Measuring the Effect of Discourse Structure on Sentiment Analysis

Baptiste Chardon[1,2], Farah Benamara[1], Yannick Mathieu[3], Vladimir Popescu[1], and Nicholas Asher[1]

[1] IRIT-CNRS, Toulouse University
[2] Synapse Développement, Toulouse
[3] LLF-CNRS, Paris 7 University
{chardon,benamara,popescu,asher}@irit.fr,
yannick.mathieu@linguist.jussieu.fr

Abstract. The aim of this paper is twofold: measuring the effect of discourse structure when assessing the overall opinion of a document and analyzing to what extent these effects depend on the corpus genre. Using Segmented Discourse Representation Theory as our formal framework, we propose several strategies to compute the overall rating. Our results show that discourse-based strategies lead to better scores in terms of accuracy and Pearson's correlation than state-of-the-art approaches.

1 Introduction

Discourse structure can be a good indicator of the subjectivity and / or the polarity orientation of a sentence. It can also be used to recognize implicit opinions and to enhance the recognition of the overall stance of texts. For instance, sentences related by a *Contrast*, *Parallel* or a *Continuation* relation often share the same subjective orientation, as in *Mary liked the movie. His husband too*, where the *Parallel* relation allows us to detect the implicit opinions conveyed by the second sentence. Polarity is reversed in case of *Contrast* and usually preserved in case of *Parallel* and *Continuation. Result* on the other hand doesn't have a strong effect on subjectivity and polarity is not preserved. For instance, in *Your life is miserable. You don't have a girlfriend. So, go see this movie*, the prior positive polarity of the recommendation follows negative opinions. Hence, *Result* can help to determine the contextual polarity of opinionated sentences. Finally, in case of *Elaboration*, subjectivity is not preserved, in contrast to polarity (It is difficult to say *The movie was excellent. The actors were bad*).

We aim in this paper to empirically measuring the effect of discourse structure on assessing the overall opinion of a document and by analyzing to what extent these effects depend on the corpus genre. To our knowledge, this is the first research effort that empirically validates the importance of discourse for sentiment analysis. Our analysis relies on manually annotated discourse information following the Segmented Discourse Representation Theory (SDRT) [1]. This is a first and a necessary step before moving to real scenarios that rely on automatic annotations

A. Gelbukh (Ed.): CICLing 2013, Part II, LNCS 7817, pp. 25–37, 2013.

(we recall that as far as we know the only existing powerful discourse parser based on SDRT theory is the one that has been developed for a dialogue corpus (Verbmobil corpus [2]). This first step allowed us to show the real added value of discourse in computing both the overall polarity and the overall rating.

2 Related Works

Although rhetorical relations seem to be very useful in sentiment analysis, most extant research efforts on both document-level and sentence-level sentiment classification do not use discourse information. Among the few research reports on discourse-based opinion analysis, let us cite the following. [3] proposed a shallow semantic representation of subjective discourse segments using a feature structure and five types of SDRT-like rhetorical relations. [4] as well as [5] have used an RST discourse parser in order to calculate semantic orientation at the document level by weighting the nuclei more heavily. [6] proposed the notion of opinion frames as a representation of documents at the discourse level in order to improve sentence-based polarity classification and to recognize the overall stance. Two sets of 'home-made' relations were used: relations between targets and relations between opinion expressions. [7] used the semantic sequential representations to recognize RST-based discourse relations for eliminating intra-sentence polarity ambiguities. [8] propose a context-based approach to sentiment analysis and show that discursive features improve subjectivity classification. [9] discuss the application of the Linguistic Discourse Model theory to sentiment analysis in movie reviews. Finally, [10] examine how two types of RST-like rhetorical relations (*conditional* and *concessive*) contribute to the expression of appraisal in movie and book reviews.

We aim here to go further by answering the following questions: (1) *What does the discourse structure tell us about opinion?* (2) *What is the impact of discourse structure when assessing the overall opinion of a document?* (3) *Does our analysis depend on the corpus genre?*. The first question is addressed in section 3 while the last two ones in section 4.

3 Discourse Structure and Opinion

Our data comes from two corpora: movie reviews (MR) taken from *AlloCiné.fr* and news reactions (NR) taken from the politics, economy and international section of *Lemonde.fr* newspaper. In order to guarantee that the discourse structure is informative enough, we only selected movies and articles that are associated to more than 10 reviews / reactions. We also filtered out documents containing less than three sentences. In addition, we balanced the number of positive and negative reviews according to their corresponding general evaluation when available (in NR users were not asked to give a general evaluation). This selection yielded a total of 180 documents for MR and 131 documents for NR.

