Automatically Determining Attitude Type and Force for Sentiment Analysis

Shlomo Argamon[1], Kenneth Bloom[1], Andrea Esuli[2], and Fabrizio Sebastiani[2]

[1] Linguistic Cognition Laboratory
Department of Computer Science
Illinois Institute of Technology
10 W. 31st Street – Chicago, IL 60616, USA
{argamon,kbloom1}@iit.edu
[2] 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 judgment of social behavior. In this chapter 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.

Keywords: Sentiment analysis, Lexicon learning, WordNet, Appraisal theory.

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 or negative 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

Z. Vetulani and H. Uszkoreit (Eds.): LTC 2007, LNAI 5603, pp. 218–231, 2009.

has relied on either raw "bag-of-words" features from which standard text classifiers can be learned, or on "semantic orientation" lexicons [1], which classify words as positive or negative (possibly with a weight), 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 [2,3] or by applying multiple text classifiers [4] 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 chapter 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 [2,3] have used the framework of Martin and White's [5] 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 Judgment 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.

Little research to date has applied such schemes in a computational context. Taboada and Grieve [2] used a small lexicon of adjectives manually classified for top-level attitude type, expanded by a technique based on pointwise mutual information (PMI) [1]. Their analysis showed that different types of review texts contain different amounts of each attitude type. Whitelaw et al. [3] and Bloom et al. [6] further showed how using attitude type, force and orientation, together with shallow parsing of evaluative adjective groups, can improve accuracy of sentiment-based text classification and also enables more detailed opinion mining than methods based only on classifying sentiment as Positive or Negative.

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. We examine here the extent to which such a semantically-complex lexicon for sentiment analysis can be learned automatically, starting from a core (manually-constructed) lexicon of adjectives and adverbs. We apply a variant of a technique [7] 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.*

2 Appraisal Theory

Appraisal Theory is a linguistic approach to analyzing how subjective language is used to express various sorts of attitudes towards various targets [5]. 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); "Graduation", which enables strengthening or weakening such expression (including force and focus); and "Engagement",

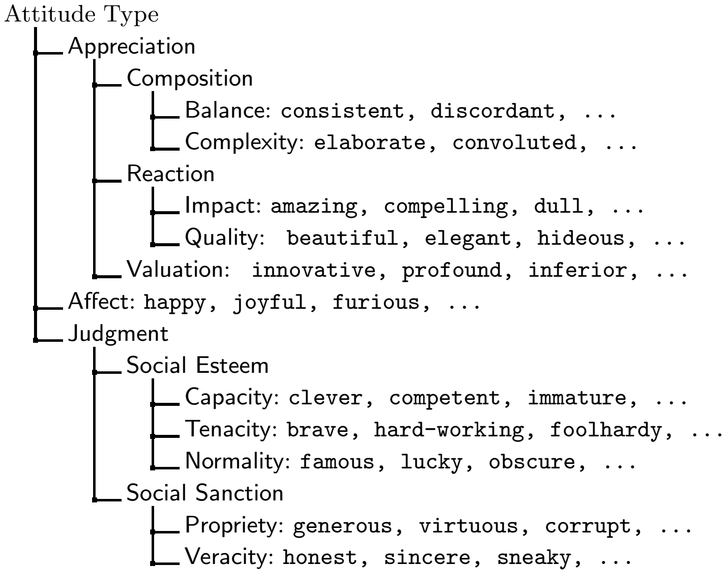
Attitude Type
Appreciation
Composition
Balance: consistent, discordant, ...
Complexity: elaborate, convoluted, ...
Reaction
Impact: amazing, compelling, dull, ...
Quality: beautiful, elegant, hideous, ...
Valuation: innovative, profound, inferior, ...
Affect: happy, joyful, furious, ...
Judgment
Social Esteem
Capacity: clever, competent, immature, ...
Tenacity: brave, hard-working, foolhardy, ...
Normality: famous, lucky, obscure, ...
Social Sanction
Propriety: generous, virtuous, corrupt, ...
Veracity: honest, sincere, sneaky, ...

Fig. 1. Options in the attitude type taxonomy, with examples of appraisal adjectives from the base lexicon described in Section 4.1

which represents different possible degrees and kinds 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 [2,3,8] has focused on three key components:

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

Attitude Type specifies the type of appraisal being expressed as one of Affect, Appreciation, or Judgment (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 Judgment (e.g., `heroic`, `idiotic`). Figure 1 gives a detailed view of the attitude type taxonomy, together with illustrative adjectives.

