Automatically Creating General-Purpose Opinion Summaries from Text

Veselin Stoyanov
Johns Hopkins University
ves@cs.jhu.edu

Claire Cardie
Cornell University
cardie@cs.cornell.edu

Abstract

We present and evaluate the first method known to us that can create rich non-extract-based opinion summaries from general text (e.g. newspaper articles). We first describe two possible representations for opinion summaries and then present our system OASIS, which identifies, and optionally aggregates, fine-grained opinions from the same source on the same topic. We propose new evaluation measures for both types of opinion summary and employ the metrics in an evaluation of OASIS on a standard opinion corpus. Our results are encouraging — OASIS substantially outperforms a competitive baseline when creating document-level *aggregate summaries* that compute the average polarity value across the multiple opinions identified for each source about each topic. We further show that as state-of-the-art performance on fine-grained opinion extraction improves, we can expect to see opinion summaries of very high quality — with F-scores of 54-78% using our OSEM evaluation measure.

1 Introduction

To date, most of the research in opinion analysis (see Related Work section) has focused on the problem of extracting opinions — both at the document level (*coarse-grained opinion information*) and at the level of sentences, clauses, or individual expressions (*fine-grained opinion information*).

In contrast, our work concerns the consolidation of fine-grained information about opinions to create non-extract-based *opinion summaries*, a rich, concise and useful representation of the opinions expressed in a document. In particular, the opinion summaries produced by our system combine opinions from the same source and/or about the same topic and aggregate multiple opinions from the same source on the same topic in a meaningful way. A simple opinion summary is shown in Figure 1. In the sample text, there are seven opinions expressed — two negative and one positive opinion from the American public on the war in Iraq, two negative opinions of Bush on withdrawal from Iraq, and so on. These are aggregated in the graph-based summary. We expect that this type of opinion summary, based on fine-grained opinion information, will be important for information analysis applications in any domain where the analysis of opinions and other subjective language is critical. Our notion of summary is fundamentally different from the extract-based textual summaries used often in Natural Language Processing. We use the term *non-extract-based summary* to make that distinction explicit, but also use *opinion summary* to refer to the summaries that we propose.

In this paper, we present and evaluate OASIS (for Opinion Aggregation and SummarIzation System), the first system known to us that can produce rich non-extract-based opinion summaries from general text.[1] The system relies on automatically extracted fine-grained opinion information and constructs fully automatic opinion summaries in a form that can be easily presented to humans or queried by other NLP applications. In addition, we discuss for the first time different forms of opinion summaries and provide novel methods for quantitative evaluation of opinion summaries.

Unlike most extract-based summarization tasks, we are able to automatically generate gold standard summaries for evaluation. As a result, our

[1] Several systems for summarizing the opinions expressed in product reviews exist (e.g. Hu and Liu (2004), Popescu and Etzioni (2005)). Due to the limited domain, summarizing opinions in product reviews constitutes a substantially different text-understanding problem; it has proven to be easier than the task addressed here and is handled using a very different set of techniques.

Proceedings of Recent Advances in Natural Language Processing, pages 202–209,
Hissar, Bulgaria, 12-14 September 2011.

[Source American public] opinion has [– *turned increasingly against*] [Topic the Iraq war]. The fourth anniversary of the Iraq war this week was marked by anti-[Topic war] [– *protests*] during the weekend. There were [Source some people] out to [+ *support*] [Topic the war] as well, fewer in number but no less vocal.
...
[Source Bush] has repeatedly [– *opposed*] [Topic setting timelines for withdrawing U.S. troops from Iraq]. [Source He] reiterated [Source the administration]'s stance that [– *premature*] [Topic troop withdrawal from Iraq] would leave security to Iraqi forces that [– *cannot yet cope*] with it on their own and allow [Topic groups like al Qaeda] to establish a base from which to [– *attack*] the US.

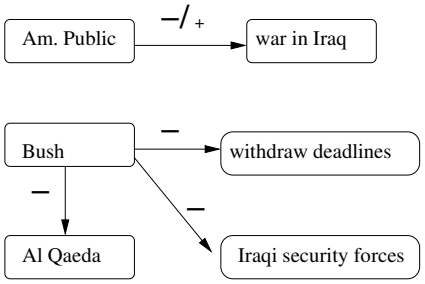

Figure 1: Example text containing opinions (above) and a summary of the opinions (below). In the text, sources and targets of opinions are bracketed; opinion expressions are shown in italics and bracketed with associated polarity, either positive (+) or negative (-). In the summary, entities involved in opinions are shown as nodes and aggregated opinions are shown as directed edges.

evaluation measures require no human intervention.

