Towards Context-Based Subjectivity Analysis

Farah Benamara
IRIT-CNRS
Toulouse
benamara@irit.fr

Baptiste Chardon
Synapse Développement
Toulouse
baptiste.chardon
@synapse-fr.com

Yannick Mathieu
LLF-CNRS
Paris
yannick.mathieu@
linguist.jussieu.fr

Vladimir Popescu
IRIT-CNRS
Toulouse
popescu@irit.fr

Abstract

We propose a new subjectivity classification at the segment level that is more appropriate for discourse-based sentiment analysis. Our approach automatically distinguish between subjective non-evaluative and objective segments and between implicit and explicit opinions, by using local and global context features.

1 Introduction

Subjectivity and polarity classification is one of the most studied research area in opinion analysis (Pang and Lee, 2004; Wiebe and Riloff, 2005; Wilson et al., 2009). The first task generally distinguish between objective and subjective statements. Polarity classification is then performed in order to extract positive, negative and possibly neutral statements. These two tasks are codependent, since subjectivity analysis filters out statements that contain no opinion.

A common approach in these tasks is to rely on the prior polarity of words and expressions as encoded in external lexical resources. However, as (Polanyi and Zaenen, 2006) stated, identifying prior polarity alone may not suffice to improve sentiment analysis at a finer grain, we need both *local* and *global* context. Context provided *locally* can help in two ways. First, it can be used in subjectivity word sense disambiguation (SWSD) in order to determine if a given word has a subjective or an objective sense (Akkaya et al., 2009). It can also be used to identify valence shifters (viz. negations, modalities and intensifiers) that strengthen, weaken or reverse the prior polarity of a word or an expression (Kennedy and Inkpen, 2006; Wilson et al., 2009). *Global* context on the other hand can be used to identify implicit opinions and to improve the recognition of the overall stance.

Few research efforts have been undertaken on using discourse as features for sentence / clause-based opinion analysis. Among them, (Pang and Lee, 2004) assume that subjective and objective sentences are more likely to appear together, (Asher et al., 2008) have developed an annotation schema for a fine-grained contextual opinion analysis using discourse relations, (Taboada et al., 2008) have used a Rhetorical Structure Theory discourse parser in order to calculate semantic orientation by weighting the nuclei more heavily, and finally, (Somasundaran, 2010) has proposed a discourse-level treatment to improve sentence-based polarity classification and to recognize the overall stance. More recently, (Zhou et al., 2011) proposed an unsupervised method to recognize RST-based discourse relations for eliminating intra-sentence polarity ambiguities. However, no work has investigated so far how discourse structure can be used to enhance subjectivity analysis (SA).

Using discourse for SA raises new issues: *Is sentence/clause subjectivity-based analysis appropriate? Is binary subjective vs. objective classification enough for capturing how opinions are expressed within discourse?* and finally, *how can rhetorical relations help to correctly identify subjective orientation at a finer-grained level?* In this paper, we aim to answer these questions.

2 Context-Based SA: New Challenges

2.1 Segment-Based SA

The sentence level is not appropriate for context-based SA, since, in addition to objective clauses, a single sentence may contain several opinion

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 1180–1188,
Chiang Mai, Thailand, November 8 – 13, 2011. ©2011 AFNLP

clauses that can be connected by rhetorical relations. Moving to the clause level is also not appropriate, since several opinion expressions can be discursively related as in *The movie is great but too long* where we have a *Contrast* relation or as in *Mr. Dupont, a rich business man, has been savagely killed* where we have an *Elaboration* because the appositive gives further information about the eventuality introduced in the main clause. Therefore, we need to move to a finer-grained analysis, at the segment level. (Somasundaran et al., 2007) have used a similar level to detect the presence of sentiment and arguing in dialogues. However, segment annotations were provided by their corpus, whereas in our case, segments are defined according to the Segmented Discourse Representation Theory (SDRT) (Asher and Lascarides, 2003) and are automatically detected.