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`.

These semantic features are also related to other analyses of term "value" or "sentiment" in the literature. Osgood's [9] 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 [10] analyses of affective meaning in WordNet. Mullen and Collier [11] 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 [12]. It classifies words as relating to various basic "values", such as wealth, power, respect, rectitude, skill, enlightenment, affection, and wellbeing. Some 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 chapter for determining the attitude type and force of terms is inspired to the method proposed by Esuli and Sebastiani [7] 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.

Note that when such expansion is used, nothing prevents a term from belonging *both* to Tr and Te. To see this, remember that the training set Tr is the union of a set L of manually labelled terms and a set U of automatically labelled ones. While, conforming to good machine learning practice, we do need to ensure that $L \cap Te = \emptyset$, there is nothing wrong if $U \cap Te \neq \emptyset$.

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. (In general, a term t_i may have more than one gloss, since it may have more than one sense.) Each such representation is converted into vectorial form by standard text indexing techniques.

In addition, *negation propagation* is performed on each gloss, by replacing all the terms that occur in the context of a negation with a synthetic term representing the negated term. For example, the vector for the gloss "*the act of moving hurriedly and in a careless manner*" (for the word "rushing"), will comprise elements for *act*, *moving*, *hurriedly*, *careless*, and *manner*, while that for the gloss "*not moving quickly*" (for the word "slow") will comprise elements for the synthetic features $\neg$*moving* and $\neg$*quickly.*

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 classification of *any* term, independent of its part-of-speech, provided there is a gloss for it in the lexical resource.

The basic 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.

In this chapter 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, High, and Max) are ordered along a scale of value, deciding which one applies to a given term is an *ordinal regression* problem [13]. 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 or not the term belongs to any of the eleven fine-grained attitude types given in Figure 1. Note that a single term may be semantically ambiguous, and thus labeled by more than one attitude type (e.g., `fair` is labeled, in the base lexicon described in Section 4.1, with attitude types Quality, Propriety, and Veracity)[1]. This means this is an *at-least-1-of-n* task with $n = 11$, since we only work on terms that carry appraisal, and which thus belong to at least one of the attitude type classes. Since the eleven attitude types are leaves in a hierarchy, we may instead 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

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

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 [14]. 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.

To produce a vector for a given term, we collate all glosses for the term into a single document; note that only glosses of synsets having certain parts-of-speech are considered (see Section 4.3). From the resulting documents we then remove stop words, stem terms, and compute term weights by cosine-normalized *tf·idf*, a standard method in information retrieval.

When performing training set expansion on seed sets $Tr^1 = \{Tr^1_1, \ldots, Tr^1_n\}$ and expand them into the final n training sets $Tr = Tr^K = \{Tr^K_1, \ldots, Tr^K_n\}$ after K iterations. For expansion, synonyms *and* antonyms of a training term are added to the training set of the same class, as antonyms will differ in *orientation* but neither in *attitude type* nor in *force* e.g., Balance includes both `consistent` and `discordant`, while Tenacity includes both `brave` and `foolhardy`. (This contrasts with expansion for binary *orientation* classification [7], where antonyms were added to the training set of the *opposite* class.)

4 Experiments

We examined the use of two base learners for this task: (i) multinomial Naive Bayes, using Andrew McCallum's Bow implementation[2], and (ii) (linear kernel) Support Vector Machines, using Thorsten Joachims' SVMlight implementation[3]. Note that we used the *tf·idf* weighted representations only when using the SVM learner, since Naive Bayes requires binary input. Actually, the use of *multinomial* Naive Bayes ensures that raw term frequencies are *de facto* taken into account.

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 the preferred approaches for classifying by 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

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

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

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 chapter were evaluated by running 10-fold cross validation on the eleven seed sets $Tr = \{Tr_1, \ldots, Tr_{11}\}$ for attitude type and on the four seed sets $Tr = \{Tr_1, \ldots, Tr_4\}$ for force. To guarantee that each category c_i is adequately represented both in the training and the testing sets, we use *stratified* cross-validation, where we split *each* set Tr_i into 10 roughly equal parts, each of which is used in turn as a test set.