3.1 Annotation Scheme

Our basic annotation level is the Elementary Discourse Units (EDU). We chose to automatically identify EDUs and then to manually correct the segmentation if necessary. We relied on an already existing discourse segmenter [8] that yields an average F-measure of 86.45%. We have a two-level annotation scheme: *at the segment level* and *at the document level.* Annotators used the GLOZZ platform (www.glozz.org) which provides a discourse graph as part of its graphical user interface.

EDU Annotation Level. For each EDU, annotators were asked to specify its subjectivity orientation as well as polarity and strength; Subjectivity can be either one of the following: **SE** – EDUs contain explicitly lexicalized *subjective and evaluative* expressions, as in *very bad movie*; **SI** – EDUs do not contain any explicit subjective cues but opinions are inferred from the context, as in *The movie should win the Oscar*; **O** – EDUs do not contain any lexicalized subjective term, neither do an implied opinion. **SN** – subjective, but non-evaluative EDUs that are used to introduce opinions, as in the segment *a* in *[I suppose]$_a$ [that the employment policy will be a disaster]$_b$*; and finally **SEandSI** which are segments that contain both explicit and implicit evaluations on the same topic or on different topics, as in *[Fantastic pub !]$_a$ [The pretty waitresses will not hesitate to drink with you]$_b$*. Polarity can be of four different values: +, −, **both** which indicates a mixed polarity as in *this stupid President made a wonderful talk*, and **no polarity** which indicates that the segment does not convey any sentiment. Finally, strength has to be stated on a four-level scale going from 0 to 3 where 0 is the score associated to *O* segments, 1, 2 and 3 respectively indicates a weak, a medium and a strong strength.

Document Annotation Level. First, annotators have to give the overall opinion orientation of the document (the initial star ratings in *MR* corpus were removed) by using a six-level scale, going from −3 to −1 for negative opinion documents and from +1 to +3 for positive ones. Then, they have to build the discourse structure of the document by respecting the structural principles of SDRT, such as the right frontier principle and structural constraints involving complex discourse units (CDUs) (which are build from EDUs in recursive fashion). It's important to recall that SDRT allows for the creation of full discourse graphs (and not trees as in the RST [11]) which allow to capture complex discourse phenomena, such as long-distance attachments and long-distance discourse pop-ups, as well as crossed dependencies.

During the elaboration of our manual, we faced a dilemma: should we annotate opinion texts using a small set of discourse relations, as already done by [3], [6] and [10] or should we use a larger set of discourse relations? Given our goals, we chose the second solution. We used 17 oriented and mostly backward-looking relations grouped into coordinating relations that link arguments of equal importance (*Contrast*, *Continuation*, *Conditional*, *Narration*, *Alternative*, *Goal*, *Result*, *Parallel*, *Flashback*) and subordinating relations that link an important argument to a less important one (*Elaboration*, *E* − *Elab*, *Correction*, *Frame*, *Explanation*, *Background*, *Commentary*, *Attribution*). To deal with the situation where the annotators are not able to decide which relation is more appropriate to link two constituents, we added a relation labeled *Unknown*.

3.2 Results of the Annotation Campaign

Each document of our corpus was doubly annotated by three undergraduate linguistic students who were provided with a complete and revised annotation manual as well as an annotation guide explaining the inner workings of GLOZZ. Annotators were first trained on 12 movie reviews and then they were asked to annotate separately 168 documents from *MR*. Then, they were trained on 10 news reactions. Afterwards, they continued to annotate separately 121 documents from *NR*. The training phase for *MR* was longer than for *NR* since annotators had to learn about the annotation guide and the annotation tool.

Results at the EDU Level. Table 1 gives a quantitative overview (in percents) of the annotations provided by our three annotators. We get a total number of 3478 annotated segments for *MR* and 2150 for *NR*.

Table 1. Quantitative overview of the annotated data (in percents)

	SE	SN	SI	O	SEandSI	+	−	both	no polarity
MR	50	2	29	14	5	45.48	33.78	4	16.74
NR	22	6	49	2	12	17.40	55	4	23.60

The Cohen's Kappa on segment type averaged over the three annotators was 0.69 for *MR* and 0.44 for *NR*. For segment polarity we get 0.74 for *MR* and 0.49 for *NR*. Since the "both" and the *SEandSI* category are very rare in our data, they have been counted with "+" (resp. *SE*). For *MR*, we get very good results for both *SE* (0.79) and the polarity (positive (0.78) and negative (0.77)) of the segment. *SN* class's kappa is also very good (0.73). However, the agreements for the *SI* and *O* classes are moderate (resp. 0.62 and 0.61) because annotators often fail to decide whether a segment is purely objective and thus if it conveys only facts or if a segment holds an implicit opinion. This can also explain the lower kappa measure we get for "no polarity" (0.66). Nonetheless, these figures are well in the range of state-of-the-art research reports in distinguishing between explicit and implicit opinions (see [12]).