2.2 Beyond Binary Classification

SA can not be simply reduced to binary subjective vs. objective classification. The following examples extracted from our corpus of French movie reviews illustrate this (they are translated in English and discourse segments are between []):

(1) [The movie is not bad,]$_a$ [although some persons left the auditorium]$_b$
(2) [Laborious]$_a$ [and copy/paste of the first part.]$_b$
(3) [This movie is poignant,]$_a$ [and the actors excellent.]$_b$ [It will remain in your DVD closet.]$_c$
(4) [I suppose]$_a$ [that the government policy failed]$_b$

Segments (1.a), (2.a), (3.a), (3.b) and (4.b) are *explicit opinions*. (1.b), (2.b) and (3.c) convey *implicit opinions* and (4.a) is *subjective*, but *non-evaluative*. (Wiebe et al., 2005) have already proposed an expression-level annotation scheme that distinguishes between explicit mentions of private states, speech events expressing private states, and expressive subjective elements. (Liu, 2010) has also observed that subjective sentences and opinionated sentences (which are objective or subjective sentences that express implicit positive or negative opinions) are not the same, even though opinionated sentences are often a subset of subjective sentences. We follow the same observations and we propose a new subjectivity classification at the segment level that is more appropriate for discourse-based sentiment analysis. We automatically classify each segment into four classes, namely *S*, *OO*, *O* and *SN*, as defined below.

Definition 1. *S segments* are segments that contain explicitly lexicalized *subjective and evaluative* expressions. Their polarity can be positive (as in (1.a)), negative (as in (2.a)) or neutral in the sense that their positivity/negativity depends on the context (as in (3.a)).

Definition 2. *OO segments* are positive or negative opinions implied in an objective segment. They do not contain any explicit subjective clues and are objective out of context [1].

Definition 3. *O segments* do not contain any lexicalized subjective term, neither do an implied opinion.

Definition 4. *SN segments* are subjective, but non-evaluative segments that are used to introduce opinions. In general, these segments contain verbs that are used to report the speech and opinions of others. It is important to note that *SN* does not cover the cases of neutral opinion.

These classes have several advantages over standard binary classification. First, they allow us to distinguish between purely subjective expressions (*S*) and implicit subjective expressions (*OO*). Secondly, our classes can be used to enhance polarity classification, since they allow for the removal of the *O* and *SN* segments, which do not convey any positive, negative or neutral opinion. Finally, our classes can also be used to enhance the overall opinion strength assessment. *SN* segments, especially in news articles, can play an important role since they convey the degree of veracity of the information and the degree of the commitment of the author and of the writer.

Recently, some efforts have been done on the automatic identification of implicit sentiments. For example, (Greene and Resnik, 2009) used lexical semantics and syntax. (Muşat and Trăuşan-Matu, 2010) investigated the influence of valence shifters on the identification of implicit sentiment in economic texts. However, to our knowledge, as yet no work proposed to automatically distinguish between evaluative and non-evaluative segments on the one hand, and between implicit and explicit opinions on the other hand, by using contextual features.

2.3 Rhetorical relations and SA

Using SDRT as a formal framework, we have the following discourse relations: *Contrast*(*a*,*b*) in (1) marked by *although*, *Continuation*(*a*,*b*) in

[1] This definition does not take into account implicit opinions conveyed in subjective segments, such as metaphors.

(2) marked by *and*, and *Attribution*(a,b) in (4).We observe in our corpus that segments related by a *Contrast* or *Continuation* relation often share the same subjective orientation (about 80 %). However, discourse connectors are not the only indicator for deciding whether a segment is opinionated or not. Indeed, some connectors can introduce several discourse relations. In addition, relations are not always explicitly marked, as in (3) where the implicit opinion conveyed in segment c is linked to the subjective segments a and b by a *Result* relation. Another problem is how segments are attached within the discourse structure. In (3), we have *Continuation*(a,b) and *Result*([a,b], c) where [a,b] is a complex segment. Therefore, the subjectivity of segment c depends on [a,b].

Using discourse in opinion analysis is thus a complex task. As a preliminary step, we propose to study the influence of contextual features using mostly lexically-marked discourse relations, and, crucially, without relying on any existing discourse relation annotated corpora. We do not use complex segments and we assume that each segment is only attached to the nearby segment on the left or on the right. In the next sections, we first present the data and our subjective lexicon. Sections 5 and 6 detail respectively the segmentation algorithm and the classification strategies. Section 7 presents the experiments we carried out and discusses the results.

