Enhancing Opinion Extraction by Automatically Annotated Lexical Resources

(Extended Version)

Andrea Esuli and Fabrizio Sebastiani

Istituto di Scienza e Tecnologia dell'Informazione
Consiglio Nazionale delle Ricerche
Via Giuseppe 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, complete with statistical significance tests and inter-annotator agreement data, on two manually annotated corpora, one of English and one of Italian texts. We evaluate our results using standard evaluation measures and also using a new evaluation measure we have recently 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.) [1,2,3,4,5]. 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 [6,7], who focus on *annotating* texts, either manually or automatically, by the *expressions of private state* (EPSs) contained in them[1], 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" [7, 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 [8], produces a noteworthy improvement in effectiveness with respect to the baseline system and, more importantly, with respect to

[1] In our OE task, with respect to EPSs we will use the verbs *to annotate* and *to extract* as synonyms, with the intended meaning of recognizing the relevant EPSs in a document.

Z. Vetulani (Ed.): LTC 2009, LNAI 6562, pp. 500–511, 2011.

the use of other lexical resources currently available. We run our experiments on two benchmarks: (i) the well-known MPQA corpus [7], and (ii) I-CAB Opinion [9], 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 that we have proposed in [10], 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, such as a comma followed by a blank) composing the text as a distinct entity to be labeled according to a given annotation tag. This new measure allows (as argued in [10]) for a more fine-grained evaluation of IE.

2 Related Work

The most relevant work to date on the annotation of opinions in text is probably [7]. The authors 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 the expressions of opinion along with (among others) 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 manually annotate a corpus of English news, the MPQA corpus (see Section 4.2), which has then become the reference corpus for OE experimentation.

[4] uses 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 [3] uses MPQA to work on the identification of the holder of an opinion. The authors 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 the tokens are based on a rich set of syntactic features, plus subjectivity features extracted from various lexical resources specifically developed for opinion mining (OM), and on extraction patterns automatically generated using the AutoSlog system [11]. The authors test their system on the MPQA corpus, measuring the effectiveness in recognizing the AGENT tag, which identifies the opinion holder in text. In a subsequent work [2], the same authors investigate the effects of *jointly* extracting opinion holders and opinion expressions, and test their system on the MPQA corpus. They use a global inference approach in which entities involved in opinion expressions (i.e., the opinion holder and the opinion itself) are extracted separately by means of a method similar to the one of [3], but designed to have a higher recall. Then

Token	Term	POS	Lemma	Label
$t-3$	It	PRP	it	O
$t-2$	is	VBZ	be	O
$t-1$	a	DT	a	O
t	love	NN	love	B-ES
$t+1$	hate	NN	hate	I-ES
$t+2$	relationship	NN	relationship	O
$t+3$	.	PUNC	.	O

Fig. 1. Example of YamCha input. Features included in the static and dynamic window of token t are underlined.

a global inference procedure, implemented using integer linear programming, is applied in order to produce the best pairing of opinion holders and opinion expressions, by exploiting their mutual dependencies and relations.

3 The Opinion Extraction System

3.1 The Learning and Classification System

As the learning and classification engine of our OE system we have used YamCha [12][2], a general-purpose system for performing information extraction tasks based on support vector machines (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 itself, its part of speech, its corresponding lemma, etc.) and a target classification label c_i (which is empty for all tokens belonging to test documents).

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., $F_i = \{f_{i-2}^1, \cdots, f_{i-2}^n, \cdots, f_i^1, \cdots, f_i^n, \cdots, f_{i+2}^1, \cdots, f_{i+2}^n\}$, thus allowing the learner to capture information from the context surrounding the observed token. Similarly, a *dynamic* window can be specified in order to enrich the representation with information about the tags assigned to the preceding tokens. Figure 1 shows an example of token representations, highlighting the features used to represent token t when a $[-2, +2]$ static window or a $[-2, -1]$ dynamic window are specified.

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

3.2 Lexical Resources for OM

We test the impact on the OE task of four lexical resources. The first is the General Inquirer, a manually developed list of 1,614 positive terms and 1,982

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

negative terms extracted from the lexicon of the General Inquirer text analysis system [13][3]. The second is HM, a manually developed lexicon of 657 positive/679 negative adjectives originally built for a work on the identification of the polarity of terms [14]. The third is SentiWordNet 1.0 [8], an automatically developed 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 fourth and last resource is SentiWordNet 2.0 an improved version of SentiWordNet 1.0 obtained by applying *random-walk* methods to the graph defined by the relation *definiens-definiendum* between WordNet synsets [15].