Our results are encouraging — OASIS substantially outperforms a competitive baseline when creating document-level *aggregate summaries* (like the one in Figure 1). We further show that as state-of-the-art performance on fine-grained opinion extraction improves, we can expect to see opinion summaries of very high quality (F-scores of 54-77% using our OSEM evaluation measure).

2 Opinion Summary Formats

In this section we discuss our notion of opinion summary as motivated by the needs of different applications and uses. In general, we presume the existence of automatically extracted fine-grained opinions, each of which has the following four attributes:

1. Trigger – the word or phrase that signals the expression of opinion in the text.
2. Source – the entity to which the opinion is to be attributed. More precisely, the span of text (usually a noun phrase or pronoun) that specifies the entity to which the opinion is to be attributed.
3. Topic – the topic of the opinion – either an entity (e.g. "Sue dislikes **John**") or a general topic (e.g. "I don't think that **lending money to friends** is a good idea").
4. Polarity – the sentiment (favorability) expressed in the opinion – either positive, negative, or neutral (a non-judgmental opinion that does not express a favorable or unfavorable attitude).

We expect that applications will use summaries of fine-grained opinion information in two distinct ways, giving rise to two distinct summary formats. The two formats differ in the way multiple opinions from the same source about the same topic are combined.

Aggregate opinion summary In an *aggregate opinion* summary, multiple opinions from a source on a topic are merged into a single aggregate opinion that represents the accumulated opinions of the source on that topic considering the document as a whole. Figure 1 depicts an aggregate opinion summary for the accompanying text.

Aggregate opinion summaries allow applications or users to access opinions in a standardized form. They will be needed by applications such as multi-perspective question answering (QA) (Stoyanov et al., 2005; Balahur et al., 2009), for example, which might need to answer questions such as "What is X's opinion toward Y?"

Opinion set summary In an *opinion set* summary, multiple opinions from a source on a topic are collected into a single set (without analyzing them for the overall trend). An opinion set summary of the example in Figure 1 would include, for example, three directed links from *American public* toward *war in Iraq* — one for each of the three expressions of opinion.

Opinion set summaries support fined-grained information extraction of opinions as well as user-directed exploration of the opinions in a document.

3 Related Work

Our works falls in the area of fine-grained subjectivity analysis concerned with analyzing opinions at, or below, the sentence level. Recent work, for example, indicates that systems can be trained to recognize opinions and their polarity, strength, and sources to a reasonable degree of accuracy (e.g. Dave et al. (2003), Riloff and Wiebe (2003), Bethard et al. (2004), Wilson et al. (2004), Yu and Hatzivassiloglou (2003), Choi et al. (2005), Kim and Hovy (2005), Wiebe and Riloff (2005)). Our work builds on research on fine-grained opinion extraction by extracting additional information that allows the creation of concise opinion summaries. In contrast to the opinion extracts produced by Pang and Lee (2004), our summaries are not text extracts, but rather explicitly identify and characterize the relations between opinions and their sources.

Several methods for computing opinions from product reviews exist (e.g. Hu and Liu (2004), Popescu and Etzioni (2005)). Due to properties of the limited domain and genre, however, the problem and approaches have been considerably simplified. In the product domain, summaries have are computed by extracting tuples [product attribute, opinion trigger, polarity] (with the product attribute extraction typically performed as a straightforward dictionary lookup) and computing summary statistics for each attribute.

The only other opinion summarization system in the general domain that we are aware of was perfromed as part of the 2008 text understanding conference (TAC) (Dang, 2008) Opinion Summarization task. The opinion summariztion task provides systems with a target such as "Trader Joe's" and 1 or 2 questions with answers of type SQUISHY LIST. A SQUISHY LIST contains complex concepts, which can overlap, may be expressed in different ways and where boundaries of the concepts are not well defined. In response, systems are expected to produce one fluent summary per target that summarizes the answers to all the questions for the target. Summaries are scored for their content using the Pyramid F-score (Nenkova et al., 2007) borrowed from the field of summarization. Additionally, summaries are manually scored along five dimensions: grammaticality, non-redundancy, structure/coherence, overall readability and overall responsiveness (content + readability).