3 Data and Annotations

Our corpus is composed of 136 French movie reviews extracted from the Allô Ciné web site (115 are used for development (our gold) and 21 for test). Three judges performed a two step annotation: first segmentation and then segment classification. Segmentation consists in finding elementary discourse units (EDUs). EDUs typically correspond to verbal clauses, but also to other syntactic units describing eventualities, adjuncts (like appositions or frame adverbials), nonrestrictive relatives and appositions (for embedded EDUs). In case of *S* EDUs, we observe that several opinion expressions (often conjoined NPs or APs clauses) can be related by discourse relations.We resegment such EDUs into separate clauses – for instance [*the film is beautiful and powerful*] is taken to express two segments: [*the film is beautiful*][*and powerful*]. For segment annotation, we rely on an already existing annotation guide elaborated during the ANNODIS project (Afantenos et al., 2010) that shows that segmentation is a relatively easy task even for naives. In order to avoid errors in determining the basic units, segmentation relies on annotation consensus.

For segment classification, we elaborated a specific annotation guide where we ask the judges to annotate each EDU into *S*, *OO*, *O* and *SN* according to the definitions given in Section 2.2. First, the judges were trained to the task and discussed while annotating the same documents (10 reviews that were subsequently discarded from the gold). Then, they separately doubly-annotated each review. This yielded an average Cohens kappa of 0.7 for *S*, 0.72 for *O*, 0.61 for *SN* and 0.54 for *OO*. The latter two are moderate agreements and figure, we believe, an artifact of the length of the texts. Indeed, the longer a text is, the higher difficulty for human subjects is in detecting discourse context in longer texts. However, the study of this hypothesis falls out of the scope of this paper and is therefore left for future work. Nonetheless, these figures are well in the range of state-of-the-art research reports in distinguishing between explicit and implicit opinions (Toprak et al., 2010). For our experiments, the conflicting cases were resolved through discussion between annotators.

4 Subjective Lexicon

Our lexicon is composed of 270 verbs, 632 adjectives, 296 nouns, 594 adverbs, 51 interjections, 178 opinion expressions, with 95 modalities among all these. Since there is no existing free subjective lexicon for French, we have manually built our own lexicon from the study of a wide variety of corpora.Following the opinion categorization described in (Asher et al., 2008), each entry (except for adverbs) is associated to four high-level semantic categories (namely *reporting*, *judgement*, *sentiment* and *advice*) and to 24 subcategories. For adverbs, we use additional categories: *negation, affirmation, doubt, intensifier* and *manner*. Only adverbs of manner express opinions, the other adverbs are used as valence shifters.

We manage both polarity and sense ambiguities. We do not fix the polarity of entries that may have context-dependent polarity orientations. Instead, we list all possible orientations (for example, the entry *long* has both a positive and a negative polarity). In order to detect if a subjective entry from our lexicon is employed in an objective sense, we coupled our lexicon to an external French dictionary D that manually encodes the senses of more than 77 678 words and expressions depending on syntactic configurations For example, the French adjective "*noble*" (noble) has three senses: (a) "noblesse" (pertaining to the aristocracy), (b) "précieux" (precious) and (c) "élevé" (lofty). For each entry E_L in our lexicon, we manually look for its corresponding senses in D as follows: if $E_L \in D$ then if $SubjSense_{E_L} \subseteq Sense_{E_L}$ then add to our lexicon the set $SubjSense_{E_L}$. Thus, for *noble* we only retain (b) and (c) as subjective senses. This dictionary is used by the Cordial syntactic parser (Laurent et al., 2009) in order to perform SWSD. If the identified sense found by the parser is encoded in our lexicon, then the word has a subjective sense; otherwise, it has an objective sense.

5 Automatic Discourse Segmentation

The segmentation is carried out using a set of lexical and syntactic features as described in (Afantenos et al., 2010). These features include the distance from sentence boundaries, the dependency path, and the chunk start/end. Since we used a different syntactic parser, we modified certain features accordingly, and discarded others. We performed a two-level segmentation. First, we constructed a feature vector for each word token, which is classified into: R (Right) for words starting an EDU, L (Left) for tokens ending an EDU, N (Nothing) for words completely inside its EDUs, and B (Both) for tokens which constitute the only word of an EDU.

In the second step, the EDUs which contain at least one token that belongs to our subjective lexicon were retained for a further segmentation. The latter is easier than the segmentation performed at the first level, because we do not encounter embedded segments. Thus, the second-level segmentation of EDUs comes down to searching for one or more "cut points" therein. Since the proportion of EDUs that need to be resegmented is relatively low (about 12 % in the gold standard), we carried out this step by using symbolic rules. These are mainly based on discourse markers.