For *NR*, our results are moderate for the *SE* and *SN* classes (0.55 for each class) and fair for the *SI* and *O* classes (resp. 0.33 and 0.34). We have the same observations for the agreements on segment polarities where we obtain moderate kappas on all the three classes (0.49). This shows that the newspaper reactions corpus was a bit more difficult to annotate because the main topic is more difficult to determine (even by the annotators) – it can be one of the subjects of the article, the article itself, its author(s), a previous comment or even a different topic, related to various degrees to the subject of the article. Hence, implicit opinions, which are more frequent, can be of a different nature: ironic statements, suggestions, hopes and personal stances, especially for comments to political articles.

We finally compute the inter-annotator agreements on the overall document rating. After collapsing the ratings -1 to -3 and +1 to +3 into respectively positive and negative ratings, we get a kappa of 0.73 for MR and 0.58 for NR for both classes when averaged over our three annotators. We have also observed, that the agreement on extreme points of our six-level scale (namely -3 and +3) are relatively good (for example, we get respectively 0.8 and 0.72 for MR) whereas the kappa on the other points is fair. We get the same observation when computing agreements on segment's strength.

Results at the Discourse Level. Our goal here is to show the importance of discourse for opinion analysis and not to build a discourse bank that examine how well SDRT predicts the intuition of subjects, regardless of their knowledge of discourse theories. Therefore, computing inter-annotator agreements is out of the scope of this paper (for a detailed description of non-expert annotations using SDRT, see [13]). The analysis of the frequency of discourse relations per corpus genre shows that *Continuation* and *Commentary* are the most frequent relations (resp. 18% and 30% for MR and 23% and 24% for NR). However, *Explanation*, *Elaboration*, $E-Elab$ (entity elaboration), *Comment*, *Contrast*, *Result* and *Goal* also have non-negligible frequencies going from 3% to 15% for each corpus genre. These results are essentially stable from one corpus to the other. Also, *Conditional*, *Alternative* and *Attribution* are more frequent in NR than in MR, which is consistent with a logically more structured discourse structure for news reactions than for movie reviews.

We have also analysed the ratio of complex segments to the total number of rhetorical relation arguments in our annotations. We have observed that, for both corpus genres, rhetorical relation instances between EDUs only are a minority and that CDUs are yet more numerous in NR – 56%, than for MR – 53%. This underscores the importance of CDUs for our task. We have finally analysed the impact of rhetorical relations on both subjectivity and polarity of their arguments only in case of relations linking two EDUs. Table 2 gives statistics (in percent) as a / b. *a* stands for on the stability (St) (that is (*SE*, *SE*), (*SI*, *SI*), (*SE*, *SI*) and (*SI*, *SE*)) and the variation (Var) of the subjectivity class (i.e. for the (O, other) and the (other, O) couples, where "other" spans the set of subjectivity classes, other than O). *b* stands for the polarities class but only between subjective (*SN*, *SE*, *SI*) EDUs only : the (+, +) and (–, –) couples for stability and the (+, –) and (+, –) couples for polarity change. We observe that our predictions (as stated in the introduction) are by and large confirmed.

3.3 The Gold Standard

The gold standard used for our experiments was made after discussion between the three annotators. This process was supervised by two experts in discourse

Table 2. Impact of rhetorical relations on both subjectivity and polarity

	MR		*NR*	
	St	**Var**	**St**	**Var**
Continuation	81 / 97	19 / 3	79 / 90	21 / 10
Commentary	61 / 82	39 / 18	75 / 96	25 / 4
Elaboration	50 / 100	50 / 0	82 / 100	18 / 0
Contrast	76 / 15	24 / 85	76 / 59	24 / 41
Result	81 / 100	19 / 0	47 / 100	53 / 0
Attribution	14 / 50	86/ 50	18 / 100	82 / 0
Parallel	100 / 100	0 / 0	73 / 100	27 / 0
Explanation	76 / 80	24 / 20	78 / 83	22 / 17
Frame	39 / 100	61 / 0	47 / 86	53 / 14

analysis and opinion mining. At the EDU level, the main difficulty was to achieve a consensus on implicit and objective segments, especially for NR. At the discourse level, annotators often produce equivalent discourse structures (two of our annotators used to systematically group constituents in CDUs while the others often produced flat structures). While building the gold standard, annotators used CDUs as often as possible. Finally, annotators have to agree on the overall document score. The graph in Figure 1 illustrates an annotation from the gold standard. Segments 1 to 12 are EDUs while segments 13 to 16 are CDUs.

