SentiWordNet: A High-Coverage Lexical Resource for Opinion Mining *

Andrea Esuli† and Fabrizio Sebastiani‡
Istituto di Scienza e Tecnologie dell'Informazione
Consiglio Nazionale delle Ricerche
Via Giuseppe Moruzzi 1 – 56124 Pisa, Italy

Technical Report 2007-TR-02
Istituto di Scienza e Tecnologie dell'Informazione
Consiglio Nazionale delle Ricerche
Pisa, IT

Abstract. *Opinion mining* (OM) is a recent subdiscipline at the crossroads of information retrieval and computational linguistics which is concerned not with the topic a document is about, but with the opinions it expresses. OM has a rich set of applications, ranging from tracking users' opinions about products or about political candidates as expressed in online forums, to customer relationship management. In order to aid the extraction of opinions from text, recent research has tried to automatically determine the "PN-polarity" of *subjective* terms, i.e. identify whether a term that indicates the presence of an opinion has a *positive* or a *negative* connotation. Research on determining the "SO-polarity" of terms, i.e. whether a term indeed indicates the presence of an opinion (a *subjective* term) or not (an *objective*, or *neutral* term) has been instead much scarcer.

In this paper we describe SentiWordNet, a lexical resource produced by asking an automated classifier $\hat{\Phi}$ to associate to each synset s of WordNet (version 2.0) a triplet of scores $\hat{\Phi}(s,p)$ (for $p \in P = \{\mathsf{Positive}, \mathsf{Negative}, \mathsf{Objective}\}$) describing how strongly the terms contained in s enjoy each of the three properties. The method used to develop SentiWordNet is based on the quantitative analysis of the glosses associated to synsets, and on the use of the resulting vectorial term representations for semi-supervised synset classification. The score triplet is derived by combining the results produced by a committee of eight ternary classifiers, all characterized by similar accuracy levels but extremely different classification behaviour. We present the results of evaluating the accuracy of the automatically assigned triplets on a publicly available benchmark. SentiWordNet is freely available for research purposes, and is endowed with a Web-based graphical user interface.

Keywords: Lexical resources, opinion mining, sentiment classification, gloss analysis, supervised learning

* This paper is a version, substantially revised and extended with new results, of (Esuli and Sebastiani, 2006b).

† E-mail: andrea.esuli@isti.cnr.it
‡ E-mail: fabrizio.sebastiani@isti.cnr.it

1. Introduction

Opinion mining (OM – also known as "sentiment classification") is a recent subdiscipline at the crossroads of information retrieval and computational linguistics which is concerned not with the topic a text is about, but with the opinions it expresses. Opinion-driven content management has several important applications, such as determining critics' opinions about a given product by classifying online product reviews, or tracking the shifting attitudes of the general public towards a political candidate by mining online forums or blogs. Within OM, several subtasks can be identified: for example,

1. *determining the SO-polarity of a text*, as in deciding whether a given text has a factual nature (i.e. describes a given situation or event, without expressing a positive or a negative opinion on it) or expresses an opinion on its subject matter. This amounts to performing binary text categorization under categories Subjective and Objective (Pang and Lee, 2004; Yu and Hatzivassiloglou, 2003);

2. *determining the PN-polarity of a text*, as in deciding if a given Subjective text expresses a Positive or a Negative opinion on its subject matter (Pang and Lee, 2004; Turney, 2002);

3. *determining the strength of the PN-polarity of a text*, as in deciding e.g., whether the Positive opinion expressed by a text on its subject matter is Weakly Positive, Mildly Positive, or Strongly Positive (Pang and Lee, 2005; Wilson et al., 2004);

4. *extracting opinions from a text*, as in determining whether a given linguistic expression within a text conveys an opinion or not, and (if positive) determining who holds this opinion, who or what is the object of this opinion, and what type of opinion it is (Kim and Hovy, 2006).

To aid these tasks, several researchers have attempted to automatically determine whether a *term* that indicates the presence of an opinion has a Positive or a Negative connotation (Esuli and Sebastiani, 2005; Hatzivassiloglou and McKeown, 1997; Kamps et al., 2004; Kim and Hovy, 2004; Takamura et al., 2005; Turney and Littman, 2003), since considering the contribution of these terms is helpful in order to solve Tasks 1–4. The conceptually simplest approach to this latter problem is probably Turney's (2002), who has obtained interesting results on Task 2 by considering the algebraic sum of the PN-polarities of terms as representative of the PN-polarity of the document they belong to; but

more sophisticated approaches have also been taken (Hatzivassiloglou and Wiebe, 2000; Riloff et al., 2003; Whitelaw et al., 2005; Wilson et al., 2004). The task of determining whether a term indeed indicates the presence of an opinion (i.e. is Subjective or Objective) has instead received much less attention (Esuli and Sebastiani, 2006a; Riloff et al., 2003; Vegnaduzzo, 2004).

Note that in these works no distinction between different senses of a word is attempted, so that the term, and not its senses, are classified (although some such works (Hatzivassiloglou and McKeown, 1997; Kamps et al., 2004) distinguish between different POSs of a word).

1.1. Our proposal

In this paper we describe SentiWordNet, a lexical resource produced by asking an automated classifier $\hat{\Phi}$ to associate to each synset s of WordNet (version 2.0), a triplet of numerical scores $\hat{\Phi}(s,p)$ (for $p \in P$ ={Positive, Negative,Objective}) describing how strongly the terms contained in s enjoy each of the three properties[1].

A WordNet synset represents a unique sense, which is defined by a unique gloss and is associated to a set of terms all with the same POS, each one associated to a sense number (e.g., the adjectives `blasphemous(2)`, `blue(4)`, `profane(1)` are all contained in the same synset, whose sense is defined by the gloss "`characterized by profanity or cursing`"). The assumption that underlies our switch from terms to synsets is that different senses of the same term may have different opinion-related properties.

Each of the three $\hat{\Phi}(s,p)$ scores ranges from 0.0 to 1.0, and their sum is 1.0 for each synset s. This means that a synset may have nonzero scores for all of the three categories, which would indicate that the corresponding terms have, in the sense indicated by the synset, each of the three opinion-related properties only to a certain degree[2]. For exam-

[1] Consistently with most mathematical literature we use the caret symbol (ˆ) to indicate estimation. In fact, a classifier $\hat{\Phi}(s,p)$ has to be understood as an estimation, or approximation, of an unknown "target function" (or "gold standard") $\Phi(s,p)$.

[2] Note that associating a graded score to a synset for a certain property (e.g., Positive) may have (at least) three different interpretations: (i) the terms in the synset are Positive only to a certain degree; (ii) the terms in the synset are sometimes used in a Positive sense and sometimes not, e.g., depending on the context of use; (iii) the annotator is uncertain whether the terms in the synset are Positive. Interpretation (i) has a *fuzzy* character, implying that each instance of these terms, in each context of use, has the property to a certain degree, and that the annotator is certain of this degree. Interpretation (ii) has a *probabilistic* nature (of a frequentistic, "objective" type), implying that membership of a synset in the set denoted by the property must be computed by counting the number of contexts of use in which the terms

ple, SentiWordNet attributes to the synset [estimable(3)][3], corresponding to the sense "may be computed or estimated" of the adjective estimable, an Objective score of 1.0 (and Positive and Negative scores of 0.0), while it attributes to the synset [estimable(1)], corresponding to the sense "deserving of respect or high regard", a Positive score of 0.75, a Negative score of 0.0, and an Objective score of 0.25 (see Figure 2).