4.1 The Lexicon

The lexicon[4] 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 type values by Martin and White [5], finding more candidate terms and phrases using WordNet and two online thesauruses[5]. Candidates were then manually checked and assigned attribute values. Very infrequent 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.

Note that while lexicon entries also include values for orientation, we only consider here classification by attitude and by force. For a thorough study of the problem of determining orientation by means of methods similar to the ones discussed here please refer to [7,15].

4.2 Evaluation Measures

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

$$\pi = \frac{TP}{TP + FP} \tag{1}$$

$$\rho = \frac{TP}{TP + FN} \tag{2}$$

$$F_1 = \frac{2\pi\rho}{\pi + \rho} = \frac{2TP}{2TP + FP + FN} \tag{3}$$

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

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

Table 1. Summary of averaged cross-validation results for attitude type, 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. The fixed variable in each row and the highest value in each column for each set of comparable results are **boldfaced** for ease of reading.

Alg.	**Model**	**Method**	**POS**	π^{μ}	ρ^{μ}	F_1^{μ}	π^{M}	ρ^{M}	F_1^{M}
baseline	*n/a*	*n/a*	*n/a*	0.086	1.000	0.158	0.085	1.000	0.155
NB	–avg–	–avg–	–avg–	**0.320**	**0.397**	**0.332**	0.362	**0.376**	**0.305**
SVM	–avg–	–avg–	–avg–	0.254	0.237	0.223	**0.464**	0.233	0.186
–avg–	**flat**	–avg–	–avg–	**0.381**	**0.421**	**0.371**	0.389	**0.401**	**0.345**
–avg–	**hier**	–avg–	–avg–	0.192	0.213	0.184	**0.437**	0.208	0.147
–avg–	–avg–	**m-of-n**	–avg–	**0.334**	0.222	0.237	**0.509**	0.225	0.207
–avg–	–avg–	**at-least-1**	–avg–	0.243	**0.375**	0.285	0.388	**0.357**	0.253
–avg–	–avg–	**1-of-n**	–avg–	0.284	0.353	**0.310**	0.343	0.331	**0.277**
–avg–	–avg–	–avg–	**Adj,Adv**	0.286	**0.318**	0.277	0.411	0.305	0.245
–avg–	–avg–	–avg–	**Adj,Adv,V**	0.285	**0.318**	0.277	0.412	**0.306**	0.246
–avg–	–avg–	–avg–	**Adj,Adv,N**	**0.289**	0.317	**0.279**	**0.417**	0.303	**0.247**
–avg–	–avg–	–avg–	**Adj,Adv,V,N**	0.287	0.315	0.277	0.413	0.303	0.245

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 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, 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 a classifier which assigns every label to every document, which is the standard baseline classifier for the F_1 measure. Tables 1 through 4 summarize our results. We first note that in both cases we obtained substantial improvements in accuracy with respect to the baseline.

Table 2. Summary of best results for attitude type classification, showing, for each setting for each variable, the settings of the other variables that give the highest microaveraged F_1 value. In each row, the fixed variable value is given in boldface.

Alg.	Model	Method	POS	π^μ	ρ^μ	F_1^μ	π^M	ρ^M	F_1^M
NB	flat	1-of-n	Adj,Adv,N	**0.416**	**0.490**	**0.449**	0.429	**0.450**	**0.417**
SVM	flat	1-of-n	Adj,Adv,V	0.413	0.411	0.412	**0.430**	0.388	0.386
NB	**flat**	1-of-n	Adj,Adv,V,N	0.418	**0.483**	**0.448**	0.431	**0.442**	**0.413**
NB	**hier**	n-of-m	Adj,Adv	**0.482**	0.214	0.295	0.**521**	0.184	0.240
SVM	flat	**at-least-1**	Adj,Adv,V,N	0.404	0.409	0.406	0.410	0.382	0.379
NB	flat	**1-of-n**	Adj,Adv,N	**0.416**	**0.490**	**0.449**	**0.429**	0.450	**0.417**
NB	flat	**n-of-m**	Adj,Adv	0.338	0.484	0.398	0.306	**0.502**	0.380
NB	flat	1-of-n	**Adj,Adv**	0.408	0.489	0.444	0.419	0.450	**0.413**
SVM	flat	1-of-n	**Adj,Adv,N**	0.412	0.410	0.411	0.428	0.384	0.383
NB	flat	1-of-n	**Adj,Adv,V**	0.409	0.482	0.442	0.424	**0.444**	0.411
NB	flat	1-of-n	**Adj,Adv,V,N**	**0.418**	**0.483**	**0.448**	**0.431**	0.442	**0.413**

Table 3. Summary of averaged cross-validation results for force, as in Table 1. Note that only the flat classification model is applicable here.