We performed a supervised learning by using the MegaM software package[2], based on the Maximum Entropy model (Berger et al., 1996) in order to classify each segment into the R, L, Nothing or Both classes, as described above. We carried out a 10-fold cross-validation on our gold standard and an evaluation on the Test data. Table 1 shows first-level EDUs segmentation results for the Right, the Left and Nothing boundaries. For the symbolic segmentation, we evaluated our results (i) on the gold and (ii) on the Test data. The F-measures for boundary recognition are 97.88 % in (i) and 98.65 % in (ii); for the internal-boundaries, we have 84.17 % in (i) and 84.68 % in (ii). Finally, for the new EDU recognition we obtain 77.23 % in (i) and 76.31 % in (ii).

6 Classifiers

The classes, *S*, *O*, *SN* and *OO* are unbalanced in the development corpus. Besides, getting the *OO* segments right is far from obvious, sometimes even for humans. This is why we have defined two orthogonal binary sets of classes:

(a) S_NC vs. O_NC where S_NC = $S \cup SN$ and O_NC = $O \cup OO$ which distinguish between *subjective non-contextual* segments, which are intrinsically subjective, irrespective of their context of occurrence and *objective non-contextual* segments, which, in the absence of any context, are intrinsically objective.

(b) Eval_Op vs. Non_EvalOp where Eval_Op = $S \cup OO$ and Non_EvalOp = $O \cup SN$ which distinguish between *evaluative* and *opinionated* segments, which, given the appropriate context, contain an explicit or an implicit opinion and *non-evaluative* and *non-opinionated* segments, which, irrespective to the context, are not evaluative.

On the gold standard, this grouping yields 919 S_NC EDUs, 511 O_NC EDUs and 1083 Eval_Op EDUs, 347 Non_EvalOp EDUs. Two binary classifiers are constructed, one for each of the two binary sets of classes, defined above: an "S" classifier for **(a)** and an "Op" classifier for **(b)**. Given that the binary sets of classes represent two mutually independent re-partitionings of the segment space, the classifiers are independent of one another. Hence, they can be run in parallel. Then, their outputs are used, via a simple set of four rules, to yield the original four EDU classes. The rules are:

- IF an EDU is S_NC AND Eval_Op, then it is *S*;
- IF an EDU is S_NC AND Non_EvalOp, then it is *SN*;
- IF an EDU is O_NC AND Eval_Op, then it is *OO*;
- IF an EDU is O_NC AND Non_EvalOp, then it is *O*.

The two orthogonal classifiers introduced above

[2] http://www.cs.utah.edu/~hal/megam/

	Precision			**Recall**			**F-measure**		
	R	**L**	**Nothing**	**R**	**L**	**Nothing**	**R**	**L**	**Nothing**
Gold	94.97	93.80	97.88	93.41	94.14	98.10	94.18	93.97	97.99
Test	92.61	93.47	94.05	81.79	78.69	98.15	86.86	85.45	96.06

Table 1: First-level EDU segmentation results (in percents)

operate on the same input text. Hence, to get the class of each EDU in the input text, it suffices to perform a fusion of the results of these classifiers, via the four rules shown above. Each classifier is based on SVMs ("Support Vector Machines") (Burges, 1998). From each EDU a distinct feature vector is computed for each classifiers.

6.1 Feature Set

The features used are described in Table 2. They have been grouped in "Local" and "Contextual" features, according to whether they rely on adjacent EDUs or not. All the features are binary.

Local Features. They have been grouped in "Lexical", "Stylistic" and "Syntactic", according to whether they rely on lexical information only, on stylistic or on syntactic information. We have three lexical features. The first one concerns the presence in an EDU of a noun, adjective, adverb of manner, verb, expression or interjection that belongs to the lexicon, excluding entries expressing modalities and negations. The second feature refines the previous one by taking into account only those lexical entries that have subjective senses as described in section 4. The last lexical feature checks the presence of modals in the lexicon.

"Stylistic" features look for emoticons, words in capital letters and for specific punctuation marks. For emoticons, we rely on a dictionary of 79 emoticons. Capitalization is extracted by taking care to filter out certain standard acronyms, such as DVD. For punctuations, we look for sequences of punctuation marks, such as "??", "!!", "?!", or "!?".