A similar intuition had previously been presented in (Kim and Hovy, 2004), whereby a term could have both a Positive and a Negative PN-polarity, each to a certain degree; however, (Kim and Hovy, 2004) deal with terms, and not with their senses. Non-binary scores are attached to opinion-related properties also in (Turney and Littman, 2003); the authors' interpretation of these scores is related to the confidence in the correctness of the labelling, rather than in how strongly the term is deemed to possess the property. A related point has recently been made in (Andreevskaia and Bergler, 2006a), in which terms that possess a given opinion-related property to a higher degree are claimed to be also the ones on which human annotators asked to assign this property agree more.

We believe that a graded (as opposed to binary) evaluation of the opinion-related properties of terms can be helpful in the development of opinion mining applications. A binary classification method will probably label as Objective any term that has no strong SO-polarity, e.g., terms such as short or alone. If a sentence contains many such terms, a resource based on a binary classification will probably miss its subtly subjective character, while a graded lexical resource like SentiWordNet may provide enough information to capture such nuances. SentiWordNet is freely available for research purposes, and is endowed with a Web-based graphical user interface.

The method we have used to develop SentiWordNet is based on our previous work on determining the opinion-related properties of *terms* (Esuli and Sebastiani, 2005; Esuli and Sebastiani, 2006a). The method relies on the quantitative analysis of the *glosses* associated to synsets, and on the use of the resulting vectorial term representa-

have the property. Interpretation (iii) has, again, a *probabilistic* nature, but of a "subjective" type, i.e. related to the degree of confidence that the annotator has in the membership of the synset in the set denoted by the property. We do not attempt to take a stand on this distinction, which (to our knowledge) had never been raised in sentiment analysis and that requires an in-depth linguistic study, but we tend to believe that the interpretation embodied in SentiWordNet may be seen as a combination of (i-iii).

[3] We here adopt the standard convention according to which a term enclosed in square brackets denotes a synset; thus [poor(7)] refers not just to the term poor but to the synset consisting of {inadequate(2), poor(7), short(4)}.

tions for semi-supervised synset classification. The triplet of scores is derived by combining the results produced by a committee of eight ternary classifiers, each of which has demonstrated, in our previous tests, similar accuracy but different characteristics in terms of classification behaviour. Two versions of SentiWordNet are discussed and evaluated in this paper, which are obtained by two different methods of generating the eight classifiers and combining their results.

The rest of the paper is organized as follows. Section 2 describes the semi-supervised learning method by which SentiWordNet was built, describing how the classifiers were trained (Section 2.1) and combined (Section 2.2). Section 3 describes the results of an evaluation exercise by which we attempt to estimate the accuracy with which synsets were automatically tagged. Section 4 discusses related work, while Section 5 concludes, pointing at some avenues for future research.

2. Building SentiWordNet

The method we have used to develop SentiWordNet relies on automatically training eight individual *synset classifiers* $\hat{\Phi}_1(s,p)$, ..., $\hat{\Phi}_8(s,p)$, and then gathering them into a *(synset) classifier committee* $\hat{\Phi}(s,p)$.

Synset classifiers (be them individual classifiers or classifier committees) are ternary classifiers, i.e., they attempt to predict whether a synset is Positive, Negative, or Objective. By an *n-ary classifier* we here mean a function $\hat{\Phi} : S \times P \rightarrow [0,1]$ that, given an object s and a class $p \in P = \{p_1, \ldots, p_n\}$, returns a numerical score $\hat{\Phi}(s,p)$.

Scores can be binary-valued or real-valued. In the former case, $\hat{\Phi}(s,p)$ must equal 1 for a single $p_i \in P$ and 0 for all $p \in P/\{p_i\}$; this corresponds to deciding that s belongs to class p_i and does not belong to any class in $P/\{p_i\}$.

In the latter case $\hat{\Phi}(s,p)$ denotes the confidence, or degree of belief, that the classifier has in the fact that s has indeed property p (the higher the value, the higher the confidence). If a binary decision needs to be taken, synset s is deemed to belong to the class

$$\arg\max_{p} \hat{\Phi}(s,p)$$

that has received the highest score.

Section 2.1 will deal with the method we have used for training the individual $\hat{\Phi}_i$'s, while Section 2.2 will discuss the issue of building a classifier committee $\hat{\Phi}$ out of them.

2.1. Training synset classifiers

The method we have used to develop the individual classifiers $\hat{\Phi}_1, ..., \hat{\Phi}_8$ is an adaptation to synset classification of our semi-supervised method for deciding the PN-polarity (Esuli and Sebastiani, 2005) and SO-polarity (Esuli and Sebastiani, 2006a) of *terms*. A *semi-supervised* method (see e.g. (Nigam et al., 2000)) is a learning process whereby only a small subset $L \subset Tr$ of the training data Tr have been *manually* labelled. In origin the training data in $U = Tr - L$ were instead unlabelled; it is the process itself that has labelled them, automatically, by using L (with the possible addition of other available resources) as input. A semi-supervised method thus trades accuracy for reduced costs: while supervised methods guarantee higher accuracy when many training examples are available, semi-supervised methods attempt to bring about reasonable accuracy by making the most of unlabelled examples too, when the labelled ones are few or are too expensive to collect.

Our method defines L as the union of three *seed* (i.e. training) sets L_p, L_n and L_o of known Positive, Negative and Objective synsets, respectively.

L_p and L_n are two small sets, which we have defined by manually selecting the intended synsets[4] for 14 "paradigmatic" Positive and Negative terms (e.g., the Positive term `nice`, the Negative term `nasty`) which were used as seed terms in (Turney and Littman, 2003). The process has resulted in 47 Positive and 58 Negative synsets. L_p and L_n are then iteratively expanded, in K iterations, into the final training sets Tr_p^K and Tr_n^K. At each iteration step k two sets Tr_p^k and Tr_n^k are generated, where $Tr_p^k \supset Tr_p^{k-1} \supset \ldots \supset Tr_p^1 = L_p$ and $Tr_n^k \supset Tr_n^{k-1} \supset \ldots \supset Tr_n^1 = L_n$. The expansion at iteration step k consists

1. in adding to Tr_p^k (resp. Tr_n^k) all the synsets that are connected to synsets in Tr_p^{k-1} (resp. Tr_n^{k-1}) by WordNet lexical relations (e.g., *also-see*) such that the two related synsets can be assumed to have *the same* PN-polarity;

2. in adding to Tr_p^k (resp. Tr_n^k) all the synsets that are connected to synsets in Tr_n^{k-1} (resp. Tr_p^{k-1}) by WordNet lexical relations (e.g., *direct antonymy*) such that the two related synsets can be assumed to have *opposite* PN-polarity.

The relations we have used in (Esuli and Sebastiani, 2005; Esuli and Sebastiani, 2006a) are synonymy (for use in substep 1) and direct

[4] For example, for the term `nice` we have removed the synset relative to the French city of Nice.

antonymy (for use in substep 2) between terms, as is common in related literature (Kamps et al., 2004; Kim and Hovy, 2004; Takamura et al., 2005). In the case of synsets, synonymy cannot be used because it is the relation that defines synsets, thus it does not connect different synsets. We have then followed the method used in (Valitutti et al., 2004) for the development of WordNet-Affect, a lexical resource that tags WordNet synsets by means of a taxonomy of affective categories (e.g. Behaviour, Personality, CognitiveState): after hand-collecting a number of labelled terms from other resources, Valitutti and colleagues generate WordNet-Affect by adding to them the synsets reachable by navigating the relations of *direct antonymy*, *similarity*, *derived-from*, *pertains-to*, *attribute*, and *also-see*, which they consider to reliably preserve/invert the involved labels. Given the similarity with our task, we have used exactly these relations in our expansion. The final sets Tr_p^K and Tr_n^K, along with the set Tr_o^K described below, are used to train the ternary classifiers.