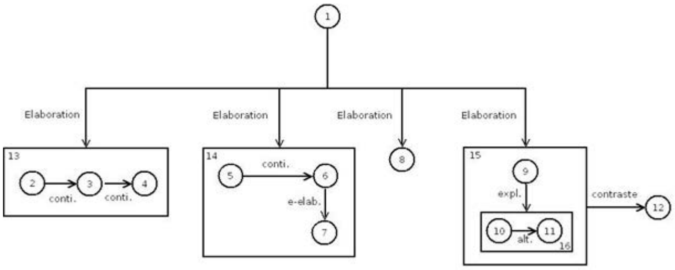

Fig. 1. Two examples of produced discourse annotation

In order to measure the effects of topic information (also called target) to compute the overall opinion, we have asked the annotators to specify, within each EDU, text spans that correspond to the topic. Topic can be of three types: the **main** topic of the document, such as *the movie*, a **partof** topic in case of features related to the main topic, such as *the actors*, and finally an **other** topic that has no mereological relation with the main topic. Once all the topics have been identified, the next step is to link them to the subjective segments of the document. For example, in *[I saw (Grey's Anatomy)_*t_1 *yesterday]_1. [It was boring]_2 [and (the actors)_*t_2 *were bad]_3*, we get $topic(2,\ t_1 : main)$ and *topic*(3, t_2 : *partof*). This annotation was made by consensus due to the difficulty of the task, especially for NR. For MR, the gold standard contains 151 documents, 1905 EDUs (*SE*: 53.85%, *SI*: 26.20%), 1766 discourse relations and 1386 topics (main: 26.26%, partof: 62.62%, other: 11.11%). For NR, we have 112 documents, 835 EDUs (*SE*: 20.24%, *SI*: 51.25%), 924 relations and 586 topics (main: 5.63%, partof: 59.55%, other: 34.81%). The distribution of the overall rating is: 37% positive opinion and 63% negative opinion for MR, versus 33% positive opinion and 67% negative opinion for NR.

4 Computing the Overall Opinion

For each document D, we aim at computing the overall opinion *score_D* of D such as $score_D \in [-3, +3]$. We consider D as an oriented graph ($\aleph$, $\Re$) such that: $\aleph = E \cup C$ is the set of EDUs and CDUs of D and $\Re$ is the set of rhetorical relations that link elements from $\aleph$. $\forall edu \in E$, $edu =< T, S, Val >$ where $T = topic(edu)$ denotes the topic of edu and $T \in \{main,\ partof,\ other\}$, $S = subj(edu)$ is the subjectivity orientation of edu and $S \in \{SE,\ SI,\ O\ , SN\}$ (*SEandSI* segments are considered to be of the *SE* type) and $Val = score(edu)$ is the opinion score of edu stated on the same discrete interval as $score_D$. Each $cdu \in C$ has the same properties as an edu i.e. $cdu =< T_{cdu}, S_{cdu}, Val_{cdu} >$, however, T_{cdu}, S_{cdu} and Val_{cdu} (which is in this case a set of scores) are not given by the annotations but are the result of a reduction process of the cdu to an edu (see Section 4.3).

We propose three strategies to compute $score_D$: (1) **Bag-of-segments (BOS)** that does not take into account the discourse structure. The overall rating is computed using a numerical function that takes the set E as argument and outputs the value $score_D$. (2) **Partial discourse** which takes the discourse graph as input and then prunes it in order to select a subset $\aleph' \subseteq \aleph$ of nodes that are relevant for computing $score_D$. This score is then computed by applying a numerical function only to $\aleph'$. (3) **Full discourse** which is based on the full use of discourse structure where a rule-based approach guided by the semantics of rhetorical relations aggregates the opinion scores of all the elements in $\aleph$ in a bottom-up fashion.

4.1 Bag-of-Segments

Here we consider $D = E = \{edu_1, \ldots, edu_i\}$. In order to evaluate the impact of segments' subjectivity and topic on our task, we propose to filter out some elements of D by applying a subjectivity filter and / or a topic filter. We have three subjectivity filters: $\emptyset$ that keeps all the segments (i.e the filter is not activated), **se** and **si** that respectively keep *SE* and *SI* segments. We also have four topic filters: $\emptyset$ where the filter is not activated, **m** and **p** that respectively keep segments that contain main and part-of topics, and finally **mp** that keeps segments that contain main or part-of topics. Each filter can be applied alone or in sequence with other filters. For example, if we apply *se* and then *m*, we get the subset $D' \subseteq D$

such that $D' = \{edu_i \in D\ ; topic(edu_i) = main\ and\ subj(edu_i) = SE\}$. Filtering can drastically reduce the number of segments in D' ($D' = \emptyset$ or $\forall edu \in D'$ $subj(edu) = O\ or\ subj(edu) = SN$). Hence, some filters are relaxed if necessary.