Our work differs from the 2008 TAC Opinion tasks in several ways: We are always grouping together opinions that belong to the same **source**, while TAC 2008 tasks do not require that sources of opinions are identified. We are interested in grouping together opinions that are on the same **topic**, while the topics for the 2008 TAC Opinion tasks are pre-specified and involve a single named entity. TAC tasks do not always require **polarity** or aggregating polarities of individual opinions. We aim for an **abstract, graph-based representation** of opinions, while the TAC Opinion Summary task aims for extractive summaries.

4 Opinion Summarization System

In this section we describe the architecture of our system, OASIS.

Fine-grained Opinion Extraction OASIS starts with the output of Choi et al.'s (2006) extractor, which recognizes opinion sources and triggers. These predictions can be described as a tuple [opinion trigger, source] with each component representing a span of text in the original document. We enhance these fine-grained opinion predictions by using the opinion polarity classifier from Choi and Cardie (2009), which adds polarity predictions as one of three possible values: *positive*, *negative* or *neutral*. This value is added to the opinion tuple to obtain [opinion trigger, source, polarity] triples.

Source Coreference Resolution Given the fine-grained opinions, our system uses *source coreference resolution* to decide which opinions should be attributed to the same source. For this task, we rely on the partially supervised learning approach of Stoyanov and Cardie (2006). Following this step, OASIS produces opinion triples grouped according to their sources.

Topic Extraction/Coreference Resolution Next, our system labels fine-grained opinions with their topic and decide which opinions are on the same topic. Here, we use the *topic coreference resolution* approach proposed in Stoyanov and Cardie (2008). As a result of this step, OASIS produces opinion four-tuples [opinion trigger, source, polarity, topic name] that are grouped both

Component	Measure	Score
Fine-grained op. extractor	F1	59.7
Polarity classifier	Acc.	65.3
Source coreference resolver	B^3	83.2
Topic coreference resolver	B^3	54.7

Table 1: Performance of components of the opinion summarization system (Acc. refers to Accuracy).

according to their source and their topic. This four-tuple constitutes an opinion set summary.

Aggregating Multiple Opinions Finally, to create an aggregate opinion summary like that of Figure 1, OASIS needs to combine the multiple (possibly conflicting) opinions from a source on the same topic that appear in the opinion set summary. This is done in a straightforward way: the polarity of the aggregate opinion is computed as the average of the polarity of all the opinions from the source on the topic.

Performance of the different subcomponents of our system as it applies to our data (see Section 6) are shown in Table 1. F1 refers to the harmonic average of precision and recall, while the B^3 evaluation metric for coreference resolution (Bagga and Baldwin, 1998) is described in Section 5.[2]

5 Evaluation Metrics

Scientific approach to opinion summarization requires evaluation metrics to quantitatively compare summaries produced by different systems. We propose two new evaluation metrics for opinion summaries inspired by metrics used for coreference resolution and information extraction.

5.1 Doubly-linked B^3 score

Opinion set summaries are similar to the output of coreference resolution – both target grouping a set of items together. Thus, our first evaluation metric is based on a popular coreference resolution measure, the B^3 score (Bagga and Baldwin, 1998). B^3 evaluates the quality of a an automatically generated clustering of items (the system response) as compared to a gold-standard clustering of the same items (the key). It is computed as the recall for each item i: $Recall_i = |R_i \cap S_i|/|S_i|$, where R_i and S_i are the clusters that contains i in the response and the key, respectively. The recall for a document is the average over all items. Precision is computed by switching the roles of the key and the response and the reported score is the harmonic average of precision and recall (the F score).

Opinion summaries differ from coreference resolution in an important way: opinion sets are doubly linked – two opinions are in the same set when they have the same source **and** the same topic. We address this difference by introducing a modified version of the B^3 algorithm – the Doubly Linked B^3 (DLB^3) score . DLB^3 computes the recall for each item (opinion) i as an average of the recall with respect to the source ($recall_i^{src}$) and the recall with respect to the topic ($recall_i^{topic}$). More precisely:

$$\begin{aligned} DLB^3\, recall_i &= (recall_i^{src} + recall_i^{topic})/2 \\ recall_i^{src} &= |R_i^{src} \cap S_i^{src}|/|S_i^{src}| \end{aligned}$$

5.2 Opinion Summary Evaluation Metric

We propose a novel Opinion Summary Evaluation Metric (OSEM) that combines ideas from the ACE score (ACE, 2006) (used for information extraction) and Luo's (2005) CEAF score (used for coreference resolution). OSEM can be used for both opinion set and aggregate summaries.