The L_o set is treated differently from L_p and L_n, because of the inherently "complementary" nature of the Objective category (an Objective term can be defined as a term that does *not* have either Positive or Negative characteristics). We have heuristically defined L_o as the set of synsets that (a) do not belong to either Tr_p^K or Tr_n^K, and (b) contain terms not marked as either Positive or Negative in the General Inquirer lexicon (Stone et al., 1966); this lexicon was chosen since it is, to our knowledge, the largest manually annotated lexicon in which terms are tagged according to the Positive or Negative categories[5]. The resulting L_o set consists of 17,530 synsets; for any K, we define Tr_o^K to coincide with L_o.

We give each synset a vectorial representation, obtained by applying a standard text indexing technique (cosine-normalized $tf * idf$ preceded by stop word removal) to its gloss, which we thus take to be a textual representation of its semantics. Our basic assumption is that synsets with similar polarity tend to have "similar" glosses: for instance, that the glosses for synsets {`good(6)`, `fine(1)`} and {`pleasure(2)`, `joy(2)`, `delight(2)`} will both contain appreciative expressions, while the glosses for synsets {`badly(2)`, `poorly(1)`, `ill(1)`} and {`improper(3)`, `unsuitable(5)`, `wrong(3)`} will both contain derogative expressions.

In Section 2.2 we discuss two different methods (called Combination Method A and Combination Method B) by which we combine the individual classifiers $\hat{\Phi}_1(s,p)$, ..., $\hat{\Phi}_8(s,p)$ into a committee $\hat{\Phi}(s,p)$.

[5] As a consequence, the General Inquirer is the *de facto* benchmark in the literature on classifying terms according to their opinion-related properties; see e.g. (Esuli and Sebastiani, 2005; Esuli and Sebastiani, 2006a; Turney and Littman, 2003).

Combination Method A requires the $\hat{\Phi}_i$'s to output binary scores, while Combination Method B requires them to output real-valued scores. As a consequence, we use the vectorial representations in two different ways[6], dubbed Learning Method A and Learning Method B, according to whether Combination Method A or Combination Method B are going to be used:

1. In Learning Method A, the $\hat{\Phi}_i$'s are obtained by means of supervised learners that generate binary classifiers. The vectorial representations of the training synsets are input to a supervised learner which generates two binary classifiers $\hat{\Phi}_i^p$ and $\hat{\Phi}_i^n$: $\hat{\Phi}_i^p$ must discriminate between terms that belong to the Positive category and ones that belong to its complement (**not** Positive), while $\hat{\Phi}_i^n$ must discriminate between terms that belong to the Negative category and ones that belong to its complement (**not** Negative).

   In the training phase, the terms in $Tr_n^K \cup Tr_o^K$ are used as training examples of category (**not** Positive), and the terms in $Tr_p^K \cup Tr_o^K$ are used as training examples of category (**not** Negative).

   Terms that have been classified *both* into Positive by the $\hat{\Phi}_i^p$ and into (**not** Negative) by $\hat{\Phi}_i^n$ are deemed to be positive, and terms that have been classified *both* into (**not** Positive) by $\hat{\Phi}_i^p$ and into Negative by $\hat{\Phi}_i^n$ are deemed to be negative. The terms that have been classified (i) into both (**not** Positive) and (**not** Negative), or (ii) into both Positive and Negative, are taken to be Objective.

   The two binary classifiers $\hat{\Phi}_i^p$ and $\hat{\Phi}_i^n$ working together thus implement, as is often the case in the supervised learning literature, a ternary classifier $\hat{\Phi}_i$ which returns a triplet of binary scores for the $p \in P$.

2. In Learning Method B, the $\hat{\Phi}_i$'s are obtained by means of supervised learners that directly generate n-ary classifiers, where the resulting classifiers return a triplet of real-valued scores for the $p \in P$. In the training phase, the terms in Tr_p^K, Tr_n^K, and Tr_o^K are used as positive examples of Positive, Negative, and Objective, respectively.

The resulting classifiers are then applied to the vectorial representations of all WordNet synsets s (including those in $Tr^K - L$).

The main difference between Learning Methods A and B is that in Learning Method B Objective is seen as a category, or concept, in its own right, while in Learning Method A objectivity is viewed as

[6] These two different ways are called Approach II and Approach III in (Esuli and Sebastiani, 2006a).

a nonexistent entity, i.e. as the "absence of subjectivity" (in fact, in Learning Method A the training examples of Objective are only used as training examples of the *complements* of Positive and Negative).

Note also that, while for Learning Method A we can use well-known learners for binary classification (support vector machines using linear kernels, and the Rocchio learner), for Learning Method B we need to use their n-ary versions[7].

Note that other approaches to learning a ternary classifier are possible; for instance, (Esuli and Sebastiani, 2006a) test three different approaches, of which the present ones are dubbed Approaches II and III (Approach I consists in a different way of combining binary classification technology). Out of the three we have chosen Approach II since it is the one that, in the experiments of (Esuli and Sebastiani, 2006a), yielded the best effectiveness, and Approach III since for Combination Method B we needed the $\hat{\Phi}_i$ to output non-binary scores for each $p \in P$.

2.2. Defining the committee of classifiers

In (Esuli and Sebastiani, 2006a) we point out how different combinations of training set and learner behave in a radically different way, even though with similar levels of accuracy. The main three observations we recall here are the following:

- Low values of K produce small training sets for Positive and Negative, which produces binary classifiers with low recall and high precision for these categories. By increasing K these sets get larger, and the effect of these larger numbers is to increase recall but to also add "noise" to the training set, which decreases precision.

- Learners that use information about the prior probabilities of categories (e.g., SVMs), which estimate these probabilities from the training sets, are sensitive to the relative cardinalities of the training sets, and tend to classify more items into the categories that have more positive training items. Learners that do not use this information, like Rocchio, do not exhibit this kind of behaviour.

- The difference in behaviour mentioned above does not affect the overall accuracy of the method, but only the relative proportions

[7] The Rocchio learner we have used is from Andrew McCallum's *Bow* package (`http://www-2.cs.cmu.edu/~mccallum/bow/`), while the SVMs learner we have used is Thorsten Joachims' SVM^{light} (`http://svmlight.joachims.org/`), version 6.01. Both packages allow the respective learners to be run in n-ary (aka "multiclass") fashion.

of items classified as Positive ∪ Negative and items classified as Objective, while the accuracy in discriminating between Positive and Negative items tends to be constant.

It is a well-known fact of computational learning theory that the more independent from each other a set of classifiers are, the better they perform once assembled into a committee (Tumer and Ghosh, 1996). Since the above-mentioned difference in behaviour among our classifiers is a witness of their independence, we have decided to combine different configurations of training set and learner into a committee.

Specifically, we have defined four different training sets, by choosing four different values of K (0, 2, 4, 6), and we have alternatively used two learners (Rocchio and SVMs); this yields a total of eight ternary classifiers. With $K = 0$, SVMs produced very "conservative" binary classifiers for Positive and Negative, i.e. classifiers characterized by very low recall and high precision. For $K = 6$, SVMs produced instead very "liberal" binary classifiers for Positive and Negative, i.e. classifiers that tend to classify many synsets as Positive or Negative even in the presence of very little evidence of subjectivity. The Rocchio learner has a similar behaviour, although not dependent on the prior probabilities of categories.

As mentioned above, we experiment with two different combination methods for computing the final triplets of $\hat{\Phi}(s, p)$ scores:

– In Combination Method A, we use ternary classifiers $\hat{\Phi}_i$ generated by Learning Method A, which thus return a triplet of binary scores; the final scores $\hat{\Phi}(s, p)$ are determined by the (normalized) proportion of ternary classifiers that have assigned the corresponding label to s, i.e.,

$$\hat{\Phi}(s, p) = \frac{1}{8} \sum_{i=1}^{8} [\![\hat{\Phi}_i(s) = p]\!] \tag{1}$$

where $[\![\pi]\!]$ indicates the characteristic function of predicate π (i.e. the function that returns 1 if π is true and 0 otherwise). If all the $\hat{\Phi}_i$'s agree in assigning the same label to a synset s, that label will have a score of 1.0 for s, otherwise each label will have a score proportional to the number of classifiers that have assigned it.

