 is undefined when $TP + FP + FN = 0$. However, in our lexicon there is at least one positive example for each category, thus $TP + FN > 0$ and F_1 is always defined.

We compute both *microaveraged* F_1 (denoted by F_1^{μ}) and *macroaveraged* F_1 (F_1^M). F_1^{μ} is obtained by (i) computing the category-specific values $TP(c_i)$, $FP(c_i)$, and $FN(c_i)$, (ii) obtaining TP as the sum of the $TP(c_i)$'s (same for FP and FN), and then (iii) applying Equation 3. F_1^M is obtained by (i) computing the category-specific precision and recall scores $\pi(c_i)$ and $\rho(c_i)$, (ii) computing $F_1(c_i)$ values for the individual categories c_i, applying Equation 3, and (iii) computing F_1^M as the unweighted average of the category-specific values $F_1(c_i)$; macroaveraged precision and macroaveraged recall (π^M and ρ^M) are computed similarly.

4.3. Results

We ran evaluations for all combinations of learning algorithm (NB and SVM), classification model (flat and hierarchical), and classification method (m-of-n, at-least-1-of-n,

[6] Available at: `http://lingcog.iit.edu/arc/appraisal_lexicon_2007b.tar.gz`

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

Table 1: Summary of averaged cross-validation results, showing microaveraged (π^μ, ρ^μ, F_1^μ) and macroaveraged (π^M, ρ^M, F_1^M) statistics. Each row shows the average over all runs (see text) for given values for certain independent variables (such as the learning algorithm, classification model, and so on), averaging over all others (indicated by –avg–). The baseline trivial acceptor result is reported for comparison.

Dimension	Algorithm	Model	Method	POS	π^μ	ρ^μ	F_1^μ	π^M	ρ^M	F_1^M
attitude	baseline	*n/a*	*n/a*	*n/a*	0.086	1.000	0.158	0.085	1.000	0.155
attitude	NB	–avg–	–avg–	–avg–	**0.320**	**0.397**	**0.332**	0.362	**0.376**	**0.305**
attitude	SVM	–avg–	–avg–	–avg–	0.254	0.237	0.223	**0.464**	0.233	0.186
attitude	–avg–	flat	–avg–	–avg–	**0.381**	**0.421**	**0.371**	0.389	**0.401**	**0.345**
attitude	–avg–	hier	–avg–	–avg–	0.192	0.213	0.184	**0.437**	0.208	0.147
attitude	–avg–	–avg–	m-of-n	–avg–	**0.334**	0.222	0.237	**0.509**	0.225	0.207
attitude	–avg–	–avg–	at-least-1-of-n	–avg–	0.243	**0.375**	0.285	0.388	**0.357**	0.253
attitude	–avg–	–avg–	1-of-n	–avg–	0.284	0.353	**0.310**	0.343	0.331	**0.277**
attitude	–avg–	–avg–	–avg–	Adj,Adv	0.286	**0.318**	0.277	0.411	0.305	0.245
attitude	–avg–	–avg–	–avg–	Adj,Adv,V	0.285	**0.318**	0.277	0.412	**0.306**	0.246
attitude	–avg–	–avg–	–avg–	Adj,Adv,N	**0.289**	0.317	**0.279**	**0.417**	0.303	**0.247**
attitude	–avg–	–avg–	–avg–	Adj,Adv,V,N	0.287	0.315	0.277	0.413	0.303	0.245
force	baseline	*n/a*	*n/a*	*n/a*	0.201	1.000	0.334	0.158	1.000	0.239
force	NB	*n/a*	–avg–	–avg–	0.585	**0.732**	**0.634**	0.281	**0.614**	**0.352**
force	SVM	*n/a*	–avg–	–avg–	**0.586**	0.498	0.499	**0.662**	0.214	0.187
force	–avg–	*n/a*	m-of-n	–avg–	**0.755**	0.759	**0.757**	**0.501**	0.404	**0.305**
force	–avg–	*n/a*	at-least-1-of-n	–avg–	0.591	**0.806**	0.661	0.476	**0.487**	0.288
force	–avg–	*n/a*	1-of-n	–avg–	0.688	0.688	0.688	0.473	0.406	0.280
force	–avg–	*n/a*	–avg–	Adj,Adv	0.677	0.750	0.701	0.489	0.432	0.290
force	–avg–	*n/a*	–avg–	Adj,Adv,V	0.677	0.750	0.701	0.479	0.430	0.291
force	–avg–	*n/a*	–avg–	Adj,Adv,N	**0.680**	**0.753**	**0.704**	**0.490**	**0.434**	0.291
force	–avg–	*n/a*	–avg–	Adj,Adv,V,N	0.679	**0.753**	**0.704**	0.475	0.433	**0.292**

and 1-of-n); we also considered the effect of using glosses from parts of speech other than adjectives and adverbs, to see how stable our method is in the face of the ambiguity introduced. For comparison we computed also F_1 as obtained by a trivial baseline consisting of the *trivial acceptor*[8] classifier, which is the baseline classifier for the F_1 measure. Table 1 summarizes our results, comparing the effects of different values for each independent variable by averaging over results for the other variables.