The OSEM metric compares two opinion summaries – the key, K, and the response, R, containing a number of "summary opinions", each of which is comprised of one or more fine-grained opinions. Each summary opinion is characterized by three attributes (the source name, the polarity and the topic name) and by the set of fine-grained opinions that were joined to form the summary opinion. OSEM evaluates how well the key's summary opinions are extracted in the response by establishing a mapping $f : K \rightarrow R$ between the summary opinions in the key and the response. A value is associated with each mapping defined as: $value_f(K, R) = \sum_{A \in K} match(A, f(A))$, where $match(A, B)$ is a measure of how well opinions A and B match (discussed below). Similarly to the ACE and CEAF score, OSEM relies on the globally optimal matching $f^* = argmax_f(value_f(K, R))$ between the key and the response. OSEM takes CEAF's approach to compute precision

[2]Our scores for fine-grained opinion extraction differ from published results (Choi et al., 2006) because we do not allow the system to extract speech events that do not signal expressions of opinions (i.e. the word "said" when used in objective context: "John said his car is blue.").

Fine-grained opinions	System	DLB^3	OSEM				
			$\alpha = 0$	$\alpha = .25$	$\alpha = .5$	$\alpha = .75$	$\alpha = 1$
Automatic	Baseline	29.20	50.78	37.32	27.90	21.12	25.47
	OASIS	31.24	49.75	41.71	35.82	31.52	41.50
Manual	Baseline	51.12	78.67	60.72	47.04	36.60	28.59
	OASIS	59.82	78.69	69.04	61.47	55.59	54.80
	OASIS + manual src coref	79.85	82.65	79.39	76.68	74.61	74.95
	OASIS + manual tpc coref	80.80	82.40	78.14	74.53	71.56	71.03

Table 2: Scores for the summary system with varying levels of automatic information.

as $value_{f^*}(K,R)/value(R,R)$ and recall as $value_{f^*}(K,R)/value(K,K)$ and report OSEM score as the harmonic average (F-score) of precision and recall. The optimal matching is computed efficiently using the Kuhn-Munkres algorithm.

Finally, $match(A,B)$, the score for a match between summary opinions A and B is computed as a combination of how well the attributes of the summary opinion are matched and how well the individual opinion mentions (i.e. the fine-grained opinions in the text that form the aggregate opinion) are extracted. More precisely we define,

$$match(A,B) = attrMatch(A,B)^{\alpha} * mentOlp(A,B)^{(1-\alpha)},$$

where $attrMatch(A,B) \in [0,1]$ is computed as an average of how well each of the three attributes (source name, topic name and polarity) of the two summary opinions match. $mentOlp(A,B) = (2 * |A \cap B|)/(|A| + |B|)$ is a measure of how well fine-grained opinions that make up the summary opinion are extracted. Lastly, $\alpha \in [0,1]$ is a parameter that controls how much weight is given to identifying correctly the attributes of summary opinions vs. extracting all fine-grained opinions.

The α parameter allows us to tailor the OSEM score toward either type of opinion summary. For example, OSEM_0 (we will use OSEM_0 to refer to the OSEM score with $\alpha = 0$) reflects only how well the response groups together fine-grained opinions from the same source and on the same topic and makes no reference to the attributes of summary opinions. Thus, this value of α is suitable to evaluating opinion set summaries. On the other hand, OSEM_1 ($\alpha = 1$) puts all weight on how well the attributes of each summary opinion are extracted, which is suitable for evaluating aggregate opinion summaries. However, OSEM_1 does not require summary opinions to be connected to any fine-grained opinions in the text. This can lead to inconsistent summaries getting undeserved credit. For instance, in the example of Figure 1 a system could incorrectly infer that there is a neutral opinion from Bush toward the American public. OSEM_1 will give partial credit to such a summary opinion when compared to the negative opinion from Bush toward Al Qaeda, for example. At any other value ($\alpha < 1$) the $mentOlp$ for such an opinion will be 0 giving no partial credit for opinions that are not grounded to a fine-grained opinion in the text. The influence of the α parameter is studied empirically in the next section.