– In Combination Method B, we use ternary classifiers $\hat{\Phi}_i$ generated by Learning Method B, which thus return a triplet of real-valued scores; the final scores $\hat{\Phi}(s, p)$ are obtained by simply adding the corresponding real-valued scores from the $\hat{\Phi}_i$'s and then normaliz-

ing them, i.e.,

$$\hat{\Phi}(s,p) = \frac{\sum_{i=1}^{8} \hat{\Phi}_i(s,p)}{\sum_{p \in P} \sum_{i=1}^{8} \hat{\Phi}_i(s,p)} \quad (2)$$

Note that Combination Method B is "finer-grained" than Combination Method A, since the scores produced by Equation 1 range on the discrete set $\{0, \frac{1}{8}, \ldots, \frac{7}{8}, 1\}$, while the scores produced by Equation 2 range on the full real-valued [0,1] interval. Note also that, while Combination Method A only brings to bear the binary decisions of the individual classifiers $\hat{\Phi}_i$, Combination Method B also brings to bear the real-valued scores $\hat{\Phi}_i(s,p)$ that these classifiers have produced, i.e., the degrees of confidence that the $\hat{\Phi}_i$'s have in the correctness of their binary decisions. All in all, Combination Method B seems *a priori* conceptually more interesting than Combination Method A; we will experimentally evaluate them in Section 3.

Hereafter, by SentiWordNet 1.0 (resp. SentiWordNet 1.1) we will denote the result of classifying WordNet according to Learning and Combination Methods A (resp. B)[8].

2.3. Visualizing SentiWordNet

Given that the sum of scores in a triplet is always 1.0, it is possible to display this triplet in a triangle whose vertices correspond to a 1.0 score for one $p \in P$ and a 0.0 score for the other two. Figure 2.3 shows the graphical model we have designed to display the score triplet associated to a synset. This model is used in the Web-based graphical user interface through which SentiWordNet can be accessed at `http://patty.isti.cnr.it/~esuli/software/SentiWordNet`. Figures 2 and 3 show two screenshots of the output for the synsets that include the terms `estimable` and `short`.

3. Results

How reliable are the opinion-related scores attached to synsets in SentiWordNet? Fully testing the accuracy of our annotation method

[8] This naming convention is due to the fact that the first version of SentiWordNet we have discussed in the literature and publicly released was SentiWordNet 1.0; at that time, SentiWordNet 1.1 had not been developed yet. The example in Section 1.1 is drawn from SentiWordNet 1.0.

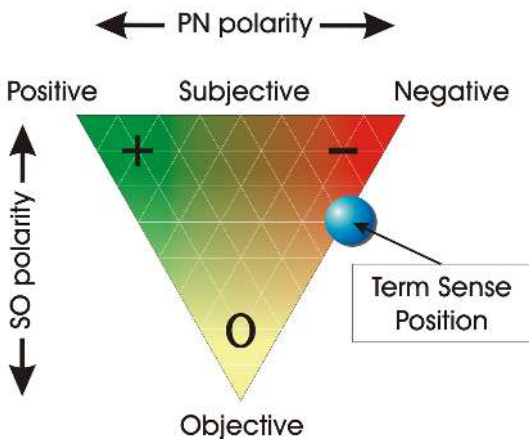

Figure 1. The graphical representation adopted by SentiWordNet for representing the opinion-related properties of a synset.

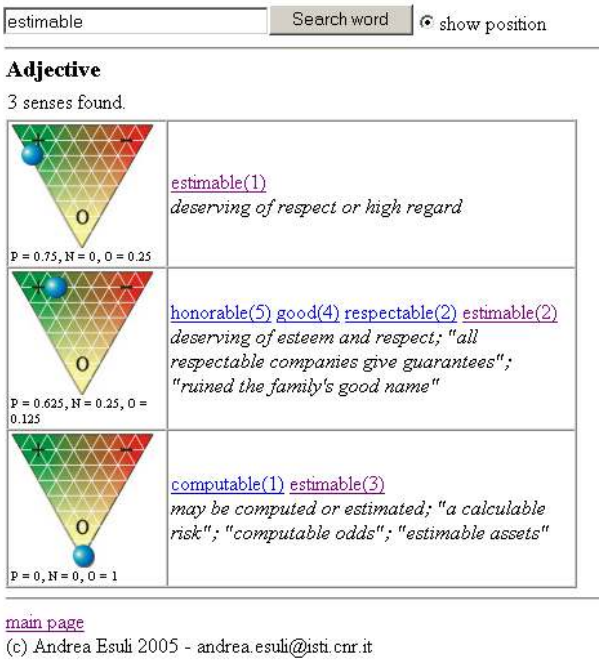

Figure 2. SentiWordNet visualization of the opinion-related properties of the synsets that include the term `estimable` (actual scores are from SentiWordNet 1.0).

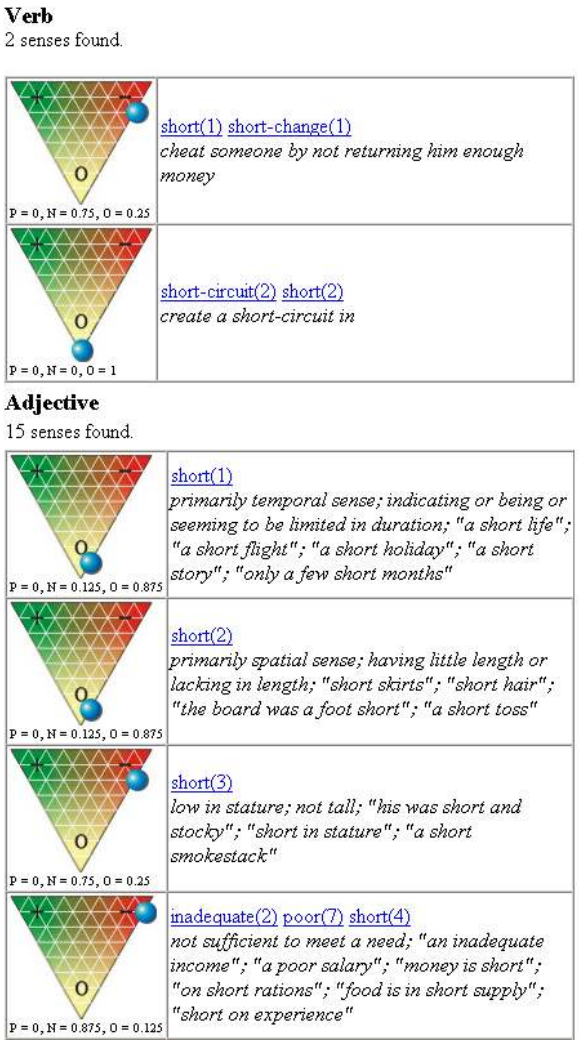

Figure 3. SentiWordNet visualization of the opinion-related properties of the synsets that include the term `short` (actual scores are from SentiWordNet 1.0).

experimentally is impossible, since for this we would need a version of WordNet manually annotated according to our three properties of interest, and the unavailability of such a manually annotated resource is exactly the reason why we are interested in generating it automatically.

A first, approximate indication of the quality of SentiWordNet can be gleaned by looking at the accuracy obtained by our method in classifying the General Inquirer (Stone et al., 1966), a lexicon which is instead fully tagged according to the three properties we have been discussing; the results of this classification exercise are reported in (Esuli

and Sebastiani, 2006a). The reader should however bear in mind a few differences between the method used in (Esuli and Sebastiani, 2006a) and the one used here: (i) we here classify entire synsets, while in (Esuli and Sebastiani, 2006a) we classified terms, which can sometimes be ambiguous, and can thus be more difficult to classify correctly; (ii) as discussed in Section 2.1, the WordNet lexical relations used for the expansion of the training set are different. The effectiveness results reported in (Esuli and Sebastiani, 2006a) may thus be considered only approximately indicative of the accuracy of the SentiWordNet labels.