We have five "syntactic" features. The first one looks for comparatives and relative superlatives using a set of manually-built French language-specific comparative and superlative patterns. The second one checks the presence of verbs in the "reporting" category, or in the "advice" category that do not have prior polarity, and by seeing whether the arguments the verbs are in the EDUs or not. The next feature gets the scoping of the modals via a syntactic (dependence) analysis of the EDU. The fourth local syntactic feature is extracted by using a set of manually-built typical French syntactic patterns that bear a subjective meaning. These patterns also allow for some flexibility since other words might be intercalated. The last feature is also detected via a syntactic analysis of the text in order to check if an EDU is left-detached place or time (circumstantial) complement (CC) since these EDUs are mainly objective.

Contextual Features. These features have been grouped in two subtypes, "Non-discursive" and "Discursive". Non-discursive features test the presence of a reporting or non-polar advice verband we test it on the EDU that occurs *before* (i.e., to the left of) the current EDU. The second group of contextual features refer to discourse constraints. The first one checks for the *unmarked Commentary* relation between the current EDU and the next one, when the latter contains an emoticon. If this happens, then that previous EDU is Eval_Op. The next feature checks for the simultaneous presence of two marked SDRT rhetorical relations, in the set {*Continuation*, *Parallel*, *Contrast*, *Alternation*}, one between the current EDU and the previous one, and the other between the current EDU and the next one, with the previous and the next EDUs being in the Eval_Op class. In order to determine the presence of these rhetorical relations, we rely on a French lexicon of discourse connectors, developed with the SDRT rhetorical relations in mind (Roze et al., 2010). The feature that follows is a relaxation of the previous one, in that it applies when at least one marked rhetorical relation is found, between the previous EDU and the current one, or between the latter and the next one. The last feature is based on the empirically-motivated intuition that, in general, in reviews, the last EDU tends to be the second argument of a *Result* relation between (some EDUs in) the rest of the document and itself. As such, this last EDU tends to be in the Eval_Op class.

6.2 Getting the Discursive Features

For computing the two discursive features in a current EDU that are based on discourse markers (henceforth, "DFM"), we rely on an already

Scope	Type	Description	S	Op
Local	Lexical	Subjective expression from the lexicon	√	√
		Semantically-disambiguated subjective expression	√	√
		Modal	—	√
	Stylistic	Emoticon	√	√
		Punctuation marks	—	√
		Word in capital letters	—	√
	Syntactic	Relative superlative or comparative	—	√
		Reporting, or non-polar advice verb, with the argument not in the EDU	√	√
		Word modified by a modal	—	√
		Syntactic pattern	—	√
		EDU left-detached place or time complement	—	√
Contextual	Non-discursive	Reporting, or non-polar advice verb, in the *previous* EDU	—	√
	Discursive	Emoticon in the *next* EDU	—	√
		(Discourse marker in the current EDU and *previous* EDU is Eval_Op) and (discourse marker in the *next* EDU and the *next* EDU is Eval_Op)	—	√
		(Discourse marker in the current EDU and *previous* EDU is Eval_Op) or (discourse marker in the *next* EDU and the *next* EDU is Eval_Op)	—	√
		Current EDU is the last in a document	—	√

Table 2: Features for the two classifiers: S and Op

available Op classification of the previous and/or next EDUs. Of course, for a raw input text, such a classification is not available. Hence, we have devised an iterative procedure for the Op classifier, which first starts with an Op classification by using all the features in Table 2, except for DFM. This provides a first Op classification of the input EDUs, which is used for *bootstrapping* a second iterative Op classification of the EDUs, this time by using all the features in Table 2.

In order to guarantee the convergence of the procedure, we rely on the idea that the goal of the Op classification is mainly to detect *OO* EDUs among intrinsically *O* EDUs. Thus, from the perspective of the Op partitioning of the EDU space, the classification is supposed to start with all the input EDUs as Non_EvalOp, and then to move the appropriate ones into the Eval_Op class. This boils down to imposing a constraint on the second Op classification in the iterative procedure, namely, that it does not alter the class of the EDUs which had already been classified as Eval_Op. The stopping criterion consists in the stabilization of the F-measure of the classifier with respect to the initial test data. The procedure assumes that both classifiers (the bootstrapping one and the iterative one) have been trained on the same data, except that the feature vectors are defined as appropriate for each classifier; the DFM features are determined, in the training phase, by relying on the gold annotation of the training EDUs.