Alg.	Method	POS	π^μ	ρ^μ	F_1^μ	π^M	ρ^M	F_1^M
baseline	*n/a*	*n/a*	0.201	1.000	0.334	0.158	1.000	0.239
NB	–avg–	–avg–	0.585	**0.732**	**0.634**	0.281	**0.614**	**0.352**
SVM	–avg–	–avg–	**0.586**	0.498	0.499	**0.662**	0.214	0.187
–avg–	**m-of-n**	–avg–	**0.755**	0.759	**0.757**	**0.501**	0.404	**0.305**
–avg–	**at-least-1**	–avg–	0.591	**0.806**	0.661	0.476	**0.487**	0.288
–avg–	**1-of-n**	–avg–	0.688	0.688	0.688	0.473	0.406	0.280
–avg–	–avg–	**Adj,Adv**	0.677	0.750	0.701	0.489	0.432	0.290
–avg–	–avg–	**Adj,Adv,V**	0.677	0.750	0.701	0.479	0.430	0.291
–avg–	–avg–	**Adj,Adv,N**	**0.680**	**0.753**	**0.704**	**0.490**	**0.434**	0.291
–avg–	–avg–	**Adj,Adv,V,N**	0.679	**0.753**	**0.704**	0.475	0.433	**0.292**

Attitude Type: Table 1 shows the overall effects of different values for each independent variable on attitude type classification, by averaging over results for the other variables. Table 2 shows the best results for various variable values—we repeatedly fixed the value of one variable and present the settings of the other variables that gave the highest microaveraged F_1. This was repeated for each value of each variable to give the results in the table.

For attitude type classification, when we consider the averaged results, we see that overall best results are achieved by Naive Bayes. When considering the best settings relative to other system variables in Table 2, we see a similar pattern, though the difference in F_1 performance is less. (Note that two specific sets of variable values, one using Naive Bayes and one using SVM, dominate the results in this table.)

Table 4. Summary of best individual results (for macroaveraged F_1) for force classification, arranged as in Table 2

Alg.	Method	POS	π^{μ}	ρ^{μ}	F_1^{μ}	π^{M}	ρ^{M}	F_1^{M}
NB	m-of-n	Adj,Adv,V	0.737	0.746	0.741	0.296	**0.549**	**0.384**
SVM	1-of-n	Adj,Adv,N	**0.771**	**0.770**	**0.771**	**0.715**	0.253	0.232
NB	**at-least-1**	Adj,Adv,N	0.414	0.844	0.555	0.286	**0.729**	0.350
NB	**1-of-n**	Adj,Adv,V,N	0.466	**0.880**	0.609	0.275	0.562	0.334
NB	**n-of-m**	Adj,Adv,V	**0.737**	0.746	**0.741**	**0.296**	0.549	**0.384**
NB	m-of-n	**Adj,Adv**	**0.740**	**0.751**	**0.746**	0.287	**0.562**	0.380
NB	m-of-n	**Adj,Adv,V**	0.737	0.746	0.741	**0.296**	0.549	**0.384**
NB	m-of-n	**Adj,Adv,N**	0.739	0.750	0.745	0.286	0.559	0.378
NB	m-of-n	**Adj,Adv,V,N**	0.736	0.747	0.741	0.286	0.560	0.379

Next, we find surprisingly that the flat classification model works noticeably better than the hierarchical model. This likely indicates that the shared semantics of siblings in the attitude type taxonomy is not well-represented in the WordNet glosses.

Regarding classification methods, when averaging over other variables, we see that while m-of-n and at-least-1-of-n 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. When considering the best individual runs, we see that the preferred classification method is nearly always 1-of-n as well.

Finally, we note that including glosses from parts-of-speech other than those in the lexicon did not appreciably change results.