3.1. The Micro-WNOp gold standard

A second, more direct route to evaluating SentiWordNet is by using a manually annotated *subset* of WordNet as a "gold standard" against which to evaluate the scores attached to the same synsets in SentiWordNet. A subset of this kind, called Micro-WNOp, indeed exists (Cerini et al., 2007): it consists of 1105 synsets manually annotated by a group of five human annotators (hereafter called J1, ..., J5); each synset is assigned a score for each of the three categories Positive, Negative, and Objective, with the scores in the triplet summing up to 1 for each synset. Synsets 1-110 (here dubbed Micro-WNOp(1)) were tagged by all the annotators working together, so as to develop a common understanding of the semantics of the three categories; then, J1, J2 and J3 independently tagged each synsets 111–606 (Micro-WNOp(2)), while J4 and J5 independently tagged synsets 607–1105 (Micro-WNOp(3)).

It is also noteworthy that Micro-WNOp as a whole, and each of its subsets, are representative of the distribution of parts of speech in WordNet: this means that, e.g., if x% of WordNet synsets are nouns, also x% of Micro-WNOp synsets (and of Micro-WNOp(1) synsets, and ...) are nouns.

The Web-based graphical user interface that was used by the annotators is based on the same graphical model as discussed in Section 2.3. In this interface each annotator was presented with a synset and was asked to place a bullet within the triangle in the position that represented, according to him/her, the mix of the three opinion-related properties as possessed by the synset.

See (Cerini et al., 2007) for further details on how Micro-WNOp and its subsets were designed.

Note that 1,005 synsets correspond to less than 1% of the total 115,000 WordNet synsets; this clarifies that, again, the accuracy obtained on this gold standard may be considered only indicative of the

(unknown) level of accuracy with which SentiWordNet has been produced. Notwithstanding this, Micro-WNOp will prove a useful tool in the comparative evaluation of future systems that, like ours, tag WordNet synsets by opinion, including possible future releases of SentiWordNet.

3.2. Evaluating SentiWordNet

For the evaluation of our experiments we have defined an effectiveness measure based on the graphical model presented in Section 2.3. Specifically, given the score triplet $\hat{\Phi}(s)$=($\hat{\Phi}(s, \mathsf{Positive})$, $\hat{\Phi}(s, \mathsf{Negative})$, $\hat{\Phi}(s, \mathsf{Objective})$) assigned to s by SentiWordNet, given the analogous score triplet $\Phi(s)$ assigned to s by Micro-WNOp, and assuming that the length of each side of the triangle is 1, we compute the reciprocal of the normalized Euclidean distance

$$\Psi(s) = 1 - \frac{D(\hat{\Phi}(s), \Phi(s))}{\sqrt{2}}$$

between the two points in the triangle representing the score triplets, and use it as our effectiveness measure. This measure is maximized (i.e., it equals 1) when $\hat{\Phi}(s)$ and $\Phi(s)$ occupy the same point in the triangle, and is minimized (i.e., it equals 0) when they are as far apart as possible in the triangle (this happens when one of them is on a vertex and the other is on the mid point of the opposite side of the triangle[9]).

We measure effectiveness for multiple synsets in terms of *agreement as a function of maximum distance*; that is, we compute the percentage of synsets on which $\hat{\Phi}$ and Φ agree "up to a certain distance", i.e.

$$A(x) = \frac{|s \in S \ : \ \Psi(s) \leq x|}{|S|}$$

where S is the set of synsets on which the evaluation is conducted and $|\cdot|$ indicates the cardinality of a set. Note that $A(x)$ is a monotonically increasing function, and that $A(1) = 1$ by definition.

Figures 4 to 6 plot the $A(x)$ measure for both SentiWordNet 1.0 and 1.1 against the three subsets of Micro-WNOp. Since the two subsets of Micro-WNOp referred to in Figures 5 to 6 were independently annotated by more than one annotator, these two figures also plot the agreement between each pair of different annotators, also measured by the $A(x)$ measure[10]. Concerning the suitability of $A(x)$ to serve as a

[9] The use of $\sqrt{2}$ as normalization factor depends on the fact that, in this latter case, the non-normalized distance is $\sqrt{2}$.

[10] The reader may notice that, while the curves representing SentiWordNet 1.0 and 1.1 are fairly smooth, the ones representing the agreement between human

measure of interannotator agreement, note that $A(x)$ is a symmetric measure, i.e., the agreement between annotator J_x and annotator J_y is the same as the agreement between J_y and J_x; that is, it does not matter who among J_x and J_y plays the "gold standard".

Looking at these figures we may observe that the effectiveness values of SentiWordNet 1.0 and 1.1 are very close to each other. SentiWordNet 1.1 seems decidedly better for small values of x (especially on Micro-WNOp(1)) while SentiWordNet 1.0 seems better for larger values of x. Indeed, the small values of x are the most important, since agreement for large values of x tends to be scarcely significant from an application point of view. Also, Micro-WNOp(1) can arguably be considered (despite its small size) the most reliable of the three subsets of Micro-WNOp, since the annotations are the result of an agreement between all five annotators. As a result, we can consider SentiWordNet 1.1 a more accurate lexicon than SentiWordNet 1.0; this confirms our expectations (as expressed at the end of Section 2.2) that Combination Method B, by means of which SentiWordNet 1.1 has been generated, would turn out a clearly superior alternative to Combination Method A, by means of which SentiWordNet 1.0 has been generated.

Figure 7 looks at another aspect, namely, at how accurate SentiWordNet is on synsets belonging to different parts of speech: each curve plots, for each set of Micro-WNOp(2) synsets belonging to a given POS, the agreement (as usual, expressed in terms of $A(x)$) between (a) the average of SentiWordNet 1.0 and 1.1, and (b) the average of the three human annotators who annotated Micro-WNOp(2) (results for Micro-WNOp(3) are analogous and will thus not be reported). The plots clearly show that not all parts of speech are equally difficult: adjectives and verbs seem decidedly more difficult to classify than adverbs and nouns. That this is the case is shown by the fact that also the human annotators tend to agree less among themselves when annotating adjectives and verbs than adverbs and nouns; on this see Figure 8, which reports the agreement between annotators J1 and J2 on different parts of speech for Micro-WNOp(2) (results on the agreement between J2 and J3, J1 and J3, as well as the agreement between J4 and J5 on Micro-WNOp(3), are analogous and will thus not be reported).

---

annotators are stairs-shaped. The reason for this is that in the graphical user interface used by the annotators there is a small, grid-shaped, finite number of points within the triangle that the annotators can place a synset on. As a consequence, the distance between the points chosen by two different annotators for the same synset can range only on a small, finite number of values.

