Enhancing Opinion Extraction by Automatically Annotated Lexical Resources

Andrea Esuli and Fabrizio Sebastiani

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

In this paper we tackle an *opinion extraction* (OE) task, i.e., identifying in a text each expression of subjectivity, the subject expressing it, and its possible target. We especially focus on how lexical resources specifically developed for opinion mining could be used to improve the performance of an opinion extraction system. We report results on two manually annotated corpora, one of English and one of Italian texts. We evaluate our results using standard evaluation measures and also using two new evaluation measures we have proposed.

1. Introduction

An emerging task in opinion mining is *opinion extraction* (OE), a specialization of *information extraction* (IE) which consists in detecting, within a sentence or a document, the expressions denoting the key components of an opinion (e.g., the opinion holder, the object of the opinion, the type of opinion, the strength of the opinion, etc.). OE is harder than other IE tasks, basically because the same opinion may be expressed in many subtly different forms.

In this paper we deal with OE as defined in (Wiebe et al., 2003; Wiebe et al., 2005), who focus on *annotating* texts, either manually or automatically, by the *expressions of private state* (EPSs) contained in them, i.e., by expressions denoting "an internal state that cannot be directly observed by others", and that as such includes "opinions, beliefs, thoughts, feelings, emotions, goals, evaluations, and judgments" (Wiebe et al., 2005, pp. 168).

The goal of our experiments is to comparatively evaluate the impact of using, in an OE task, lexical resources explicitly devised for OM. We use an IE system based on support vector machines (SVMs), and test the impact on extraction accuracy of several lexical resources. We show that the use of one such resource, SENTIWORDNET (Esuli and Sebastiani, 2006b), produces a noteworthy improvement in effectiveness with respect to the baseline system and, more importantly, with respect to the use of other lexical resources currently available. We run our experiments on two benchmarks: (i) the well-known MPQA corpus, and (ii) I-CAB Opinion (Esuli et al., 2008), a corpus of Italian news that we have manually annotated by EPSs using the same annotation language of the MPQA corpus. The experiments on I-CAB Opinion allow us to illustrate a "cross-language" use of SENTIWORDNET, since SENTIWORDNET is an English-language resource.

For the evaluation of our experiments we use standard evaluation measures for IE and OE, based on a model that considers each annotated textual expression as a single entity. We also use a new evaluation measure, based on viewing each *token* (i.e., any distinct alphanumeric expression, separated form the others by blanks and punctuation) and each *separator* (i.e., each string of symbols that separates two subsequent tokens) composing the text as a distinct entity to be labeled according to a given annotation tag. This new measure allows for a more rigorous evaluation of IE, capturing all the aspects that determine the effectiveness of an IE system.

1.1. Related work

(Wiebe et al., 2005) is currently the most relevant work on the annotation of opinions in text. They focus on the definition of an annotation language capable of capturing the various expressions of subjectivity in text. They propose (what we here call) the WWC opinion markup language, which is used to annotate in text the expressions of opinion along with the *opinion holder*, i.e., the subject that has the opinion, and the (possible) *opinion target*, i.e., the entity about which the opinion is expressed. They use this language to annotate a corpus of English news, the MPQA corpus (see Section 2.1.), which has then become the reference corpus for OE experimentation.

(Kim and Hovy, 2005) use MPQA to work on the task of recognizing the opinion holder in an opinion expression. Their work is focused on recognizing opinion holders for use in a question answering system. Given as input a question such as "`What does X think about Y?`", the recognition of the opinion holder allows one to eliminate from the candidate answers all the opinions about Y which are not by X.

Also (Choi et al., 2005) used MPQA to work on the identification of opinion holder. They model the task as an IE problem, in which each token composing a sentence has to be classified as belonging or not to an expression identifying an opinion holder. The vectorial representations of tokens are based on a rich set of syntactic features, plus subjectivity features extracted from various OM-specific lexical resources. In a subsequent work, (Choi et al., 2006) investigate the effects of *jointly* extracting of opinion holders and opinion expressions.