6 Experimental Evaluation

For evaluation we use the MPQA (Wiebe et al., 2005) and MPQA^{Topic} (Stoyanov and Cardie, 2008) corpora.[3] The MPQA corpus consists of 535 documents from the world press, manually annotated with phrase-level opinion information following the annotation scheme of Wiebe et al. (2005). The corpus provides annotations for opinion expressions, their polarities, and sources as well as source coreference. The MPQA^{Topic} corpus consists of 150 documents from the MPQA corpus, which are also manually annotated with opinion topic information, including topic spans, topic labels, and topic coreference.

Our gold-standard summaries are created automatically for each document in the MPQA^{Topic} corpus by relying on the manually annotated fine-grained opinion and source- and topic-coreference information. For our experiments, all components of OASIS are trained on the 407 documents in the MPQA corpus that are not part of the MPQA^{Topic} corpus, with the exception of topic coreference, which uses 5-fold cross-validation on the MPQA^{Topic} corpus.

[3]The MPQA corpus is available at http://nrrc.mitre.org/NRRC/publications.htm.

Taipei, Sept. 26 (CNA) – It is unlikely that the Vatican will establish diplomatic ties with mainland China any time soon, judging from their differences on religious issues, Ministry of Foreign Affairs (MOFA) spokeswoman [Source Chang Siao-yue] [neu *said*] Wednesday.
[Source Chang]'s [neu *remark*] came in response to a foreign wire [neu *report*] that mainland China and the Vatican are preparing to bridge their differences and may even pave the way for full diplomatic relations.
[Source Beijing authorities] are [neu *expected*] to take advantage of a large religious meeting slated for October 14 in Beijing to develop the possibility of setting up formal relations with the Vatican, [neu *according*] to the report.
...
[Source The MOFA spokeswoman] [+ *affirmed*] that from the angle of Eastern and Western cultural exchanges, the sponsoring of similar conferences will be instrumental to [Source mainland Chinese people]'s [+ *better understanding*] of Catholicism and its contributions to Chinese society.
As for the development of diplomatic relations between mainland China and the Vatican, [Source Chang] [− *noted*] that differences between the Beijing leadership and the Holy See on religious issues dates from long ago, so it is impossible for the Vatican to broach this issue with Beijing for the time being.
[Source Chang] also [+ *reaffirmed*] the solid and cordial diplomatic links between the Republic of China and the Vatican.

KEY SUMMARY:

♯ source	opinion	topic
k1. **Chang Siao-yue**	**neutral**	**diplomatic links**
	said	
	remark	
	noted	
	reaffirmed	
k2. **foreign wire**	**neutral**	**diplomatic links**
	report	
	according to	
k3. **Chinese people**	**positive**	**Catholicism**
	better understanding	
k4. **Chang**	**positive**	**conferences**
	affirmed	
k5. **author**	**neutral**	**Beijing authorities**
	are expected	

RESPONSE SUMMARY:

♯ **source**	opinion	topic
r1. **Chang Siao-yue**	**positive**	**pave bridge vatican**
	said	
	remark	
	noted	
	reaffirmed	
r2. **MOFA spokeswoman**	**positive**	**sponsor conference**
	affirmed	**Catholicism**
r3. **Chinese people**	**neutral**	**sponsor conference**
	better understanding	**Catholicism**
r4. **Beijing authorities**	**neutral**	**Beijing authorities**
	are expected	

Figure 2: An opinion summary produced by OASIS. The example shows the original article with gold-standard fine-grained opinion annotations above, the key opinion summary in the middle and the summary produced by OASIS below.

6.1 Example

We begin our evaluation section by introducing an example of an output summary produced by OASIS. The top part of Figure 2 contains the text of a document from the MPQA^{Topic} corpus, showing the fine-grained opinion annotations as they are marked in the MPQA corpus. The middle part of Figure 2 shows the gold-standard summary produced from the manual annotations. The summary is shown as a table with each box corresponding to an overall opinion. Each opinion box shows the source name on the left (each opinion is labeled with a unique string, e.g. $k1$ for the first opinion in the key) and the topic name on the right (string equivalence for the source and topic name indicate the same source/topic for the purpose of the example). The middle column of the opinion box shows the opinion characterized by the computed overall opinion shown in the first row and all opinion mentions that were combined to produce the overall opinion shown in subsequent rows (for the purpose of presentation mentions are shown as strings, but in reality they are represented as spans in the original text by the summaries). Finally, the summary produced by OASIS is shown in the bottom part of Figure 2 following the same format.