Computing $score_D$ in the BOS strategy consists in applying a numerical function to all elements of D or to a subset of D obtained after filtering. Let D' be a subset of D. We have seven functions on it: (1) $A(D')$ and (2) $M(D')$, which respectively compute the average and the median of the scores associated to each EDU *edu* in D'. Unlike the average, the median is more suitable in case of skewed distributions. (3) $MSc(D')$ computes the maximum positive scores Max_Pos and the maximum negative scores Max_Neg of elements of D' and then returns $Max(Max\_Pos, Max\_Neg)$. In case of equality, we choose the scores with positive polarity for MR and with negative polarity for NR which correspond to the general polarity orientation of each corpus genre (see section 3). (4) $MSc\_A(D')$ computes $Sc(D')$ when the elements of D' have the same polarity orientation and $A(D')$ otherwise. (5) $Fr(D')$ returns the most frequent opinion score found in D'. In case of equality, it chooses the score that is the closest to the second most frequent score in D'. (6) $Frt(D')$ and (7) $Lst(D')$ returns the score of the first and the last element *edu* of D' such that $subj(edu) = SE$ or $subj(edu) = SI$. We consider here that the order of elements in D' follows the reading order of the document.

4.2 Partial Discourse (PD)

This strategy takes the discourse graph D as input and proceeds by pruning it in order to select the most important nodes for computing the overall rating. We consider two main types of pruning: (a) one based on the distinction between *subordinating* and *coordinating* relations and (b) another one based on *top-level constituents*. (a) can be done either by a **Sub1** pruning that selects from $\aleph$ only EDUs (or CDUs) that are the first argument of a subordinating relation or by a **Sub2** pruning where the selected segments are the first argument of a subordinating relation and at the same time do not appear as the second argument of a subordinating relation. The aim here is to deal with a 'cascade' of subordinations. On the other hand, (b) aims at deleting from $\aleph$ nodes that are right arguments of subordinating relations or nodes that are left arguments of already pruned constituents. Pruning in (b) can be done either by using a **Top1** strategy that preserves all the constituents of the CDUs or by using a **Top2** strategy that reduces CDUs by recursively applying **Top1** to all the elements of the CDU. The resulting set of segments $\aleph' \subseteq \aleph$, obtained after using one of the previous four pruning strategies, can be filtered by using either a subjectivity and / or a topic filter (see Section 4.1).

As in *BOS*, some filters can be relaxed if necessary. It is important to notice that our pruning / filtering process guarantees the connectivity of the graph since the non-selected nodes are not physically removed. Instead, their subjectivity type is set to *O*. $score_D$ is then computed by applying to all the elements of $\aleph'$ one of the seven numerical functions lastly presented.

4.3 Full Discourse (FD)

The third strategy as well has the discourse graph D as input. FD does not prune the graph and does not use any filter but it recursively determines the topic, the subjectivity and the score of each node in a bottom-up fashion. This process is guided by a set of rules that are associated to each rhetorical relation $r(a,\ b) \in \Re$. A rule merges the opinion information of a (i.e $<$ *topic*(a), *subj*(a), *score*$(a) >$) and b (i.e $<$ *topic*(b), *subj*(b), *score*$(b) >$) and computes a triple $< T_{ab}, S_{ab}, Val_{ab} >$ depending on the semantics of r. Since the rules are recursively applied to all nodes of the graph, they thus have to deal with CDUs. For instance, in case we have $r(a,\ b)$ where $a \in C$ and / or $b \in C$, we first need to reduce the complex segment a and / or b by computing its corresponding triple $< T_{cdu}, S_{cdu}, Val_{cdu} >$ using the rules associated to each relation which links the segments belonging to a (resp. b). Let $cdu \in C$ and let $\aleph_{cdu} = E_{cdu} \cup C_{cdu}$ be the set of nodes of the segment cdu and let $\Re_{cdu}$ be the set of relations that link elements of $\aleph_{cdu}$. The reduction process is done in a depth-first traversal of the sub-graph of cdu according to the functions *reduce*(cdu) and *merge*(cdu) defined below:

```
reduce(cdu){
While ($C_{cdu} \neq \emptyset$)
  $\forall cdu' \in C_{cdu}$ reduce($cdu'$)
Let $e \in \aleph_{cdu}$ a left-most node
return(merge($e$)) }
```

```
merge(cdu){
Let $e' \in \aleph_{cdu}$
$\forall r(e,\ e') \in \Re_{cdu}$ and $r$ is subordinating {
    merge($e'$)
     $e$ = ApplyRule($r, e, e'$) }
  If ($\exists r(e,\ e') \in \Re_{cdu}$ and $r$ is coordinating {
  $e$= ApplyRule($r, e, e'$)}
  return($e$) }
```