2. Annotating expressions of private state

WWC provides five types of *tags* (here indicated in SMALL CAPS) that identify the various components involved in EPSs.

In WWC every EPS is mapped into a *private state frame*, i.e., a structured object in which the real-world entities that play a role in the EPS are annotated by means

of the tags and further qualified by means of the attributes. In each private state a *source agent* holds a private state, possibly towards a *target agent*. WWC identifies three kinds of private states: (i) the explicit mention of a private state (e.g., "`I` **`fear`** `the Greeks, even when they bring presents`"); (ii) a speech event expressing a private state (e.g., "`You` **`said`** `you love her.`"); and (iii) an expressive subjective element (e.g., "`He is a` **`nice`** `person`").

A textual expression (*text span*, in WWC terminology) identifying the source agent or the target agent of a private state is annotated with the AGENT tag, which assigns a unique (at the document level) identifier to the entity denoted by the expression.

The explicit mention of a private state (Type (i) above), or a speech event expressing a private state (Type (ii) above) are annotated using the DIRECT-SUBJECTIVE tag.

Reported speech about objective facts is also annotated (e.g., "`John` **`said`** `he is 30`"), using the OBJECTIVE-SPEECH-EVENT tag.

Subjective expressions in text are annotated using the EXPRESSIVE-SUBJECTIVITY tag, which qualifies the annotated text by means of three attributes: source agents chain, intensity, and polarity of the expression.

2.1. MPQA and I-CAB Opinion

WWC has been used in (Wiebe et al., 2005) to manually annotate EPSs in the MPQA corpus. MPQA consists of 535 documents (10,657 sentences), which are English versions of news articles collected from 187 press sources around the world, and dating from Jun 2001 to May 2002.

Our experiments adopt the document split used by previous works on MPQA (see Section 1.1.): an *optimization* set, consisting of the first 135 documents, used for parameter optimization, and a test set, consisting of the remaining 400 documents, on which the final experiments, with optimized parameters, are run. On both sets experiments are done by using a 10-fold cross validation protocol.

I-CAB Opinion (Esuli et al., 2008) is the result of annotating the Italian Content Annotation Bank (Magnini et al., 2006) by EPSs using the WWC markup language. I-CAB is a corpus of newspaper articles in Italian, manually annotated with semantic information of various types, including TEMPORAL EXPRESSIONS, NAMED ENTITIES, and RELATIONS between such entities. I-CAB consists of 525 articles from an Italian newspaper, subdivided into a training set of 335 articles and a test set of 190 articles.

3. Evaluation models and measures

We evaluate the results of our experiments using two different evaluation models.

The first is a widely used model, that we call *annotation-based model*, and that considers each annotated text span as a single entity. The evaluation is based on comparing the matches among the sets of true annotations, from the benchmark corpus, with the set of predicted annotations, from the OE system.

The other model, that we call *token & separator model* (Esuli et al., 2009), is based on considering each *token* and each *separator* composing the text as a single entity belonging or not to an annotation (this is repeated for each possible annotation type).

3.1. Annotation-based model

The annotation-based model considers each *annotated text span* (hereafter: "annotation") as a single entity. The evaluation thus consists in identifying, for each tag X in the annotation language, the matches between the set of gold-standard annotations $G_X = \{g_1, \ldots, g_n\}$ and the set of the *predicted* annotations $P_X = \{p_1, \ldots, p_m\}$.

A first point to note is that P may contain an arbitrary number of elements, not necessarily equal to the number of elements in G. Moreover, the annotations in P may obviously span any portion of text in the annotated documents, without any relation with G. A consequence of this facts is that it is impossible to establish any one-to-one relation between the elements of G and P. The typical approach (Lavelli et al., 2008) is to define a predicate $match(g, p)$, with values in $\{True, False\}$, which determines if there is a match between two annotations $g \in G$ and $p \in P$, and then use this predicate to compute an approximate version of precision (π) and recall (ρ):

$$\pi(G, P) = \frac{|\{p | p \in P \wedge \exists g \in G : match(g, p)\}|}{|P|} \quad (1)$$

$$\rho(G, P) = \frac{|\{g | g \in G \wedge \exists p \in P : match(g, p)\}|}{|G|} \quad (2)$$

after which F_1 (their harmonic mean) can be computed. Three widely adopted definitions for the $match$ predicate are (i) **overlap**, defined as $match_{overlap}(g, p) = True$ iff the two annotations have *any overlap* in text; (ii) **head**, defined as $match_{head}(g, p) = True$ iff the two annotations *start* from the same position in text; and (iii) **exact**, defined as $match_{exact}(g, p) = True$ iff the two annotations *start and end* at the same positions in text.