Force: Table 3 shows the overall effects of different values for each independent variable on force classification, by averaging over results for the other variables. Table 4 shows the best results for macroaveraged F_1 for the various variable values, in the same format as in Table 2.

For force, when averaged over other variable settings, Naive Bayes achieves better recall and F_1, while SVMs achieve better precision under macroaveraging. The same pattern held for macroaveraged results for the best individual runs, though microaveraged results were similar for the two algorithms.

Also similar to the case of attitude type 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 better (macroaveraged) F_1 than m-of-n, but the difference is small. When we consider the best individual runs for each method, we see at-least-1-of-n classification increases macroaveraged recall at the expense of

Table 5. 10-fold cross-validation results for attitude type classification using Naive Bayes with flat m-of-n categorization, under three different levels of expansion (K) of the training sets ($K = 0$ means no expansion)

K	π^μ	ρ^μ	F_1^μ	π^M	ρ^M	F_1^M
0	.338	.484	**.398**	.306	.502	**.380**
1	.316	.478	.380	.293	.495	.368
2	.305	.467	.369	.287	.480	.359

precision; 1-of-n, the *a priori* preferred method for force, gives slightly better microaveraged precision, but m-of-n gives the best F_1 by a slight margin.

As in the case of attitude type, the preferred classification method appears to be correlated with the choice of classification algorithm, with m-of-n working best with Naive Bayes and at-least-1-of-n working best with SVM.

For force, as for attitude type, we find that addition of glosses from other parts-of-speech did not appreciably affect results.

Significantly we find micro- and macroaveraged F_1 to be quite different for force, showing that the majority category, Median, comprising 78% of terms, is noticeably better classified than other classes, though results do indicate that minority classes are being identified with reasonable accuracy. Treatment of force in the future as an ordinal regression problem may help.

Expansion: Table 5 reports results for attitude type of applying expansion to the training sets, as described in Section 3.2. In contrast to previous results for orientation, expansion results in decreased effectiveness: the change in F_1^μ is -5.3% for $K = 1$ and -7.3% for $K = 2$. This is likely due to the fact that the seed sets of these experiments can be considered as already "expanded"; to see this, we need only to compare their size (average: 189 terms each) with the size of those used previously for orientation (maximum: 7 terms each). Expansion thus appears to add only "noise" to the training sets under these conditions. Future work will include exploration of the effect of expansion for different seed set sizes.

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`.

The most influential technique for Task 1 is probably that of Turney [1], which determines the orientation of subjective terms by bootstrapping from two (a Positive and a Negative) small sets of "seed" terms. Their method computes the *pointwise mutual information* (PMI) 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 use of rhetorical relationships between words [16,17], WordNet path lengths and synonym sets [18,19], and WordNet glosses [7,20].

Task 2 has received less attention than Task 1 in the research community. Esuli and Sebastiani [15] have shown 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 [7] and showing that much lower performance is obtained. Other methods that have been applied to this task are those of Andreevskaia and Bergler [20], who consider WordNet paths and glosses, Baroni and Vegnaduzzo [21], who use mutual information, Riloff et al. [22], who use bootstrapped information extraction patterns, and Wiebe [23], who combined supervised learning with distributional similarity measures.

Task 4 has been addressed by Esuli and Sebastiani [24] by applying a committee of independent classifiers to the classification of each of the WordNet synsets. The sum of the scores attributed by the individual classifiers is used for the final classification decision. The magnitude of this sum is used as an indication of the strength of association of the synset to either Positive, Negative, or Objective.

Comparatively little work has been done on Task 5. The most comprehensive approach to this task that we are aware of is that by Whitelaw et al. [3] as extended by Bloom et al. [6]. Their method uses 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. The lexicon used in the experiments reported here is based on that developed in this work. Experimental results showed that using such "appraisal groups" as features for sentiment classification improved classification results. Other related work includes research on valence shifting [25,26] and contextual polarity [27].

6 Conclusion

We have shown 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* [7]. 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 third and fourth authors were 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.