Once each CDU in C is reduced, we consider the resulting graph as a unique CDU that needs to be reduced again following the same process. The result of the FD strategy is a triple $< T_D, S_D, Val_D >$ containing the overall topic, subjectivity and score of D. Finally, $score_D$ is inferred from Val_D which is a set of scores obtained after reductions. If $|Val_D| = \emptyset$, $score_D$ is not computed, because in this case, the document does not contain relevant opinion instances (e.g. opinions on a topic of the *other* category). Otherwise, if $|Val_D| = 1$, then $score_D = Val_D$, else $score_D = \Gamma(Val_D)$ such that Γ is one of our seven functions.

Drawing on the already established effect on both subjectivity and polarity of the rhetorical relations used in the annotation campaign, we have designed 17 rules (which correspond to *ApplyRule*$(r,\ e,\ e')$ in the *merge* function above). We show below the rule associated to *Contrast*$(e,\ e')$. Until now, $\forall e,\ e' \in E$, if $subj(e) = SE$ (resp. SI) and $subj(e') = SI$ (resp. SE) then $subj(e) = subj(e') = su$.

In addition to the very strong effect of this relation on opinion, we have also observed that this effect may depend on the syntactic order of its arguments. For instance, the overall opinion on the movie is more negative in *The idea is original, but there are some meaningless sequences* than in *There are some meaningless sequences but the idea is original.* Hence, the positivity / negativity of $Contrast(e,\ e')$ is determined by e'. Then, $ApplyRule(Contrast,\ e,\ e') =< T, S, Val >$ where:

– if $topic(e) = topic(e')$ then $T = topic(e)$, if $topic(e) = main$ or $topic(e') = main$ then
$T = main$ (as in *The idea is original, but the movie was bad*),
– $S = su$ if $subj(e) = su$ or $subj(e') = su$, $S = O$ otherwise.
– If $topic(e) = topic(e') = main$ or $topic(e) = topic(e') = partof$, then, if $(score(e) > score(e'))$ then $Val = score(e')$, otherwise $Val = Int^{-}(score(e))$. Finally, if $topic(e) = main$ then $Val = Int^{-}score(e)$, if $topic(e) = partof$ then $Val = Int^{-}score(e')$.

5 Evaluation

We have used these three strategies for each document D of our gold standard. For BOS and FD, we have first applied a subjectivity filter ($\emptyset$, *se* and *si*). Then for each subjectivity filter, we have applied a topic filter ($\emptyset$, *m, p* and *mp*) (the order of application of our two filters does not matter). Consequently, we get 12 configurations corresponding to 12 subsets $D' \subseteq D$ for the BOS and to 12 subsets $\aleph' \subseteq \aleph$ for the PD. If one of these sets is empty, or if it only contains objective segments, we proceed by relaxing some filters (see Section 4.1). For each subset, we have applied one of the seven functions described in Section 4.1. We have thus computed 84 scores per strategy. For the FD strategy, the result set Val_D can be reduced by the same set of functions, thus yielding 7 different computed scores.

We have assessed the reliability of our three strategies by comparing their results (namely $score_D$) against the score given in the gold standard. We have also compared our results against a baseline which consists in applying *BOS* with the subjectivity filter *se* followed by the topic filter $\emptyset$. This baseline is similar to state-of-the-art approaches in rating-inference problems [14] that aggregate the strengths of the opinion words in a review with respect to a given polarity and then assign an overall rating to the review to reflect the dominant polarity. We used two evaluation metrics: accuracy and Pearson's correlation. Accuracy corresponds to the total number of correctly classified documents divided by the total number of documents while Pearson's correlation (r) reflects the degree of linear relationship between the set of scores computed by our strategies and the set given by the gold standard. The closer r is to +1 (or to -1), the better the correlation.

We have performed two experiments: (1) *an overall polarity rating* where we consider that the overall ratings -3 to -1 represent the -1 score (i.e. negative documents) and the ratings +1 to +3 correspond to the +1 score (positive documents); and (2) a *an overall multi-scale rating* where the ratings are considered to be in the continuous interval $[-3, \ +3]$. Among the 84 experiments made for BOS and PD and among the 7 experiments made for FD, Tables 3 and 4 give the configuration that leads to the best results for polarity and multi-scale ratings, respectively. For a strategy s, the notation $(a,\ b,\ c)$ indicates that the given accuracy (resp. correlation) is computed when applying to s the subjectivity filter a followed by the topic filter b and by using the function c. The results below are statistically significance since we get a p-value < 0.01 for reviews corpus and < 0.05 for news reactions.