4 Experiments

4.1 The WWC Opinion Markup Language

We use WWC as a markup language for our experiments. 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 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 bearing gifts`"); (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`").

WWC provides five different *tags* (here indicated in small caps) that identify the various components involved in EPSs. 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. Since EPSs can be nested, it is natural to identify the outermost source of every EPS in a given text as the author of the text itself; by convention, the identifier denoting the author of the text is "writer".

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. WWC also includes an inside tag, used for identifying the scope of a speech event (e.g., "`Mary said I love pizza`"). This tag has not been used in MPQA (except for automatically marking an entire sentence as an inside for "writer").

[3] `http://www.wjh.harvard.edu/~inquirer/`

4.2 The Datasets: MPQA and I-CAB Opinion

WWC has been used in [7] 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 on MPQA adopt the document split used by previous works (see Section 2): a validation set, consisting of the first 135 documents, used as held-out data for parameter optimization, and a test set, consisting of the remaining 400 documents (8,297 sentences), on which the final experiments are run via 10-fold cross validation.

I-CAB Opinion [9] is the result of annotating by EPSs the Italian Content Annotation Bank (I-CAB) [16] 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.

4.3 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 the *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. For what counts as a match we adopt three widely used definitions [1,17,18], i.e., (i) **overlap**, defined as $match_{overlap}(g,p) = True$ iff the two annotations have any overlap; (ii) **head**, defined as $match_{head}(g,p) = True$ iff the two annotations start from the same token; and (iii) **exact**, defined as $match_{exact}(g,p) = True$ iff the two annotations start and end at the same tokens.

The other model, that we call the *token & separator model* [10], is based on considering each *token* and each inter-token *separator* composing the text as entities that may belong or not to a tag.

The evaluation is performed by applying to each adopted model the standard IR evaluation measures of precision, recall and F_1.

4.4 The 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 token itself as it appears in the text; (ii) the lowercased version of the token;

(iii) a feature that specifies the capitalization properties of the token, ranging on {AllLowerCase, AllUpperCase, Mixed, NotWord[4]}; (iv) the POS of the token, 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 [13]; 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 heuristically 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 [19] 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.

5 Results and Conclusions

5.1 The Results

The results of the OE experiments are summarized in Tables 1 and 2.

A globlal analysis of results indicates that the use of OM-specific lexical resources improves effectiveness, producing a high gain in recall, which largely compensates for 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.

[4] Numbers and alpha-numeric strings.

Table 1. Results of the automatic annotation of EPSs on the MPQA corpus

Model	Annotation									Token & Separator		
Predicate	Overlap			Head			Exact					
	π	ρ	F_1	π	ρ	F_1	π	ρ	F_1	π	ρ	F_1
	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%)

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 the inaccuracies due to the fact that they were automatically, and not manually, generated.

Table 2. Results of the automatic annotation of EPSs on the I-CAB Opinion corpus

Model	Annotation									Token & Separator		
Predicate	Overlap			Head			Exact					
	π	ρ	F_1	π	ρ	F_1	π	ρ	F_1	π	ρ	F_1
	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%)

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 the literature (e.g., [2,3]). This is due to the fact that we have designed our experiments with the only aim of creating an "isolated" environment for the evaluation of the impact of OM-specific lexical resources on OE. We have thus 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 (see Section 5.3, in which IAA is much higher on MPQA than on I-CAB Opinion).

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; these values are lower than those measured on MPQA, probably due to the limited coverage of MultiWordNet (the synsets of this latter are strictly a subset of those of WordNet. As for the MPQA experiments, the increase in recall is higher than the loss in precision (if any) and determines the increase in overall effectiveness. The highest increase in performance is again observed on the EXPRESSIVE-SUBJECTIVITY tag).

5.2 Statistical Significance Tests

In order to check whether the obtained results are statistically significant we have subjected them to thorough statistical significance testing. A statistical significance test takes in input two classifications, generally produced by two independent classifiers, and outputs a probability value P that indicates the probability that the observed differences in the two classifications are due to chance. A low P value is thus an indication that the observed differences are significant (i.e., *not* due to chance) and that the classifiers that have produced them are substantially different. Two common threshold values for P used in literature are $P \leq 0.05$ and $P \leq 0.01$, identifying the recognition of a statistically significant difference in the compared experiments, with increasing confidence. When $P > 0.05$ the difference is considered to be *not* statistically significant.