Attitude type: Here, best results are clearly achieved by Naive Bayes; this result holds also for the non-averaged results of individual runs (omitted for lack of space). Surprisingly, the flat classification model works noticeably better than the hierarchical model, which may indicate that the shared semantics of siblings in the taxonomy is not well-represented in the WordNet glosses. Regarding classification methods, while the m-of-n and at-least-1-of-n methods achieve the highest precision and recall, respectively, the 1-of-n method achieves the best balance between the two, as measured by F_1—this may be explained by the relatively low average ambiguity (1.12 – defined as the average number of categories per term) of the lexicon, which makes this m-of-n task similar to an 1-of-n task. In practice, the higher recall method should probably be preferred, since incorrect category assignments could be weeded out at the text analysis stage. Finally, we note that including glosses from POS other than those in the lexicon did not appreciably change results.

Force: Here, as for attitude type, Naive Bayes dominates for recall and F_1, while SVMs achieve better precision. Also similar is that at-least-1-of-n classification increases recall at the expense of precision; 1-of-n, which is the a priori optimal method for force, achieves slightly better (macroaveraged) F_1 than m-of-n, but the difference is slight. More significant, however, is that micro- and macroaveraged F_1 are quite different for force, showing that the majority category, Median, comprising 78% of terms, is better classified than other classes, though results still indicate that minority classes are being identified with reasonable accuracy. Treatment of force in the future as an ordinal regression problem may help with this issue.

In both cases the improvement in accuracy with respect to the baseline is substantial, especially in terms of F_1^μ.

[8] A classifier which assigns every label to every document.

5. Previous Work

Most previous work dealing with the properties of terms from the standpoint of sentiment analysis has dealt with five main tasks:

1. Determining *orientation*: i.e., deciding if a given Subjective term (i.e. a term that carries evaluative connotation) is Positive or Negative.
2. Determining *subjectivity*: i.e., deciding whether a given term has a Subjective or an Objective (i.e. neutral, or factual) nature.
3. Determining the *strength* of term sentiment: i.e., attributing degrees of positivity or negativity.
4. Tackling Tasks 1–3 for term *senses*; i.e., properties such as Subjective, Positive, or Mildly Positive, are predicated of individual term senses, taking into account the fact that different senses of the same ambiguous term may have different sentiment-related properties.
5. Tackling Tasks 1–3 for *multiword terms*: i.e., properties such as Subjective, Positive, or Mildly Positive are predicated of complex expressions such as `not entirely satisfactory`.

Concerning Task 1, the most influential work is probably (Turney and Littman, 2003), who determine the orientation of subjective terms by bootstrapping from two (a Positive and a Negative) small sets of subjective "seed" terms. Their method computes the *pointwise mutual information* of the target term t with each seed term t_i, as a measure of their semantic association. PMI is a real-valued function, and its scores can thus be used also for Task 3. Other efforts at solving Task 1 include (Andreevskaia and Bergler,

2006; Esuli and Sebastiani, 2005; Hatzivassiloglou and McKeown, 1997; Kamps et al., 2004; Kim and Hovy, 2004; Takamura et al., 2005).

Task 2 has received less attention than Task 1 in the research community. Esuli and Sebastiani (2006a) show it to be much more difficult than Task 1, by employing variants of the method by which they had obtained state-of-the-art effectiveness at Task 1 (2005) and showing that much lower performance can be obtained. Other works dealing with this task are those of Andreevskaia and Bergler (2006), Baroni and Vegnaduzzo (2004), Riloff et al. (2003), and Wiebe (2000).

Task 4 has been addressed by Esuli and Sebastiani (2006b) by applying a committee of independent classifiers to the classification of each of the WordNet synsets.

The only work we are aware of on Task 5 is that of Whitelaw et al. (2005), who developed a method for using a structured lexicon of appraisal adjectives and modifiers to perform chunking and analysis of multi-word adjectival groups expressing appraisal, such as `not very friendly`, analysed as having Positive orientation, Propriety attitude type, and Low force. Experimental results showed that using such "appraisal groups" as features for movie review classification improved sentiment classification.