Unfortunately, these predicates have drawbacks. The $match_{overlap}$ predicate overestimates the performance of a system that produces long annotations: a trivial system that annotates the entire document with a single annotation obtains a perfect score. On the opposite side, the $match_{head}$ and $match_{exact}$ predicates are too strict in their evaluation, because they treat many or all (respectively) approximate matches as full errors. The problem with all these predicates is that they do not take into account the *degree* of overlap between the two annotations. For instance, two annotations that are each 10 words long and overlap by 1 word only, receive the same partial credit as two annotations that are each 10 words long and overlap by 9 words, which is unintuitive.

A further problem with the annotation-based model is that it is not possible to compute a full contingency table, due to the fact that there is no notion of a "true negative" annotation. This prevents using evaluation measures that require a full contingency table, such as Cohen's κ (which is typically used to measure inter-annotator agreement).

3.2. Token & separator model

The evaluation model we have proposed in (Esuli et al., 2009) is based on considering each *token* and each *separator* in the text as entities belonging or not to a tag. In practice, we reformulate the problem of evaluating annotations as a problem of evaluating the classification of tokens and separators, where the various tags represent the categories that tokens and separators can be assigned to.

In order to give formal definition we first analyze a simplified version of this "token & separator model": the *token model*. In the token model, given a document d consisting of a sequence $\{t_1, \ldots, t_k\}$ of tokens, we say token t_i belongs to tag X iff there exists at least one annotation $g \in G_X$ which includes t_i. The predictions by the system are interpreted similarly: we say token t_i is predicted to belong to tag X iff there is an annotation $p \in P_X$ which includes t_i. This generates two token classifications which can be compared using any standard evaluation measure (e.g., precision, recall, F_1, Cohen's κ, etc.).

A potential problem with the token model is that it is not able to recognize if two adjacent tokens classified with the same tag X belong to the same annotation or to two distinct adjacent annotations. The *token & separator* model extends the token model in order to solve this problem. In the token & separator model the document d is considered as an alternating sequence of tokens and separators $\{t_1, b_1, t_2, \ldots, t_{k-1}, b_{k-1}, t_k\}$, where by "separator" we mean any sequence of characters that separates one token from the other (e.g., a comma followed by a blank). When tokens $t_i, \ldots, t_j$ are part of the *same* annotation for tag X, also separators $b_i, \ldots, b_{(j-1)}$ are considered to be part of the same annotation for X. When two consecutive tokens t_i and $t_{(i+1)}$ belong instead to two *separate* annotations for tag X, then separator b_i is considered not to belong to X. The evaluation is performed as in the token model, the only difference being that both tokens and separators concur to the final result.

With respect to the annotation-based model discussed above, the token & separator model has many advantages: (i) the number of entities under evaluation is constant; (ii) it is possible to compute a full contingency table; (iii) values in the contingency table are robust with respect to the role of the two classifications being compared, i.e., switching gold standard and predictions just swaps false positives with false negatives, leaving unchanged the numbers of true positives and true negatives; (iv) token-based evaluations are tolerant to minor errors, such as adding a spurious token to a long annotation; (v) at the same time, token-based evalutations are strict on assigning high scores, i.e., the perfect score is returned by the token & separator model only when the gold standard and the prediction are exactly the same.

4. The opinion extraction system

As the learning and classification engine of our OE system we have used YamCha[1], a general-purpose system for performing text chunking tasks based on SVMs.