We have applied to the results of our experiments the *s-test* and the *p-test*, two significance tests designed for text classification systems (see [20, Section 4]). The *s-test* is a sign test [21, Chapter 17] which compares two classifiers $\hat{\Phi}_1$ and $\hat{\Phi}_2$ by analyzing their binary decisions on each document/category pair. The *p-test*, on π and ρ, is a t-test which compares two classifiers $\hat{\Phi}_1$ and $\hat{\Phi}_2$ by analyzing the microaveraged precision and recall values that the two systems have obtained. Both s-test and p-test are designed to compare two systems at the ("micro") level of individual classification decisions. Given the focus of our task on token classification, they are thus the ideal tests to evaluate the statistical significance of our experiments.

We have first compared all the experiments using the BASE feature set with all the others, thus performing 21 statistical significance tests[5], obtaining a "highly statistically significant" judgment ($P \leq 0.01$) for 13 of them, a "statistically significant" judgment ($P \leq 0.05$) for 5, and a "not statistically significant" judgment ($P > 0.05$) for the remaining 3 (all related to the GI or HM feature sets). These tests confirm that the use of OM-specific lexical resources has a significant impact on the overall performance, not due to chance.

We have then compared the SWN1 and SWN2 experiments with the GI and HM experiments, obtaining, in the 12 tests, 6 "highly statistically significant" judgments, 3 "statistically significant" judgments, and 3 "not statistically significant" judgments, confirming that SentiWordNet-based feature better results are not due to chance.

Last, 5 tests out of 6 comparing the experiments based on the two versions of SentiWordNetreturned a "not statistically significant" answer, leaving unanswered the question on which of the two versions is better when put in use.

5.3 Inter-Annotator Agreement

Wiebe et al. [7] report an IAA study on the MPQA corpus. They evaluate IAA by using the AGR measure, which estimates the agreement between two independent annotators C_1 and C_2 by computing the mean between the precision

[5] The five experiments using the subjectivity features for the MPQA corpus plus the two on the I-CAB Opinion corpus, all multiplied by the three significance tests performed: the s-test, and the p-test on precision and recall.

Table 3. IAA study on I-CAB Opinion. The first two columns indicate the number of annotations for the various tags, while the remaining columns indicate the IAA results according to various IAA models.

	# annotations		Overlap		Token&Separator	
	A	B	AGR	F_1	κ	F_1
AGENT	1239	859	.539	.521	.439	.472
DIRECT-SUBJ.	263	246	.507	.507	.414	.422
EXPRESSIVE-SUBJ	924	467	.602	.537	.339	.357
OBJECTIVE-SPEECH-EV.	132	144	.501	.500	.462	.465
INSIDE	491	563	.767	.763	.718	.791

$\pi(C_1, C_2)$ and recall $\rho(C_1, C_2)$. They report IAA values only for two tags out of the five defined in the WWC markup language. They measure IAA by averaging AGR, calculated pairwise among three annotators, on a set of 13 documents. The measured agreement is 0.72 on the EXPRESSIVE-SUBJECTIVITY tag and 0.82 on the DIRECT-SUBJECTIVE tag.

In Table 3 we report IAA results on I-CAB Opinion, which can be useful for the interpretation of the overall performance obtained in our experiments. The IAA investigation on I-CAB Opinion corpus consisted in having an intern (a third-year student in Computers and the Humanities – Annotator C_2) independently annotate 127 (94 training and 33 test) articles of the total 525 articles articles already annotated by the third author of [9] (Annotator C_1). IAA is here measured using various measures: AGR, the F_1 measure computed using the **overlap** match predicate between annotations, Cohen's κ, and F_1 computed using the token & separator model. Cohen's κ [22,23] is a widely adopted IAA measure computed as

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)} \tag{1}$$

where $P(A)$ is the observed probability of agreement between the two annotators and $P(E)$ is the probability of agreement by chance; they are defined as

$$P(A) = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$P(E) = \frac{(TP + FP)(TP + FN) + (TN + FP)(TN + FN)}{TP + TN + FP + FN} \tag{3}$$

and are typically estimated by using values in the contingency table (see Section 4.3) computed by matching the annotations from the two annotators.

The IAA values for I-CAB Opinion, which are much lower than those obtained on MPQA in [7], indicating that I-CAB Opinion is a "harder" datasets than MPQA, which explains the lower performance of the automatic extraction system on I-CAB Opinion with respect to its performance on MPQA.