The procedure goes as described below, where $bootstrp\_vs$ is the set of bootstrapping feature vectors, and $curr\_vs(i)$ is the set of input feature vectors at iteration i. $preds(i)$ are the predicted class labels of the respective SVM classifier, at iteration i; $\leftarrow$ is the assignment operator; $F\_score(A, B)$ is the F-measure between the class labellings of a list of feature vectors, A, and a list of class labels, B, both lists having the same length. $\oplus$ is an operator that takes the same types of arguments as F_score, A and B, and implements the filter on the Eval_Op EDUs ensuring the convergence of the iterative procedure ($length(A)$ is the number of elements in list A). It is defined as:

```
$A \oplus B$ ::=   for $i$ from 0 to $length(A)$:
                       if $class(A[i]) =$ Non_EvalOp:
                           $class(A[i]) \leftarrow B[i]$.
```

We call ϵ the "convergence factor", a threshold of the F-measure variation from one iteration to another. MAX_ITER is the maximum number of iterations if convergence is not achieved before. The procedure is:

```
for an input test document $test$:

  1. compute $bootstrp\_vs$, with all features except for $DFM$;
  2. apply the bootstrapping classifier on $bootstrp\_vs$;
     obtain thus $preds(0)$ and
     $F\_score(0) \leftarrow F\_score(bootstrp\_vs, preds(0))$;
  3. compute $DFM$ by using $preds(0)$;
     obtain thus $curr\_vs(0)$;
  4. for $n$ from 1 to $MAX\_ITER$:
       4.1 $curr\_vs(n) \leftarrow curr\_vs(n-1) \oplus preds(n-1)$
       4.2 apply the iterative classifier on $curr\_vs(n)$;
           obtain thus $preds(n)$ and
           $F\_score(n)$ $\leftarrow$
```

```
        $F\_score(curr\_vs(n), preds(n))$;
    4.3 if $||F\_score(n) - F\_score(n-1)|| \leq \epsilon$:
        STOP
5. compute $F\_score(preds(n), test)$.
```

7 Experiments and Results

Several experiments were performed for testing the validity of our subjectivity classification approach, and especially of the contextual features. Thus, first the two classifiers are assessed in a 10-fold cross-validation on the development corpus. Secondly, the two classifiers are evaluated on the 21-document test corpus, with the entire development corpus used for training. For both setups, we have used the SVM-light software package[3]. Due to the fact that the feature vector spaces were found to be non-linearly-separable for both the S and Op classifiers, training was performed by using polynomial kernels. The Op classifier is evaluated in two manners when DFM features are used: first, the values of these features are drawn from the class annotation of the EDUs as given in the manually annotated corpus. Secondly, the iterative approach has been used, in order to test the approach in the real-life scenario when the contextual features cannot be detected by relying on a prior annotation of the input data. In all situations, the four rules introduced in Section 6 are used on the results of the two classifiers for inferring the finer-grained class of each EDU.

7.1 Evaluation of the Classifiers

We first present, in Table 3 the results of both classifiers, both in 10-fold cross-validation on the development corpus ("Gold"), and on the test data ("Test"). We first start with baseline feature sets, to which several features are progressively added; this is marked by the "+" sign. The best performances are marked in boldface. For the S classifier, our baseline considers only emoticons, all entries from the lexicon, except adverbs of manner or negation as well as modalities, along with the presence of a reporting verb with no argument. We observe that when adverbs of manner are added, all the performance figures improve, on both Gold and Test data (although a slight loss in precision is noticed on the new data). We also observe that adding SWSD yields the best performance figures for S.

For the Op classifier, our baseline uses local and syntactic features which rely on our lexicon: the presence of a subjective word or an emoticon or a modal or a word in the scope of a modal. Adding stylistic features provides a slight improvement of all the performance figures on the Gold but a slight degradation on the Test data. This might be due to a less regular way of using punctuation marks. The use of the feature referring to the presence of comparative and superlative patterns in the Gold slightly degrades precision, but provides the best recall of all the feature combinations for the Op classifier. On the Test data, no change in the measures is recorded. When syntactic patterns are added the recall slightly degrades but the accuracy and precision improve, thus providing the best F-score of all the feature combinations on the Gold. However, on the Test data all performance figures slightly degrade. Adding SWSD degrades the recall and, slightly, the F-measure but improves the accuracy and precision; this is true for the Gold and for the Test data. Adding contextual features that do not rely on a prior classification of the context slightly degrades accuracy and precision in the Gold, but provides a more significant improvement in the recall and yields a slightly higher F-score than without these features.