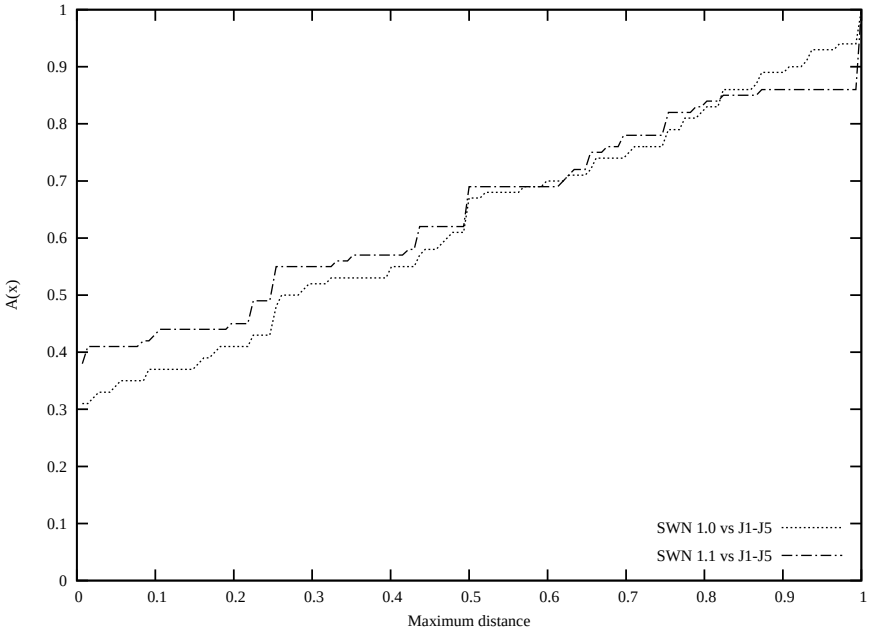

Figure 4. Agreement (on MICRO-WNOP(1)) as a function of maximum distance between score triplets. Agreement between the five human coders and SENTIWORDNET (versions 1.0 and 1.1), are reported.

3.3. SOME STATISTICS

Tables I and II show some statistics about the distribution of scores in SENTIWORDNET 1.0. Note that each table has nine rows, corresponding to the nine different values $\{0, \frac{1}{8}, \ldots, \frac{7}{8}, 1\}$ on which the $\Phi(s,p)$ classifier that generated SENTIWORDNET 1.0 ranges.

The first remarkable fact is that the synsets judged to have some degree of opinion-related properties (i.e. not fully Objective) are a considerable part of the whole WORDNET, i.e. 24.63% of it. However, as the objectivity score decreases, indicating a stronger subjectivity score (either as Positive, or as Negative, or as a combination of them), the number of the synsets involved decreases rapidly, from 10.45% for $Obj(s) <= 0.5$, to 0.56% for $Obj(s) <= 0.125$. This seems to indicate that there are only few terms that are unquestionably Positive or Negative, where "unquestionably" here indicates widespread agreement among different automated classifiers; in essence, this is the same observation which has independently been made in (Andreevskaia and Bergler, 2006a), where agreement among human classifiers is shown to correlate strongly with agreement among automated classifiers, and

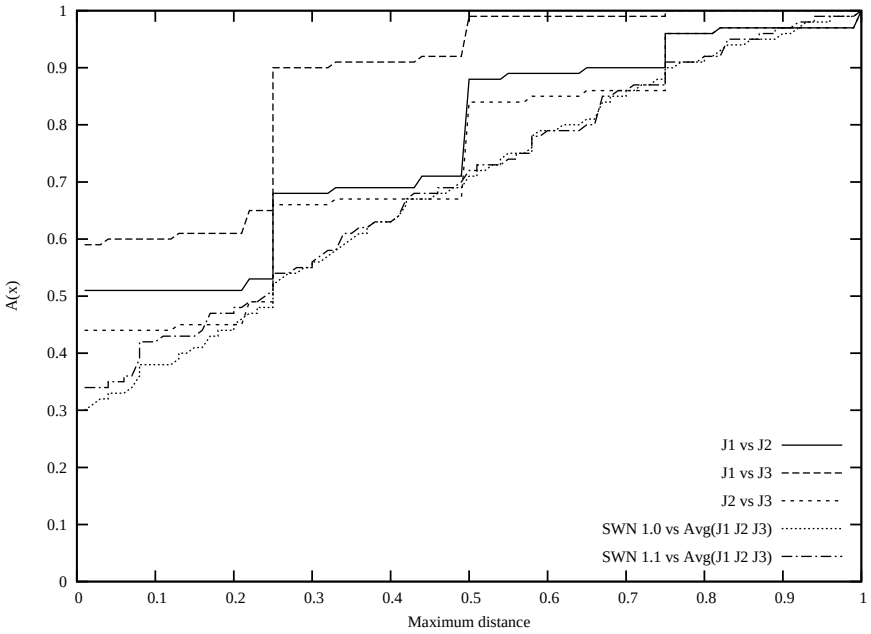

Figure 5. Agreement (on MICRO-WNOP(2)) as a function of maximum distance between score triplets. Agreement (1) between different human coders (J1, J2, and J3) and (2) between average of the three human coders and SENTIWORDNET (versions 1.0 and 1.1), are reported.

where such agreement is strong only for a small subset of "core", strongly-marked terms.

Table I reports a breakdown by POS of the scores obtained by synsets. It is quite evident that "adverb" and "adjective" synsets are evaluated as (at least partially) Subjective (i.e. $Obj(s) < 1$) much more frequently (39.66% and 35.7% of the cases, respectively) than "verb" (11.04%) or "noun" synsets (9.98%). This fact seems to indicate that, in natural language, opinions are most often conveyed by parts of speech used as modifiers (i.e. adverbs, adjectives) rather than parts of speech used as heads (i.e. verbs, nouns), as exemplified by expressions such as `a disastrous appearance` or `a fabulous game`. This intuition might be rephrased by saying that the most frequent role of heads is to denote entities or events, while that of modifiers is (among other things) to express a judgment of merit on them.

4. Related work

The only work we are aware of that deals with tagging synsets by PO- and SN-polarity is the very recent (Andreevskaia and Bergler,

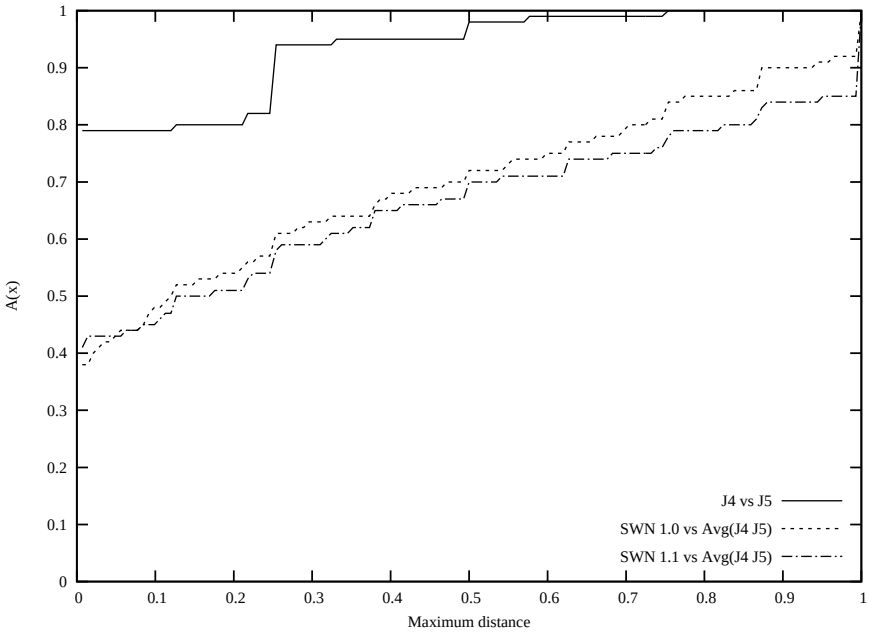

Figure 6. Agreement (on MICRO-WNOP(3)) as a function of maximum distance between score triplets. Agreement (1) between different human coders (J4 and J5) and (2) between average of the two human coders and SENTIWORDNET (versions 1.0 and 1.1), are reported.