6 Conclusions

We have presented an experimental comparison, performed in the context of an opinion extraction task, among "opinion-oriented" lexical resources. In these experiments SentiWordNet has delivered substantially better performance than the HM and GI lexicons; conversely, between two tested versions of SentiWordNet no clear winner has emerged. The simultaneous use of all tested lexical resources has brought about a noteworthy improvement with respect to the use of any single such resource; this suggests that none of the various tested resources is "complete", indicating that there is still room for their improvement.

References

1. Breck, E., Choi, Y., Cardie, C.: Identifying expressions of opinion in context. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI 2007), Hyderabad, IN, pp. 2683–2688 (2007)
2. Choi, Y., Breck, E., Cardie, C.: Joint extraction of entities and relations for opinion recognition. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006), Sydney, AU (2006)
3. Choi, Y., Cardie, C., Riloff, E., Patwardhan, S.: Identifying sources of opinions with conditional random fields and extraction patterns. In: Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005), Vancouver, CA, pp. 355–362 (2005)
4. Kim, S.M., Hovy, E.: Identifying opinion holders for question answering in opinion texts. In: Proceedings of the AAAI 2005 Workshop on Question Answering in Restricted Domains, Pittsburgh, US (2005)
5. Kim, S.M., Hovy, E.: 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 (2006)
6. Wiebe, J., Breck, E., Buckley, C., Cardie, C., Davis, P., Fraser, B., Litman, D., Pierce, D., Riloff, E., Wilson, T., Day, D., Maybury, M.: 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, pp. 12–19 (2003)
7. Wiebe, J., Wilson, T., Cardie, C.: Annotating expressions of opinions and emotions in language. Language Resources and Evaluation 39(2/3), 165–210 (2005)
8. 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)
9. Esuli, A., Sebastiani, F., Urciuoli, I.: Annotating expressions of opinion and emotion in the Italian Content Annotation Bank. In: Proceedings of the 6th Conference on Language Resources and Evaluation (LREC 2008), Marrakech, MA (2008)
10. Esuli, A., Sebastiani, F.: Evaluating information extraction. In: Agosti, M., Ferro, N., Peters, C., de Rijke, M., Smeaton, A. (eds.) CLEF 2010. LNCS, vol. 6360, pp. 100–111. Springer, Heidelberg (2010)
11. Riloff, E.: Automatically generating extraction patterns from untagged text. In: Proceedings of the 13th Conference of the American Association for Artificial Intelligence (AAAI 1996), Portland, US, pp. 1044–1049 (1996)

12. Kudo, T., Matsumoto, Y.: Fast methods for kernel-based text analysis. In: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003), Sapporo, JP, pp. 24–31 (2003)
13. 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)
14. 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)
15. Esuli, A., Sebastiani, F.: Random-walk models of term semantics: An application to opinion-related properties. In: Proceedings of the 3rd Language Technology Conference (LTC 2007), Poznań, PL, pp. 221–225 (2007)
16. Magnini, B., Pianta, E., Girardi, C., Negri, M., Romano, L., Speranza, M., Bartalesi-Lenzi, V., Sprugnoli, R.: I-CAB: The Italian content annotation bank. In: Proceedings of the 5th Conference on Language Resources and Evaluation (LREC 2006), Genova, IT, pp. 963–968 (2006)
17. Freitag, D.: Using grammatical inference to improve precision in information extraction. In: Proceedings of the ICML 1997 Workshop on Grammatical Inference, Automata Induction, and Language Acquisition, Nashville, TN (1997)
18. Lavelli, A., Califf, M.E., Ciravegna, F., Freitag, D., Giuliano, C., Kushmerick, N., Romano, L., Ireson, N.: Evaluation of machine learning-based information extraction algorithms: Criticisms and recommendations. Language Resources and Evaluation 42(4), 361–393 (2008)
19. Pianta, E., Bentivogli, L., Girardi, C.: MultiWordNet: Developing an aligned multilingual database. In: Proceedings of the 1st International Conference on Global WordNet (GWN 2002), Mysore, IN (2002)
20. Yang, Y., Liu, X.: A re-examination of text categorization methods. In: Proceedings of the 22nd ACM International Conference on Research and Development in Information Retrieval (SIGIR 1999), Berkeley, US, pp. 42–49 (1999)
21. Spiegel, M.R., Stephens, L.J.: Statistics, 3rd edn. McGraw-Hill, New York (1999)
22. Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20(1), 37–46 (1960)
23. Eugenio, B.D., Glass, M.: The kappa statistic: A second look. Computational Linguistics 30(1), 95–101 (2004)