OASIS performed relatively well on the example summary of Figure 2. This is partially due to the fact that most of the opinion mentions were identified correctly. Additionally, source coreference and topic coreference appear to be mostly accurate, but there are several mistakes in labeling the topic clusters as compared to the gold standard.

Next, we use the example of Figure 2 to illustrate the computation of the OSEM score. The first step of computing the score is to calculate the scores for how well each response opinion matches each key opinion. The four-by-five matrix of scores for matching response opinions to key opinions is shown in Table 3. Scores in the table are computed for value of the α parameter set to .5. As discussed in the previous section, all values of $\alpha < 1$ require that key and response opinions have at least one mention in common to receive a non-zero score. This is illustrated in Table 3, where only four of the 20 match scores are greater than 0.

Based on the scores in Table 3, the optimal match between key and response opinions is $r1 \rightarrow k1$, $r2 \rightarrow k4$, $r3 \rightarrow k3$, and $r4 \rightarrow k5$. The value of this score is 2.91, which translates in OSEM$_{.5}$

α	0.00	0.01	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.99	1.00
OSEM prec	51.5	50.9	47.8	44.6	41.8	39.3	37.1	35.2	33.5	32.0	30.7	29.6	42.8
OSEM recall	48.1	47.6	44.7	41.7	39.0	36.7	34.6	32.8	31.2	29.7	28.5	27.5	40.3
OSEM F1	49.8	49.2	46.2	43.1	40.4	38.0	35.8	33.9	32.3	30.8	29.5	28.5	41.5

Table 4: OSEM precision, recall and F-score as a function of α.

	k1	k2	k3	k4	k5
r1	.58	0	0	0	0
r2	0	0	0	.81	0
r3	0	0	.71	0	0
r4	0	0	0	0	.81

	k1	k2	k3	k4	k5
r1	.33	0	.33	.67	0
r2	0	0	.33	.50	0
r3	.33	.33	.50	.16	.33
r4	.33	.33	0	0	.67

Table 3: OSEM score for each response opinion as matched to key opinions in the example summary of Figure 2 with parameter $\alpha = .5$ (above) and $\alpha = 1.0$ (below).

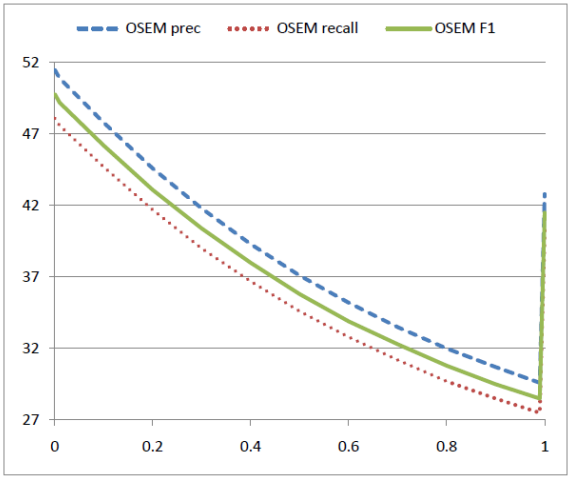

Figure 3: OSEM precision, recall and F-score (x-axis) vs. α (y-axis).

precision of .73 and recall of .58 for an overall $\text{OSEM}_{.5}$ F-score of .65.

Finally, to illustrate the different implications for the score when the α parameter is set to 1, we show the match scores for OSEM_1 in Table 3. Note that there are far fewer 0 scores in Table 3 as compared to Table 3. In the case of this particular summary, the optimal matching between key and response opinions is the same for as the setting of $\alpha = .5$, but this is not always the case. The OSEM_1 precision, recall and F-score for this summary are .50, .60 and .55, respectively.

6.2 Baseline

We compare the performance of our system to a baseline that creates one summary opinion for each fine-grained opinion. In other words, each source and topic mention is considered unique and each opinion is in its own cluster.

6.3 Results

Results are shown in Table 2. We compute $\text{DL}B^3$ score and OSEM score for 5 values of α chosen uniformly over the $[0, 1]$ interval. The top two rows of Table 2 contain results for using fully automatically extracted information.