References

1. Turney, P.D., Littman, M.L.: Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems 21(4), 315–346 (2003)
2. Taboada, M., Grieve, J.: Analyzing appraisal automatically. In: Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications (2004)
3. Whitelaw, C., Garg, N., Argamon, S.: Using appraisal groups for sentiment analysis. In: Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM 2005), Bremen, DE, pp. 625–631 (2005)
4. Wilson, T., Wiebe, J., Hwa, R.: 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 2004), San Jose, US, pp. 761–769 (2004)
5. Martin, J.R., White, P.R.: The Language of Evaluation: Appraisal in English. Palgrave, London (2005)
6. Bloom, K., Argamon, S., Garg, N.: Extracting appraisal expressions. In: Proceedings of NAACL-HLT 2007, pp. 308–315 (2007)
7. Esuli, A., Sebastiani, F.: Determining the semantic orientation of terms through gloss analysis. In: Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM 2005), Bremen, DE, pp. 617–624 (2005)
8. Argamon, S., Whitelaw, C., Chase, P., Hota, S., Garg, N., Levitan, S.: Stylistic Text Classification Using Functional Lexical Features. Journal of the American Society for Information Science and Technology 58(6), 802–822 (2007)
9. Osgood, C., Suci, G., Tannenbaum, P.: The measurement of meaning. University of Illinois Press, Urbana (1957)
10. Kamps, J., Marx, M.: Words with attitude. In: Proceedings of the 1st Global WordNet (GWC 2002) Conference, Mysore, IN, pp. 332–341 (2002)
11. Mullen, T., Collier, N.: Sentiment analysis using support vector machines with diverse information sources. In: Proceedings of the 9th Conference on Empirical Methods in Natural Language Processing (EMNLP 2004), Barcelona, ES, pp. 412–418 (2004)

12. Stone, P.J., Dunphy, D.C., Smith, M.S., Ogilvie, D.M.: The General Inquirer: A Computer Approach to Content Analysis. The MIT Press, Cambridge (1966)
13. Crammer, K., Singer, Y.: Pranking with ranking. In: Advances in Neural Information Processing Systems, vol. 14, pp. 641–647. MIT Press, Cambridge (2002)
14. Esuli, A., Fagni, T., Sebastiani, F.: Boosting multi-label hierarchical text categorization. Information Retrieval 11(4), 287–313 (2008)
15. Esuli, A., Sebastiani, F.: 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 2006), Trento, IT, pp. 193–200 (2006)
16. Hatzivassiloglou, V., McKeown, K.R.: Predicting the semantic orientation of adjectives. In: Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL 1997), Madrid, ES, pp. 174–181 (1997)
17. Takamura, H., Inui, T., Okumura, M.: Extracting emotional polarity of words using spin model. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005), Ann Arbor, US, pp. 133–140 (2005)
18. Kamps, J., Marx, M., Mokken, R.J., De Rijke, M.: Using WordNet to measure semantic orientation of adjectives. In: Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, PT, vol. IV, pp. 1115–1118 (2004)
19. Kim, S.M., Hovy, E.: Determining the sentiment of opinions. In: Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004), Geneva, CH, pp. 1367–1373 (2004)
20. Andreevskaia, A., Bergler, S.: 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 2006), Trento, IT, pp. 209–216 (2006)
21. Baroni, M., Vegnaduzzo, S.: 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 2004), Vienna, AU, pp. 17–24 (2004)
22. Riloff, E., Wiebe, J., Wilson, T.: Learning subjective nouns using extraction pattern bootstrapping. In: Proceedings of the 7th Conference on Natural Language Learning (CONLL 2003), Edmonton, CA, pp. 25–32 (2003)
23. Wiebe, J.: Learning subjective adjectives from corpora. In: Proceedings of the 17th Conference of the American Association for Artificial Intelligence (AAAI 2000), Austin, US, pp. 735–740 (2000)
24. Esuli, A., Sebastiani, F.: SentiWordNet: A publicly available lexical resource for opinion mining. In: Proceedings of the 5th Conference on Language Resources and Evaluation (LREC 2006), Genova, IT, pp. 417–422 (2006)
25. Kennedy, A., Inkpen, D.: Sentiment classification of movie reviews using contextual valence shifters. Computational Intelligence 22, 110–125 (2006)
26. Miyoshi, T., Nakagami, Y.: Sentiment classification of customer reviews on electric products. In: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, pp. 2028–2033 (2007)
27. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity in phrase-level sentiment analysis. In: HLT 2005: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, Morristown, NJ, USA, pp. 347–354. Association for Computational Linguistics (2005)