YamCha takes as input an IOB2-formatted file in which each token t_i is represented by a list of *features* $F_i = \{f_i^1, \ldots, f_i^n\}$ (e.g., the token t_i, its POS and lemmatized version, etc.) and a target classification label c_i. YamCha allows to enrich the representation of a token by adding information from the tokens contained in a specified neighborhood window. For example, when specifying a $[-2, +2]$ *static* window, the representation for token t_i also consists of the features of the two preceding and two following tokens, i.e., $F_i = \{f_{i-2}^1, \ldots, f_{i-2}^n, \ldots, f_i^1, \ldots, f_i^n, \ldots, f_{i+2}^1, \ldots, f_{i+2}^n\}$, thus allowing the learner to capture information from the context surrounding the observed token. Similarly, a *dynamic* window can be specified to enrich the representation with information about the tags assigned to the preceding tokens.

In our experiments we have considered the annotation of each tag type as a distinct task, thus running separate experiments for each tag type.

[1] http://www.chasen.org/~taku/software/

5. Lexical resources for OM

We test the impact on the OE task of four lexical resources. The first is the General Inquirer, a list of 1,614 positive terms and 1,982 negative terms extracted from the lexicon of the General Inquirer text analysis system (Stone et al., 1966). The second is HM, a lexicon of 657 positive/679 negative adjectives developed for a work on the identification of the polarity of terms (Hatzivassiloglou and McKeown, 1997). The third is SENTIWORDNET 1.0 (Esuli and Sebastiani, 2006b), a lexical resource in which each WORDNET synset s is associated to three numerical scores $Obj(s)$, $Pos(s)$ and $Neg(s)$, describing how objective, positive, and negative the terms contained in the synset are. 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 last resource is SENTIWORDNET 2.0 (Esuli, 2008) an improved version of SENTIWORDNET 1.0 obtained by applying *random-walk* methods to the graph defined by the relation *definiens-definiendum* between the synsets of WORDNET (Esuli and Sebastiani, 2007a).

6. Experiments

We have carried out our experiments with the goal of measuring the impact of the above described OM-specific lexical resources on OE. We have thus prepared various versions of the two annotated corpora we have used, each one with specific information extracted from documents by using the various OM-specific lexical resources.

More in detail, for the MPQA corpus we have tested five *feature sets*. The first is **BASE**, where each token is represented by the following features: (i) the term identifying the token, exactly as it appears in the text; (ii) the lowercased version of the term; (iii) a feature that specifies the capitalization properties of the term, ranging on {AllLowerCase, AllUpperCase, Mixed, NotWord}; (iv) the POS of the term, obtained via the Brill tagger. The second is **GI**, which consists of BASE plus a feature which indicates if the term is labeled as either Positive or Negative in the General Inquirer's lexicon (Stone et al., 1966); this resulted in tagging 1,416 distinct terms in the MPQA corpus as subjective, for a total of 98,130 occurrences. The third is **HM**, which consists of BASE plus a feature which indicates if the term appears in the HM subjectivity lexicon discussed above; this resulted in tagging 747 distinct terms in MPQA as subjective, for a total of 31,620 occurrences. The fourth is **SWN1**, which consists of BASE plus a feature that indicates if the term is one of the 2,645 distinct terms in the MPQA corpus that has a subjectivity score higher than 0.5 in SENTIWORDNET 1.0, for a total

of 171,467 occurrences. (We define the SENTIWORDNET subjectivity score for a term as the sum of positivity and negativity scores of all the synsets the term belongs to.) The fifth and last is **SWN2**, similar to SWN1 but based on SENTIWORDNET 2.0, which identifies 2,333 subjective terms in MPQA, for a total of 176,600 occurrences. We denote by **ALLSUBJ** the union of all the features defined in the previous feature sets.

For the I-CAB Opinion corpus, the problem is that we do not have any OM-specific lexical resource for the Italian language. We have then used MultiWordNet (Pianta et al., 2002) in order to map the SENTIWORDNET scores to Italian synsets. On I-CAB Opinion we have tested three feature sets. The first is **BASE** (defined analogously as for MPQA). The second is **SWN1**, which consists of BASE plus a subjectivity feature based on the Italian mapping of SENTIWORDNET 1.0, computed in the same way as for the English version; this process determined a set of 541 subjective terms in I-CAB Opinion, for a total of 19,051 occurrences. The third is **SWN2**, which is the same as SWN1 but based on SENTIWORDNET 2.0, resulting in 523 subjective terms and 17,610 occurrences.