Compared to the baseline, OASIS shows little improvement when considering opinion set summaries ($\text{DL}B^3$ improves from 29.20 to 31.20, while OSEM_0 worsens from 50.78 to 49.75). However, as α grows and more emphasis is put on correctly identifying attributes of summary opinions, OASIS substantially outperforms the baseline (OSEM_1 improves from 25.47 to 41.50).

Next, we try to tease apart the influence of different subsystems. The bottom four rows of Table 2 contain system runs using gold-standard information about fine-grained opinions (i.e. the [opinion trigger, source, polarity] triple). Results indicate that the quality of fine-grained opinion extractions has significant effect on overall system performance – scores for both the baseline and OASIS improve substantially. Additionally, OASIS appears to improve more compared to the baseline when using manual fine-grained opinion information. The last two rows of Table 2 show the performance of OASIS when using manual information for source and topic coreference, respectively. Results indicate that the rest of the errors of OASIS can be attributed roughly equally to the source and topic coreference modules.

Lastly, the OSEM score is higher at the two extreme values for α (0 and 1) as compared to values

in the middle (such as .5). To study this anomaly, we compute OSEM scores for 13 values of α. Results, shown in Table 4, and visualized in Figure 3, indicate that the OSEM score decreases as more weight is put on identifying attributes of summary opinions (i.e. α increases) with a discontinuity at $\alpha = 1$. We attribute this discontinuity to the fact that OSEM_1 does not require opinions to be grounded in text as discussed in Section 5.2. Note, however, that the $\alpha = 1$ setting is akin to the standard evaluation scenario for many information extraction tasks.

7 Conclusions

We present and evaluate OASIS, the first general-purpose non-extract-based opinion summarization system known to us. We discuss possible forms of opinion summaries motivated by application needs, describe the architecture of our system and introduce new evaluation measures for objectively judging the goodness of complete opinion summaries. Results are promising – OASIS outperforms a competitive baseline by a large margin when we put more emphasis on computing an aggregate summary.

References

ACE. 2006. Ace 2005 evaluation, November.

A. Bagga and B. Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *In Proceedings of COLING/ACL*.

A. Balahur, E. Boldrini, A. Montoyo, and P. Martínez-Barco. 2009. Opinion and generic question answering systems: a performance analysis. In *Proceedings of the ACL-IJCNLP*.

S. Bethard, H. Yu, A. Thornton, V. Hativassiloglou, and D. Jurafsky. 2004. Automatic extraction of opinion propositions and their holders. In *2004 AAAI Spring Symposium on Exploring Attitude and Affect in Text*.

Y. Choi and C. Cardie. 2009. Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification. In *Proceedings of EMNLP*.

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

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

H.T. Dang. 2008. Overview of the TAC 2008 Opinion Question Answering and Summarization Tasks. In *Text Analysis Conference (TAC-2008)*, Gaithersburg, Maryland (USA), November.

K. Dave, S. Lawrence, and D. Pennock. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of IWWWC*.

M. Hu and B. Liu. 2004. Mining opinion features in customer reviews. In *AAAI*, pages 755–760.

S. Kim and E. Hovy. 2005. Identifying opinion holders for question answering in opinion texts. In *Proceedings of AAAI Workshop on Question Answering in Restricted Domains*.

X. Luo. 2005. On coreference resolution performance metrics. In *Proceedings of EMNLP*.

A. Nenkova, R. Passonneau, and K. McKeown. 2007. The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing*, 4(2).

B. Pang and L. Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of ACL*.

A.-M. Popescu and O. Etzioni. 2005. Extracting product features and opinions from reviews. In *HLT/EMNLP*.

E. Riloff and J. Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proceesings of EMNLP*.

V. Stoyanov and C. Cardie. 2006. Partially supervised coreference resolution for opinion summarization through structured rule learning. In *Proceedings of EMNLP*.

V. Stoyanov and C. Cardie. 2008. Topic identification for fine-grained opinion analysis. In *Proceedings of COLING*.

V. Stoyanov, C. Cardie, and J. Wiebe. 2005. Multi-Perspective question answering using the OpQA corpus. In *Proceedings of EMNLP*.

J. Wiebe and E. Riloff. 2005. Creating subjective and objective sentence classifiers from unannotated texts. In *Proceedings of CICLing*.

J. Wiebe, T. Wilson, and C. Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 1(2).

T. Wilson, J. Wiebe, and R. Hwa. 2004. Just how mad are you? Finding strong and weak opinion clauses. In *Proceedings of AAAI*.

H. Yu and V. Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of EMNLP*.