Table 3. Overall polarity ratings in both corpus genres

	MR		*NR*	
	Accuracy	**Pearson**	**Accuracy**	**Pearson**
Baseline	0.89 (A)	0.81 (A)	0.88 (MSc)	0.52 (MSc)
BOS	0.92 (∅, ∅, Fr)	0.87 (∅, ∅, A)	0.94 (∅, ∅, MSc)	0.77 (∅, ∅, MSc)
Sub1	0.91 (∅, ∅, M)	0.84 (∅, ∅, M)	0.92 (∅, ∅, A)	0.74 (∅, ∅, A)
Sub2	0.91 (∅, ∅, M)	0.84 (∅, ∅, M)	0.92 (∅, ∅, A)	0.74 (∅, ∅, A)
Top1	0.96 (∅, ∅, Fr)	0.94 (∅, ∅, M)	0.92 (∅, ∅, A)	0.77 (∅, ∅, MSc)
Top2	0.90 (∅, ∅, A)	0.80 (∅, ∅, A)	0.96 (∅, ∅, MSc)	0.82 (∅, ∅, MSc)
FD	0.90 (MSc)	0.86 (MSc_A)	0.94 (Fr)	0.82 (MSc_A)

We observe that for assessing the overall polarity, the baseline results for *MR* in term of accuracy (when applying the *average*) are as good as those obtained by other strategies, whereas for *NR*, the results are worse. In terms of Pearson's correlation, we observe that the results are quite good (the baseline beats the *Top 2* strategy when applying the *average*), whereas for *NR* the correlations are not good compared to other strategies. PD strategy beats the *BOS*. For instance, for *MR*, *Top1* outperforms *BOS* by 4% for accuracy and 8% for correlation while for *NR*, *Top2* is the best with more than 2% for accuracy and 5% for correlation. The *FD* strategy is less efficient in *MR* than in *NR* when comparing its results to *BOS*. This difference shows that *FD* is very sensitive to the complexity of the discourse structure. The more elaborate the discourse is, (as in *NR*) the better the results yielded by the rule-based approach are. In addition, for both *BOS* and *PD*, the best combination of filters consists in keeping all segments' types (the *K_all* strategy) and then keeping all the types of topics (*K_all*) (similar results were obtained when applying the other topic filters i.e. *K_M, K_P* and *K_MP*). This entails that both explicit and implicit opinions are important for computing the overall polarity, whereas using topic information does not seem to be very useful. For instance, in *MR*, we get, for *BOS*, an accuracy of 0.84 when applying *K_SI* with the *MaxSc* function and hence – 4 % compared to *K_SE* while for *NR* we get 0.93 when using the *MaxSc* function and hence + 5% over applying *K_SE*. The same holds for the Pearson's correlation. This brings us to the conclusion that the importance of implicit opinions varies, depending on the corpus genre: for movie reviews, more direct and sometimes terse, explicit opinions are better correlated to the global opinion scores, whereas for news reactions, implicit opinions are more important when negative opinions are concerned. This could indicate a tendency to conceal negative opinions as apparently objective statements, which can be related to social conventions (politeness, in particular).

For overall multi-scale ratings, the baselines results are not good compared to the other strategies. In addition, we observe that discourse-based strategies yield better results for both corpus genres. For *MR*, *FD* gives a significant improvement

Table 4. Overall multi-scale ratings in both corpus genres

	MR		*NR*	
	Accuracy	**Pearson**	**Accuracy**	**Pearson**
Baseline	0.63 (Fr)	0.84 (A)	0.60 (MSc)	0.66 (MSc)
BOS	0.63 (se, ∅, Fr)	0.91 (∅, ∅, M)	0.70 (si, ∅, MSc)	0.82 (∅, ∅, MSc)
Sub1	0.63 (∅, p, MSc)	0.90 (∅, ∅, M)	0.69 (si, ∅, MSc)	0.78 (∅, p, A)
Sub2	0.63 (∅, p, Fr)	0.90 (∅, ∅, M)	0.69 (si, ∅, MSc)	0.77 (∅, p, A)
Top1	0.63 (se, ∅, M)	0.94 (∅, ∅, M)	0.70 (∅, ∅, MSc)	0.78 (si, ∅, MSc)
Top2	0.65 (si, ∅, MSc)	0.84 (∅, ∅, M)	0.68 (∅, ∅, MSc)	0.80 (si, ∅, MSc)
FD	0.75(MSc)	0.91 (Avg)	0.73 (A)	0.84 (A)

of 12% over the baseline in terms of accuracy while *NR* gets an improvement of 13%. In terms of Pearson's correlations, the best results are obtained when applying *Top1* to *MR* and *FD* to *NR*. Concerning the filters, we observe that we get different configurations than in overall polarity. Indeed, in terms of accuracy, the best results in *MR* are given by the *K_SE* followed by the topic filter *K_all* or the configuration *K_all* for subjectivity and *K_P* for topic for all the strategies (except for the *Top2*). Similar observations hold for *NR*, where we have in addition the subjectivity filter *K_SI*. Unlike for polarity overall ratings, the weight of implicit opinions seems to be less important for *MR* and more important for *NR*. On the other hand, taking into account *partof* topics has a stronger effect on multi-scale ratings, especially for *MR*. This might be because opinions focused on *partof* topics are more often used to express intensity nuances.