We have used the windowing option of YamCha specifying a $[+2, -2]$ static window and a $[-2, -1]$ dynamic window, optimizing these values with a 10-fold cross validation experiment on the validation part of MPQA.

7. Results and conclusions

The results of the OE experiments are summarized in Table 1. A general trend clearly emerges: the use of OM-specific lexical resources improves effectiveness, producing a high gain in recall, which largely compensates a small loss in precision (i.e., lexical resources allow to spot more text spans with relevant information, at the same time bringing about a minor number of additional false positives). The use of lexical resources has the best impact on the EXPRESSIVE-SUBJECTIVITY tag. This is reasonable, given the affinity between the semantics of the tag and the lexical resources.

On MPQA the average improvement, in terms of token & separator-based F_1, over the various tags with respect to the BASE feature set, is 2.23% for the GI features set, 1.35% for HM, 4.40% for SWN1, 4.30% for SWN2, and 5.79% for ALLSUBJ. The SWN1 and SWN2 feature sets always perform better than the GI and HM feature sets. Between the two versions of SENTIWORDNET -based features there is no clear winner.

The better performance of the SENTIWORDNET-based feature sets indicates that their wide coverage of the language largely compensates for their inaccuracies, due to their automatic generation. For example, in SWN1 the term `phone` is erroneously marked as subjective, but SWN1 also includes, correctly, the terms `advantageous` and `insulting`, which are missing from both GI and HM. However, the ALLSUBJ feature set always scores the best result, suggesting that none of the tested lexical resources "contains" the others, and that each contains *relevant* information about subjective language that the others do not capture.

Our results on MPQA do not reach the state-of-the-art results reported in literature (e.g., (Choi et al., 2006; Choi et al., 2005)). We point out that we have designed our experiments with the aim of creating an "isolated" environment for the evaluation of the impact of OM-specific lexical resources on OE. We have reduced the BASE features to a minimal definition and we have not used any advanced NLP tool.

I-CAB Opinion results are generally of lower quality compared to those obtained on MPQA. A possible reason for this may be found in the higher relative hardness of I-CAB Opinion with respect to MPQA, which can be hypothesized by observing the inter-annotator agreement values obtained on the two corpora (Esuli et al., 2008).

On I-CAB Opinion the SENTIWORDNET-based feature sets improve with respect to the BASE feature set. The average improvement over the various tags is 3.56% for SWN1 and 3.39% for SWN2; values are lower than those measured on MPQA, probably due to the limited coverage of MultiWordNet on the Italian language.

8. References

Choi, Y., E. Breck, and C. Cardie, 2006. Joint extraction of entities and relations for opinion recognition. In *Proceedings of EMNLP'06*. Sydney, AU.

Choi, Y., C. Cardie, E. Riloff, and S. Patwardhan, 2005. Identifying sources of opinions with conditional random fields and extraction patterns. In *Proceedings of HLT/EMNLP'05*. Vancouver, CA.

Esuli, A., 2008. Automatic Generation of Lexical Resources for Opinion Mining: Models, Algorithms and Applications. PhD in Information Engineering, PhD School "Leonardo da Vinci", University of Pisa.

Esuli, A., M. Pryczek, and F. Sebastiani, 2009. Evaluating information extraction systems. Technical report, Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa, IT. Forthcoming.

Esuli, A. and F. Sebastiani, 2006b. SENTIWORDNET: A publicly available lexical resource for opinion mining. In *Proceedings of LREC'06*. Genova, IT.

Esuli, A. and F. Sebastiani, 2007a. Random-walk models of term semantics: An application to opinion-related properties. In *Proceedings of LTC'07*. Poznań, PL.