2006b), which is based on the crude idea of tagging with a category $p \in \{\mathsf{Positive}, \mathsf{Negative}, \mathsf{Objective}\}$ all synsets whose EXTENDEDWORDNET gloss[11] contains (i) a synset that is known to belong to p, or (ii) a synset that is reachable from synsets belonging to p via WORDNET lexical relations that (similarly to what we do in Section 2.1) can be assumed to preserve opinion-related properties. However, there are key differences between (Andreevskaia and Bergler, 2006b) and our work. First, those authors limit their work to WORDNET adjectives, while we tag all WORDNET synsets, irrespectively of their POS; arguably, words other than adjectives are the hardest to work with, since they tend to be sentiment-laden to a much smaller degree than adjectives (see Table I). Second, the system of (Andreevskaia and Bergler, 2006b) tags synsets as either belonging or not belonging to a category p, while in our system membership is a matter of degrees. Last, (Andreevskaia and Bergler, 2006b) do not evaluate the accuracy of their system at

[11] EXTENDEDWORDNET (Harabagiu et al., 1999) is a version of WORDNET in which, among other things, all terms appearing in the gloss of a synset are (automatically) disambiguated, and are thus linked to the synset they pertain to.

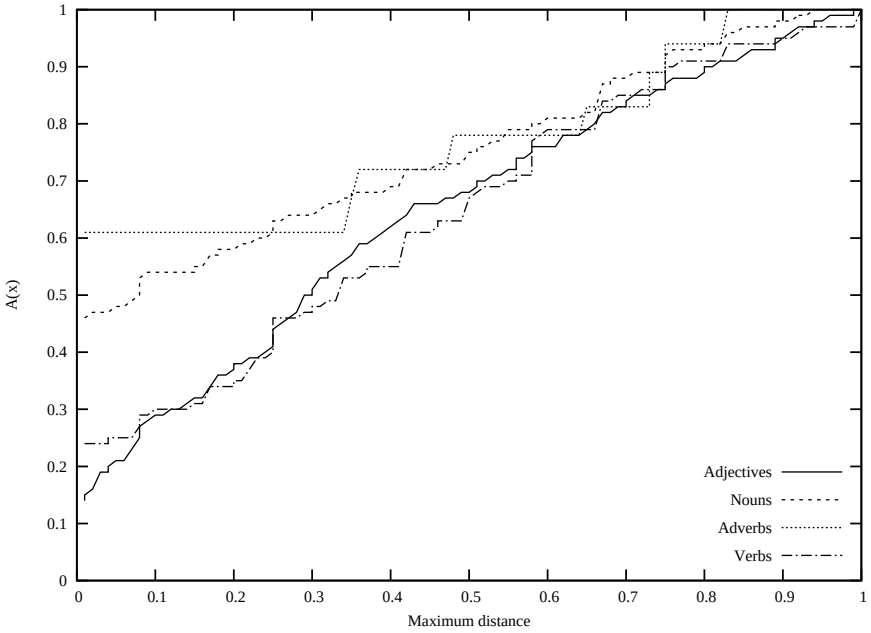

Figure 7. Agreement (on MICRO-WNOP(2)) as a function of maximum distance between score triplets. Agreement between the average of the human coders (J1, J2, and J3) and the average of SENTIWORDNET versions 1.0 and 1.1, is reported as a function of POS.

tagging *synsets* (they only indirectly evaluate their system by tagging the General Inquirer, which is a set of manually tagged *terms*).

Previous work dealing with the properties of subsentential linguistic units from the standpoint of sentiment analysis has dealt with four main tasks:

1. Determining term PN-polarity, as in deciding if a given Subjective term (i.e. a term that bestows a positive or negative connotation on its denoted entity) has a Positive or a Negative slant;

2. Determining term SO-polarity, as in deciding whether a given term has a Subjective or an Objective (i.e. neutral, or factual) nature;

3. Determining the *strength* of term attitude (either PN-polarity or SO-polarity), as in attributing to terms (real-valued) degrees of positivity or negativity;

4. Tackling Tasks 1–3 for *multiword terms*; that is, predicating properties such as Subjective, Positive, or Mildly Positive, of complex expressions such as `not entirely satisfactory`.

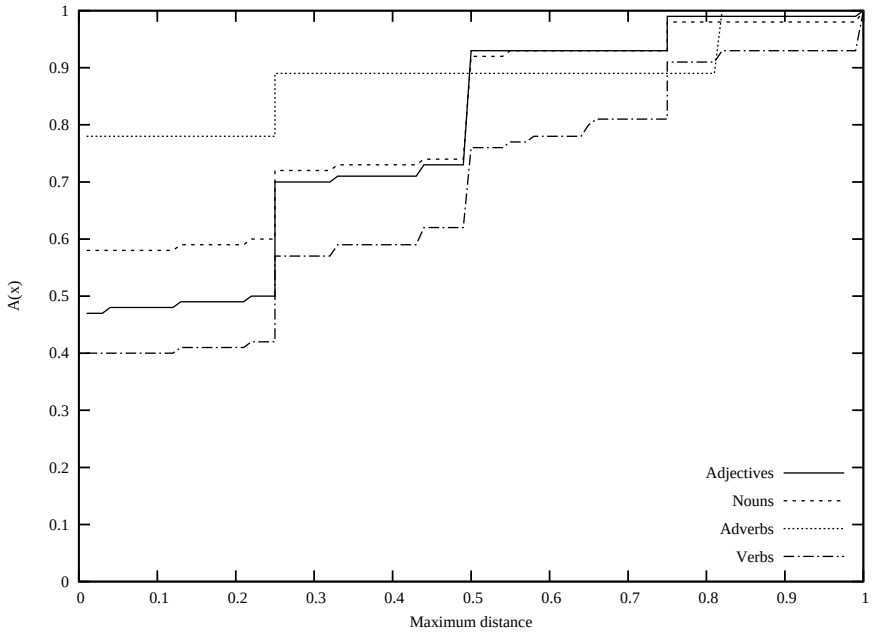

Figure 8. Agreement (on MICRO-WNOP(2)) as a function of maximum distance between score triplets. Interannotator agreement between J1 and J2 is reported as a function of POS

Concerning Task 1, the most influential work so far has probably been Turney and Littman's (2003), who determine the PN-polarity of subjective terms by bootstrapping from two (a Positive and a Negative) small sets of subjective "seed" terms. Their method is based on computing the *pointwise mutual information* (PMI) of the target term t with each seed term t_i as a measure of their semantic association. They determine the PN-polarity of a target term by checking whether its average PMI with the Positive seed terms is higher or not than its average PMI with the Negative seed terms. They query the AltaVista search engine[12] with a "t" query, a "t_i" query, and a "t NEAR t_i" query, and use the number of matching documents returned by AltaVista as estimates of the marginal and joint probabilities of occurrence needed for the computation of PMI. PMI is a real-valued function, and its scores can thus be used to provide a solution for Task 3. Other efforts at solving Task 1 are those of Andreevskaia and Bergler (2006a), Esuli and Sebastiani (2005), Hatzivassiloglou and McKeown (1997), Kamps et al. (2004), Kim and Hovy (2004), and 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 dif-

[12] `http://www.altavista.com/`

Table I. Percentages of WordNet synsets that have obtained a given score in SentiWordNet 1.0 for our three categories of interest, grouped by POS, and average scores obtained for all WordNet synsets with a given POS.