On the Test data, contextual features detect more subjective (explicit or implicit) EDUs than without them. Adding contextual features that rely on a prior classification of the EDUs, provides the best accuracy and precision of all our feature sets which shows their added value. The recall however, worsens on both the Gold and on the Test data (and so does the F-measure), because of the sparseness of the discourse markers in our corpus. Indeed, these last contextual features rely on surface cues which mark only a slight proportion of the discourse relations considered (cf. Section 2.3). This shows that although discourse information seems to be useful in detecting (mostly implicit) subjective EDUs that cannot be detected by other surface means, providing a good coverage is a caveat that could be solved, we believe, through a deeper level of analysis (for example lexical semantic). Finally, adding syntactic information pertaining the EDU being a left-detached CC, and to the presence of a reporting or non-polar advice verb without argument, yields only an improvement of the recall and F-measure, but the accuracy and precision worsen.

In the iterative Op classifier, for the convergence factor $\epsilon = 0.01$, the iterative procedure stops af-

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

Classif.	**Feature set**	**Accuracy**		**Precision**		**Recall**		**F-measure**	
		Gold	Test	Gold	Test	Gold	Test	Gold	Test
S	Baseline	74.3	68.79	77.82	65.5	83.35	79.39	80.49	71.77
	+ Adverbs	74.68	70	77.85	66.18	84.25	81.82	80.92	73.17
	+ Semantic disambiguation	**82.31**	**70.91**	**87.54**	**67.69**	84.08	80	85.77	73.33
non-iter. Op	Baseline	75.82	**73.33**	76.25	73.56	98.93	99.59	86.12	84.61
	+ Capitalized words and punctuation marks	76.51	72.73	76.8	73.39	98.94	98.77	86.48	84.2
	+ Superlatives and comparatives	76.52	72.73	76.73	73.39	99.13	98.77	86.5	84.2
	+ Syntactic patterns	76.66	72.42	76.94	73.31	98.84	98.35	86.53	84
	+ Semantic disambiguation	**77.39**	**73.33**	**81.14**	**78.6**	91.98	87.65	86.22	82.87
	+ Contextual features, no discourse markers	77.32	**75.15**	80.35	79.49	93.31	89.3	86.35	84.1
	+ Contextual features, discourse markers	**78.35**	**75.15**	**83.74**	**85.15**	88.33	80.25	85.97	82.62
	+ EDU left-detached and CC	77.18	74.55	79.15	77.89	94.71	91.36	86.23	84.08
iter. Op	+ Contextual features, discourse markers	77.68	75.45	82.78	84.32	89.02	81.89	85.79	83.08

Table 3: Results (in percents) for the S and Op classifiers

Configuration	***S***		***SN***		***O***		***OO***	
	Gold	Test	Gold	Test	Gold	Test	Gold	Test
Best S / best non-iterative non-contextual Op	80.83	70.6	97.58	96.06	79.35	75.75	72.64	66.06
Best S / best non-iterative contextual Op	81.38	73.33	97.57	92.72	79.96	74.54	79.02	72.12
Best S / best iterative (contextual) Op	81.45	73.03	97.64	93.03	79.21	74.54	77.88	70.9

Table 4: Accuracies (in percents) for the four-class classification

ter at most 2 iterations on the Gold data, and at most 3 iterations on the Test data. As expected, the accuracies and precisions worsen slightly (by around 1 %) on the Test data, since the classes of the adjacent EDUs are not provided beforehand by the gold standard. However, on the Test data the accuracy (but not the precision) very slightly improves (by less than 0.5 %). The recall increases slightly as well on both data sets (by around 1 % as well), which means that the imperfections of the iteratively-obtained classification of the adjacent EDUs somewhat compensates for the limits of these two features themselves.

7.2 Evaluation of the Four Classes

We now show the results of the classification of the EDUs in the four classes *S*, *OO*, *SN* and *O*, obtained by applying the four rules described in Section 6 on the outputs of the S and Op classifiers with the feature sets providing the best performance figures, according to the boldface results in Table 3. For the contextual Op configuration we analyze both the non-iterative (non-iter. Op) and iterative (iter. Op) performance effects on the four classes. The accuracies are synthesized in Table 4. We observe that adding contextual features improves the performance figures, except for the *SN* class, where they degrade by 0.01 % for the non-iterative contextual Op, and, only on the Test data, for the *O* class, where they degrade by 0.21 %. We especially notice the dramatic improvements, on both the Gold and the Test data, for the *OO* class, of implicit subjective EDUs. We thus see that the contextual features provide an improvement of around 5–6 % for the accuracy. The improvements are, as expected, less marked in iter. Op. However, the degradation is rather slight: less than 2 % for the *OO* class and even less for the other classes. Nonetheless, even with the iter. Op, the performance figures remain higher than with the non-contextual Op classifier. Interestingly, in iter. Op, performance figures for the *S* (on the Gold only) and *SN* classes slightly improve.