Esuli, A., F. Sebastiani, and I. Urciuoli, 2008. Annotating expressions of opinion and emotion in the Italian Content Annotation Bank. In *Proceedings of LREC'08*. Marrakech, MA.

Hatzivassiloglou, V. and K. R. McKeown, 1997. Predicting the semantic orientation of adjectives. In *Proceedings of ACL'97*. Madrid, ES.

Kim, S.-M. and E. Hovy, 2005. Identifying opinion holders for question answering in opinion texts. In *Proceedings of the AAAI'05 Workshop on Question Answering in Restricted Domains*. Pittsburgh, US.

Lavelli, A., M. E. Califf, F. Ciravegna, D. Freitag, C. Giuliano, N.s KushmE.k, L. Romano, and N. Ireson, 2008. Evaluation of machine learning-based information extraction algorithms: Criticisms and recommendations. *Language Resources and Evaluation*, 42(4):361–393.

Magnini, B., E. Pianta, C. Girardi, M. Negri, L. Romano, M. Speranza, V. Bartalesi-Lenzi, and R. Sprugnoli, 2006. I-CAB: The Italian content annotation bank. In *Proceedings of LREC'06*. Genova, IT.

Pianta, E., L. Bentivogli, and C. Girardi, 2002. MultiWordNet: Developing an aligned multilingual database. In *Proceedings of GWN'02*. Mysore, IN.