Score	Positive	Negative	Objective	Positive	Negative	Objective
	Adjectives			**Verbs**		
0.0	65.77%	62.81%	0.08%	89.98%	87.93%	0.00%
0.125	12.12%	7.32%	2.14%	4.43%	4.94%	0.21%
0.25	8.81%	8.68%	7.42%	2.66%	2.95%	0.64%
0.375	4.85%	5.19%	11.73%	1.55%	1.81%	1.35%
0.5	3.74%	5.63%	9.50%	0.84%	1.24%	2.67%
0.625	2.94%	5.53%	7.65%	0.84%	1.24%	2.67%
0.75	1.28%	3.72%	9.21%	0.10%	0.42%	4.57%
0.875	0.47%	1.07%	7.57%	0.07%	0.08%	6.11%
1.0	0.03%	0.04%	44.71%	0.00%	0.00%	81.05%
Avg	**0.106**	**0.151**	**0.743**	**0.026**	**0.034**	**0.940**
	Nouns			**Adverbs**		
0.0	90.80%	89.25%	0.00%	43.70%	76.99%	0.00%
0.125	4.53%	3.93%	0.23%	6.25%	9.66%	0.57%
0.25	2.37%	2.42%	0.87%	6.17%	5.32%	3.00%
0.375	1.25%	1.54%	1.84%	14.44%	2.51%	12.83%
0.5	0.62%	1.35%	2.32%	22.63%	2.70%	23.91%
0.625	0.24%	0.91%	2.57%	5.70%	1.72%	13.56%
0.75	0.14%	0.48%	3.27%	1.06%	0.82%	6.11%
0.875	0.05%	0.12%	5.40%	0.05%	0.27%	7.04%
1.0	0.00%	0.00%	83.50%	0.00%	0.00%	32.97%
Avg	**0.022**	**0.034**	**0.944**	**0.235**	**0.067**	**0.698**

ficult than Task 1; they do this by employing variants of the same method and corpus on which they had obtained state-of-the-art effectiveness at Task 1 (Esuli and Sebastiani, 2005), and by showing that much lower performance figures are obtained. Other works dealing with this task are those of Andreevskaia and Bergler (2006a), Baroni and Vegnaduzzo (2004), Riloff et al. (2003), and Wiebe (2000).

Concerning Task 4, the only work we are aware of 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`, which gets analysed as having Positive PN-polarity, Propriety attitude type, and Low force. Experimental results showed that using

Table II. Percentages of WordNet synsets that have obtained a given score in SentiWordNet 1.0 for our three categories of interest, and average scores obtained for all WordNet synsets (all parts of speech considered together).

Score	Positive	Negative	Objective
	All parts of speech		
0.0	85.18%	84.45%	0.02%
0.125	5.79%	4.77%	0.54%
0.25	3.56%	3.58%	1.97%
0.375	2.28%	2.19%	3.72%
0.5	1.85%	2.07%	4.20%
0.625	0.87%	1.64%	3.83%
0.75	0.35%	1.00%	4.47%
0.875	0.12%	0.27%	5.88%
1.0	0.01%	0.01%	75.37%
Avg	**0.043**	**0.054**	**0.903**

such "appraisal groups" as features for movie review classification gave a noticeable increase in sentiment classification accuracy.

5. Conclusion and future research

We have presented SentiWordNet, an automatically generated lexical resource in which each WordNet synset is tagged with a triplet of numerical scores representing how Positive, Negative, and Objective a synset is. We believe that SentiWordNet can prove a useful tool for opinion mining applications, because of its wide coverage (*all* WordNet synsets are tagged according to *each* of the three labels Objective, Positive, Negative) and because of its fine grain, obtained by qualifying the labels by means of numerical scores.

We are currently testing new algorithms for tagging WordNet synsets with opinion-related properties, and thus plan to continue the development of SentiWordNet beyond the currently released "Version 1.1", hopefully allowing us to make available to the scientific community more and more refined releases of SentiWordNet.

Acknowledgments

This work was partially supported by 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

Andreevskaia, A. and S. Bergler: 2006a, '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.

Andreevskaia, A. and S. Bergler: 2006b, 'Sentiment Tagging of Adjectives at the Meaning Level'. In: *Proceedings of the 19th Conference of the Canadian Society for Computational Studies of Intelligence (AI'06)*. Quebec City, CA, pp. 336—346.

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 2004)*. Vienna, AU, pp. 17–24.

Cerini, S., V. Compagnoni, A. Demontis, M. Formentelli, and G. Gandini: 2007, 'Micro-WNOp: A gold standard for the evaluation of automatically compiled lexical resources for opinion mining'. In: A. Sansò (ed.): *Language resources and linguistic theory: Typology, second language acquisition, English linguistics*. Milano, IT: Franco Angeli Editore, pp. 200–210.

Esuli, A. and F. 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 2005)*. Bremen, DE, pp. 617–624.

Esuli, A. and F. 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 2006)*. Trento, IT, pp. 193–200.

Esuli, A. and F. Sebastiani: 2006b, '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.

Harabagiu, S. M., G. A. Miller, and D. I. Moldovan: 1999, 'WordNet 2: A Morphologically and Semantically Enhanced Resource'. In: *Proceedings of the ACL Workshop on Standardizing Lexical Resources (SIGLEX 1999)*. College Park, US, pp. 1–8.

Hatzivassiloglou, V. and K. R. McKeown: 1997, '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.

Hatzivassiloglou, V. and J. M. Wiebe: 2000, 'Effects of adjective orientation and gradability on sentence subjectivity'. In: *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*. Saarbrücken, DE, pp. 174–181.

Kamps, J., M. Marx, R. J. Mokken, and M. De Rijke: 2004, 'Using WordNet to measure semantic orientation of adjectives'. In: *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Vol. IV. Lisbon, PT, pp. 1115–1118.

Kim, S.-M. and E. Hovy: 2004, 'Determining the Sentiment of Opinions'. In: *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*. Geneva, CH, pp. 1367–1373.

Kim, S.-M. and E. Hovy: 2006, 'Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text'. In: *Proceedings of ACL/COLING 2006 Workshop on Sentiment and Subjectivity in Text*. Sidney, AUS.

Nigam, K., A. K. McCallum, S. Thrun, and T. M. Mitchell: 2000, 'Text Classification from Labeled and Unlabeled Documents using EM'. *Machine Learning* **39**(2/3), 103–134.

Pang, B. and L. Lee: 2004, 'A Sentimental Education: Sentiment Analysis using Subjectivity Summarization based on Minimum Cuts'. In: *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL 2004)*. Barcelona, ES, pp. 271–278.

Pang, B. and L. Lee: 2005, 'Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales'. In: *Proceedings of the 43rd Meeting of the Association for Computational Linguistics (ACL 2005)*. Ann Arbor, US, pp. 115–124.

Riloff, E., J. Wiebe, and T. Wilson: 2003, 'Learning subjective nouns using extraction pattern bootstrapping'. In: *Proceedings of the 7th Conference on Natural Language Learning (CoNLL 2003)*. Edmonton, CA, pp. 25–32.

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: The MIT Press.

Takamura, H., T. Inui, and M. Okumura: 2005, '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.

Tumer, K. and J. Ghosh: 1996, 'Error Correlation and Error Reduction in Ensemble Classifiers'. *Connection Science* **8**(3/4), 385—403.

Turney, P.: 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 (ACL 2002)*. Philadelphia, US, pp. 417–424.

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

Valitutti, A., C. Strapparava, and O. Stock: 2004, 'Developing Affective Lexical Resources'. *PsychNology Journal* **2**(1), 61—83.

Vegnaduzzo, S.: 2004, 'Acquisition of Subjective Adjectives with Limited Resources'. In: *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*. Stanford, US.

Whitelaw, C., N. Garg, and S. Argamon: 2005, '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.

Wiebe, J.: 2000, '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.

Wilson, T., J. Wiebe, and R. 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 2004)*. San Jose, US, pp. 761–769.

Yu, H. and V. Hatzivassiloglou: 2003, 'Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences'. In: *Proceedings of the 8th Conference on Empirical Methods in Natural Language Processing (EMNLP 2003)*. Sapporo, JP, pp. 129–136.

Address for Offprints:
Andrea Esuli
Istituto di Scienza e Tecnologie dell'Informazione
Consiglio Nazionale delle Ricerche
Via Giuseppe Moruzzi 1 – 56124 Pisa, Italy