8 Conclusion and Further Work

In this paper, we have assessed a discourse-based approach to SA. We have proposed a method to distinguish between four types of discourse units, by using both local and global context features. In the future, we plan to annotate opinion documents with SDRT-inspired relations, in order to learn them automatically from several cues other than discourse markers. We believe that the real strength of the discourse-based approach to opinion analysis appears when assessing the global polarity of documents.

Acknowledgements

This work was supported by a DGA-RAPID project under grant number 0102906143. We also thank Stergos Afantenos for his useful help on automatic discourse segmentation.

References

Stergos D. Afantenos, Pascal Denis, Philippe Muller, and Laurence Danlos. 2010. Learning recursive segments for discourse parsing. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 3578–3584.

Cem Akkaya, Janyce Wiebe, and Rada Mihalcea. 2009. Subjectivity word sense disambiguation. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 190–199, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press.

Nicholas Asher, Farah Benamara, and Yvette Yannick Mathieu. 2008. Distilling opinion in discourse: A preliminary study. In *Proceedings of Computational Linguistics (CoLing)*, pages 7–10, Manchester, UK. Association for Computational Linguistics.

Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22:39–71.

Christopher J. C. Burges. 1998. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167.

Stephan Greene and Philip Resnik. 2009. More than words: syntactic packaging and implicit sentiment. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT – NAACL)*, pages 503–511.

Alistair Kennedy and Diana Inkpen. 2006. Sentiment classification of movie and product reviews using contextual valence shifters. *Computational Intelligence*, 22(2):110–125.

Dominique Laurent, Sophie Nègre, and Patrick Séguéla. 2009. L'analyseur syntaxique Cordial dans Passage. In *Proceedings of Traitement Automatique des Langues Naturelles (TALN)*, Senlis, France.

Bing Liu. 2010. Sentiment analysis and subjectivity. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, Boca Raton, FL.

Claudiu Muşat and Ştefan Trăuşan-Matu. 2010. The impact of valence shifters on mining implicit economic opinions. *Lecture Notes in Computer Science*, 6304:131–140.

Bo Pang and Lillian Lee. 2004. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pages 271–278, Stroudsburg, PA, USA. Association for Computational Linguistics.

Livia Polanyi and Annie Zaenen. 2006. Contextual valence shifters. In *Computing Attitude and Affect in Text: Theory and Applications*, volume 20 of *The Information Retrieval Series*, pages 1–10, Berlin-Heidelberg. Springer-Verlag.

Charlotte Roze, Laurence Danlos, and Philippe Muller. 2010. LEXCONN: a french lexicon of discourse connectives. In *Proceedings of Multidisciplinary Approaches to Discourse (MAD)*, pages 114–125, Moissac, France.

Swapna Somasundaran, Josef Ruppenhofer, and Janyce Wiebe. 2007. Detecting arguing and sentiment in meetings. In *Proceedings of the SIGdial Workshop on Discourse and Dialogue*, pages 26–34. Association for Computational Linguistics.

Swapna Somasundaran. 2010. *Discourse-level relations for Opinion Analysis*. PhD Thesis, University of Pittsburgh.

Maite Taboada, Kimberly Voll, and Julian Brooke. 2008. Extracting sentiment as a function of discourse structure and topicality. In *School of Computing Science Technical Report 2008-20*.

Cigdem Toprak, Niklas Jakob, and Iryna Gurevych. 2010. Sentence and expression level annotation of opinions in user-generated discourse. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 575–584, Stroudsburg, PA, USA. Association for Computational Linguistics.

Janyce Wiebe and Ellen Riloff. 2005. Creating subjective and objective sentence classifiers from unannotated texts. In *Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*, volume 3406 of *Lecture Notes in Computer Science*, pages 486–497, Berlin, Heidelberg. Springer-Verlag.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2–3):165–210.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2009. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35(3):399–433.

Lanjun Zhou, Binyang Li, Wei Gao, Zhongyu Wei, and Kam-Fai Wong. 2011. Unsupervised discovery of discourse relations for eliminating intra-sentence polarity ambiguities. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 162–171, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.