Model	Annotation match									Token & Separator		
Predicate	Overlap			Head			Exact					
	π	ρ	F_1	π	ρ	F_1	π	ρ	F_1	π	ρ	F_1
MPQA												
	AGENT											
BASE	.725	.526	.609	.652	.471	.547	.598	.432	.502	.634	.449	.526
GI	.715 (-1.33%)	.534 (1.61%)	.611 (0.35%)	.638 (-2.08%)	.476 (0.96%)	.545 (-0.34%)	.586 (-2.13%)	.436 (0.90%)	.500 (-0.39%)	.622 (-1.91%)	.457 (1.75%)	.527 (0.20%)
HM	.712 (-1.74%)	.538 (2.34%)	.613 (0.58%)	.638 (-2.15%)	.473 (0.41%)	.543 (-0.68%)	.586 (-2.13%)	.436 (0.90%)	.500 (-0.39%)	.622 (-1.95%)	.456 (1.59%)	.526 (0.09%)
SWN1	.711 (-1.82%)	.548 (4.21%)	.619 (1.59%)	.630 (-3.31%)	.484 (2.75%)	.548 (0.12%)	.578 (-3.45%)	.443 (2.58%)	.502 (-0.04%)	.624 (-1.65%)	.475 (5.90%)	.540 (2.63%)
SWN2	.713 (-1.63%)	.548 (4.29%)	.620 (1.71%)	.632 (-3.02%)	.484 (2.80%)	.548 (0.27%)	.578 (-3.35%)	.443 (2.42%)	.502 (-0.08%)	.621 (-2.05%)	.474 (5.52%)	.538 (2.24%)
ALLSUBJ	.701 (-3.21%)	.555 (5.52%)	.619 (1.66%)	.630 (-3.37%)	.487 (3.48%)	.550 (0.49%)	.565 (-5.67%)	.445 (3.03%)	.498 (-0.81%)	.623 (-1.79%)	.479 (6.72%)	.542 (3.02%)
	DIRECT-SUBJECTIVE											
BASE	.668	.424	.519	.555	.349	.428	.485	.305	.375	.613	.321	.422
GI	.664 (-0.59%)	.447 (5.45%)	.534 (3.02%)	.547 (-1.37%)	.365 (4.64%)	.438 (2.23%)	.476 (-2.04%)	.317 (3.95%)	.381 (1.55%)	.608 (-0.77%)	.341 (6.15%)	.437 (3.66%)
HM	.664 (-0.59%)	.447 (5.45%)	.534 (3.02%)	.540 (-2.65%)	.365 (4.64%)	.436 (1.70%)	.490 (0.93%)	.310 (1.51%)	.380 (1.29%)	.583 (-4.92%)	.330 (2.70%)	.421 (-0.06%)
SWN1	.660 (-1.19%)	.465 (9.76%)	.546 (5.23%)	.541 (-2.54%)	.377 (8.17%)	.444 (3.77%)	.472 (-2.88%)	.329 (7.81%)	.388 (3.42%)	.600 (-2.07%)	.358 (11.31%)	.448 (6.32%)
SWN2	.660 (-1.30%)	.464 (9.39%)	.545 (4.98%)	.539 (-2.79%)	.376 (7.71%)	.443 (3.40%)	.469 (-3.48%)	.327 (6.95%)	.385 (2.67%)	.599 (-2.36%)	.355 (10.57%)	.446 (5.75%)
ALLSUBJ	.654 (-2.20%)	.489 (15.31%)	.559 (7.82%)	.527 (-4.94%)	.391 (11.97%)	.449 (4.78%)	.460 (-5.25%)	.338 (10.60%)	.390 (3.89%)	.586 (-4.38%)	.378 (17.58%)	.460 (8.97%)
	EXPRESSIVE-SUBJECTIVITY											
BASE	.668	.368	.474	.445	.230	.304	.234	.121	.159	.503	.293	.370
GI	.656 (-1.83%)	.384 (4.46%)	.484 (2.14%)	.422 (-5.23%)	.242 (4.82%)	.307 (1.17%)	.229 (-2.14%)	.129 (6.77%)	.165 (3.56%)	.499 (-0.72%)	.315 (7.46%)	.386 (4.29%)
HM	.658 (-1.45%)	.374 (1.74%)	.477 (0.58%)	.430 (-3.45%)	.238 (3.29%)	.306 (0.89%)	.229 (-2.14%)	.124 (2.63%)	.161 (0.96%)	.499 (-0.72%)	.315 (7.46%)	.386 (4.29%)
SWN1	.651 (-2.56%)	.414 (12.55%)	.506 (6.68%)	.433 (-2.76%)	.260 (12.65%)	.325 (6.88%)	.224 (-4.29%)	.134 (10.93%)	.168 (5.23%)	.500 (-0.58%)	.326 (11.38%)	.395 (6.65%)
SWN2	.652 (-2.34%)	.414 (12.61%)	.506 (6.81%)	.431 (-3.33%)	.258 (12.06%)	.323 (6.29%)	.225 (-3.64%)	.135 (11.85%)	.169 (6.04%)	.503 (0.02%)	.327 (11.52%)	.396 (6.99%)
ALLSUBJ	.637 (-4.66%)	.433 (17.65%)	.515 (8.63%)	.430 (-3.45%)	.263 (13.93%)	.326 (7.34%)	.223 (-4.61%)	.139 (15.05%)	.171 (7.50%)	.497 (-1.12%)	.335 (14.28%)	.400 (8.08%)
	OBJECTIVE-SPEECH-EVENT											
BASE	.556	.432	.486	.528	.410	.461	.503	.391	.440	.546	.372	.443