The discourse-based strategies (PD and FD) fail to capture the overall score in four main cases. The first one, concerns situations where the writer expresses implicit opinions towards other topics or when he is in a position of observer or recessed relative to the discussion. Second, sometimes, opinions in a document do not reflect the writer's point of view but the feelings of other persons. Hence, identifying the holder can yield an improvement. Third, ironic and sarcasm documents, where most subjective segments in a document are implicit. Finally, other cases of errors come from documents that are neither positive nor negative towards the main or a *partof* topic (about 4% of *MR*).

6 Conclusion

In this paper, we proposed the first research effort that empirically validates the importance of discourse for sentiment analysis. Based on a manual annotation campaign conducted on two corpus genres (movie reviews and news reactions), we have first shown that discourse has a strong effect on both polarity and subjectivity analysis. Then, we have proposed three strategies to compute document overall rating, namely bag of segments, partial discourse and full discourse.

Our results show that discourse-based strategies lead to better scores in terms of accuracy and Pearson's correlation on both corpus genres. Our results are more salient for overall scale rating than for polarity rating. In addition, this added value is more important for newspaper reactions than for movie reviews. The next step is to validate our results on automatically parsed data. We attempt to do this by adapting [15]'s parser to opinion texts.

Acknowledgement. This work was supported by a DGA-RAPID project under grant number 0102906143.

References

1. Asher, N., Lascarides, A.: Logics of Conversation. Cambridge University Press (2003)
2. Baldridge, J., Lascarides, A.: Annotating discourse structures for robust semantic interpretation. In: IWCS (2005)
3. Asher, N., Benamara, F., Mathieu, Y.Y.: Distilling opinion in discourse: A preliminary study. In: CoLing, pp. 7–10 (2008)
4. Taboada, M., Voll, K., Brooke, J.: Extracting sentiment as a function of discourse structure and topicality. School of Computing Science Technical Report 2008-20 (2008)
5. Heerschop, B., Goossen, F., Hogenboom, A., Frasincar, F., Kaymak, U., de Jong, F.: Polarity analysis of texts using discourse structure. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, pp. 1061–1070 (2011)
6. Somasundaran, S.: Discourse-level relations for Opinion Analysis. PhD Thesis, University of Pittsburgh (2010)
7. Zhou, L., Li, B., Gao, W., Wei, Z., Wong, K.F.: Unsupervised discovery of discourse relations for eliminating intra-sentence polarity ambiguities. In: Proceedings of EMNLP, pp. 162–171 (2011)
8. Benamara, F., Chardon, B., Mathieu, Y., Popescu, V.: Towards context-based subjectivity analysis. In: Proceedings of the IJCNLP, pp. 1180–1188 (2011)
9. Polanyi, L., van den Berg, M.: Discourse structure and sentiment. In: Data Mining Workshops (ICDMW), pp. 97–102 (2011)
10. Trnavac, R., Taboada, M.: The contribution of nonveridical rhetorical relations to evaluation in discourse. Language Sciences 34 (3), 301–318 (2012)
11. Carlson, L., Marcu, D., Okurowski, M.E.: Building a discourse-tagged corpus in the framework of rhetorical structure theory. In: van Kuppevelt, J., Smith, R. (eds.) Current Directions in Discourse and Dialogue, pp. 85–112. Kluwer Academic Publishers (2003)
12. Toprak, C., Jakob, N., Gurevych, I.: Sentence and expression level annotation of opinions in user-generated discourse. In: ACL, Morristown, NJ, USA, pp. 575–584 (2010)
13. Afantenos, S., et al.: An empirical resource for discovering cognitive principles of discourse organisation: the annodis corpus. In: LREC (2012)
14. Leung, C.W.K., Chan, S.C.F., Chung, F.L., Ngai, G.: A probabilistic rating inference framework for mining user preferences from reviews. World Wide Web 14(2), 187–215 (2011)
15. Muller, P., Afantenos, S., Denis, P., Asher, N.: Constrained decoding for text-level discourse parsing. In: Proceedings of COLING (2012)