6. Conclusion

We have shown in this paper how information contained in dictionary glosses can be exploited to automatically determine the type and force of attitudes expressed by terms. These are challenging tasks, given that there are many classes (four levels of force and eleven of attitude type). We have used an adapted version of a method previously applied to the simpler task of recognizing *polarity* (Esuli and Sebastiani, 2005). Though effectiveness values from experiments are not high in absolute value, the improvement with respect to the baseline is relevant, showing the feasibility of automatic construction of lexicons in which a variety of sentiment-related attributes are attributed to words for use in appraisal extraction and sentiment analysis. Future work will seek to improve the methods developed here by refining feature choice and processing from glosses, as well as incorporating other sources of information, such as collocations from large, general corpora.

Acknowledgments

The work of the third and fourth authors was partially funded by the Project ONTOTEXT "From Text to Knowledge for the Semantic Web", funded by the Provincia Autonoma di Trento under the 2004–2006 "Fondo Unico per la Ricerca" funding scheme.

7. References

Andreevskaia, Alina and Sabine Bergler, 2006. Mining WordNet for fuzzy sentiment: Sentiment tag extraction from WordNet glosses. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'06)*. Trento, IT.

Baroni, M. and S. Vegnaduzzo, 2004. Identifying subjective adjectives through Web-based mutual information. In *Proceedings of the 7th Konferenz zur Verarbeitung Natürlicher Sprache (German Conference on Natural Language Processing – KONVENS'04)*. Vienna, AU.

Esuli, Andrea, Tiziano Fagni, and Fabrizio Sebastiani, 2006. TreeBoost.MH: A boosting algorithm for multi-label hierarchical text categorization. In *Proceedings of the 13th International Symposium on String Processing and Information Retrieval (SPIRE'06)*. Glasgow, UK.

Esuli, Andrea and Fabrizio Sebastiani, 2005. Determining the semantic orientation of terms through gloss analysis. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM'05)*. Bremen, DE.

Esuli, Andrea and Fabrizio Sebastiani, 2006a. Determining term subjectivity and term orientation for opinion mining. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'06)*. Trento, IT.

Esuli, Andrea and Fabrizio Sebastiani, 2006b. SENTIWORDNET: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06)*. Genova, IT.

Hatzivassiloglou, Vasileios and Kathleen R. McKeown, 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL'97)*. Madrid, ES.

Kamps, Jaap and Maarten Marx, 2002. Words with attitude. In *Proceedings of the 1st Global WordNet Conference (GWC'02)*. Mysore, IN.

Kamps, Jaap, Maarten Marx, Robert J. Mokken, and Maarten De Rijke, 2004. Using WordNet to measure semantic orientation of adjectives. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, volume IV. Lisbon, PT.

Kim, Soo-Min and Eduard Hovy, 2004. Determining the sentiment of opinions. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING'04)*. Geneva, CH.

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

Mullen, Tony and Nigel Collier, 2004. Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of the 9th Conference on Empirical Methods in Natural Language Processing (EMNLP'04)*. Barcelona, ES.

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

Riloff, Ellen, Janyce Wiebe, and Theresa Wilson, 2003. Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the 7th Conference on Natural Language Learning (CONLL'03)*. Edmonton, CA.

Stone, P. J., D. C. Dunphy, M. S. Smith, and D. M. Ogilvie, 1966. *The General Inquirer: A Computer Approach to Content Analysis*. Cambridge, US: MIT Press.

Taboada, Maite and Jack Grieve, 2004. Analyzing appraisal automatically. In *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*. Stanford, US.

Takamura, Hiroya, Takashi Inui, and Manabu Okumura, 2005. Extracting emotional polarity of words using spin model. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*. Ann Arbor, US.

Turney, Peter D. and Michael L. Littman, 2003. Measuring praise and criticism: 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 14th ACM International Conference on Information and Knowledge Management (CIKM'05)*. Bremen, DE.

Wiebe, Janyce, 2000. Learning subjective adjectives from corpora. In *Proceedings of the 17th Conference of the American Association for Artificial Intelligence (AAAI'00)*. Austin, US.

Wilson, Theresa, Janyce Wiebe, and Rebecca Hwa, 2004. Just how mad are you? Finding strong and weak opinion clauses. In *Proceedings of the 21st Conference of the American Association for Artificial Intelligence (AAAI'04)*. San Jose, US.