GI	.552 (-0.76%)	.438 (1.32%)	.488 (0.40%)	.520 (-1.45%)	.418 (1.96%)	.463 (0.44%)	.497 (-1.16%)	.391 (0.06%)	.438 (-0.48%)	.540 (-1.13%)	.380 (2.11%)	.446 (0.77%)
HM	.554 (-0.40%)	.435 (0.62%)	.487 (0.17%)	.525 (-0.50%)	.412 (0.49%)	.462 (0.05%)	.500 (-0.56%)	.395 (1.08%)	.441 (0.35%)	.540 (-1.13%)	.382 (2.65%)	.447 (1.08%)
SWN1	.550 (-1.06%)	.448 (3.63%)	.494 (1.53%)	.517 (-1.93%)	.421 (2.57%)	.464 (0.55%)	.491 (-2.32%)	.399 (2.18%)	.440 (0.16%)	.536 (-1.88%)	.390 (4.84%)	.452 (2.01%)
SWN2	.551 (-0.98%)	.448 (3.67%)	.494 (1.58%)	.519 (-1.58%)	.422 (3.04%)	.466 (0.97%)	.493 (-1.96%)	.401 (2.66%)	.442 (0.59%)	.537 (-1.67%)	.391 (5.03%)	.452 (2.21%)
ALLSUBJ	.546 (-1.89%)	.458 (5.98%)	.498 (2.39%)	.513 (-2.84%)	.430 (4.90%)	.468 (1.37%)	.485 (-3.48%)	.407 (4.22%)	.443 (0.71%)	.530 (-2.90%)	.400 (7.59%)	.456 (3.08%)
I-CAB Opinion												
	AGENT											
BASE	.476	.235	.314	.442	.216	.291	.377	.184	.248	.397	.203	.269
SWN1	.470 (-1.37%)	.240 (2.09%)	.317 (0.92%)	.441 (-0.41%)	.222 (2.83%)	.296 (1.74%)	.370 (-1.87%)	.187 (1.33%)	.248 (0.26%)	.401 (0.88%)	.205 (0.92%)	.271 (0.90%)
SWN2	.463 (-2.72%)	.248 (5.48%)	.323 (2.63%)	.447 (1.04%)	.228 (5.48%)	.302 (3.98%)	.379 (0.60%)	.177 (-3.86%)	.248 (0.26%)	.400 (0.75%)	.205 (0.92%)	.271 (0.86%)
	DIRECT-SUBJECTIVE											
BASE	.466	.171	.250	.424	.155	.227	.424	.155	.227	.415	.124	.191
SWN1	.456 (-2.17%)	.177 (3.64%)	.255 (2.01%)	.416 (-1.82%)	.161 (4.00%)	.233 (2.37%)	.416 (-1.82%)	.161 (4.00%)	.233 (2.37%)	.403 (-2.71%)	.130 (4.41%)	.196 (2.68%)
SWN2	.447 (-4.13%)	.185 (8.38%)	.262 (4.72%)	.409 (-3.48%)	.158 (1.65%)	.228 (0.22%)	.412 (-2.87%)	.165 (6.46%)	.236 (3.78%)	.409 (-1.36%)	.130 (4.41%)	.197 (3.02%)
	EXPRESSIVE-SUBJECTIVITY											
BASE	.495	.222	.306	.411	.172	.243	.333	.139	.196	.407	.152	.221
SWN1	.499 (0.91%)	.245 (10.33%)	.328 (7.23%)	.420 (2.05%)	.194 (12.77%)	.266 (9.37%)	.362 (8.95%)	.168 (20.39%)	.229 (16.77%)	.409 (0.52%)	.170 (11.99%)	.240 (8.63%)
SWN2	.508 (2.69%)	.237 (6.92%)	.323 (5.57%)	.422 (2.61%)	.202 (17.39%)	.274 (12.60%)	.366 (9.94%)	.161 (15.58%)	.224 (13.86%)	.402 (-1.12%)	.169 (11.38%)	.238 (7.68%)
	OBJECTIVE-SPEECH-EVENT											
BASE	.592	.377	.460	.586	.372	.455	.579	.368	.450	.612	.383	.471
SWN1	.600 (1.33%)	.389 (3.33%)	.472 (2.55%)	.594 (1.37%)	.385 (3.37%)	.467 (2.58%)	.587 (1.41%)	.381 (3.41%)	.462 (2.62%)	.616 (0.66%)	.394 (2.88%)	.481 (2.01%)
SWN2	.600 (1.31%)	.392 (4.14%)	.474 (3.02%)	.596 (1.76%)	.378 (1.40%)	.462 (1.54%)	.595 (2.76%)	.389 (5.54%)	.470 (4.44%)	.616 (0.52%)	.394 (2.94%)	.481 (2.00%)

Table 1: Results of the automatic annotation of EPSs on the MPQA (upper part) and the I-CAB Opinion (lower) corpora.

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.

Wiebe, J., et al., 2003. Recognizing and organizing opinions expressed in the world press. In *Proceedings of the 2003 AAAI Spring Symposium on New Directions in Question Answering*. Stanford, US.

Wiebe, J., T. Wilson, and C. Cardie, 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2/3):165–210.