Multi-Perspective Question Answering Using the OpQA Corpus

Veselin Stoyanov and **Claire Cardie**
Department of Computer Science
Cornell University
Ithaca, NY 14850, USA
{ves,cardie}@cs.cornell.edu

Janyce Wiebe
Department of Computer Science
University of Pittsburgh
Pittsburgh, PA 15260, USA
wiebe@cs.pitt.edu

Abstract

We investigate techniques to support the answering of opinion-based questions. We first present the OpQA corpus of opinion questions and answers. Using the corpus, we compare and contrast the properties of fact and opinion questions and answers. Based on the disparate characteristics of opinion vs. fact answers, we argue that traditional fact-based QA approaches may have difficulty in an MPQA setting without modification. As an initial step towards the development of MPQA systems, we investigate the use of machine learning and rule-based subjectivity and opinion source filters and show that they can be used to guide MPQA systems.

1 Introduction

Much progress has been made in recent years in automatic, open-domain question answering (e.g., Voorhees (2001), Voorhees (2002), Voorhees and Buckland (2003)). The bulk of the research in this area, however, addresses fact-based questions like: "When did McDonald's open its first restaurant?" or "What is the Kyoto Protocol?". To date, however, relatively little research been done in the area of Multi-Perspective Question Answering (MPQA), which targets questions of the following sort:

- How is Bush's decision not to ratify the Kyoto Protocol looked upon by Japan and other US allies?
- How do the Chinese regard the human rights record of the United States?

In comparison to fact-based question answering (QA), researchers understand far less about the properties of questions and answers in MPQA, and have yet to develop techniques to exploit knowledge of those properties. As a result, it is unclear whether approaches that have been successful in the domain of fact-based QA will work well for MPQA.

We first present the *OpQA* corpus of opinion questions and answers. Using the corpus, we compare and contrast the properties of fact and opinion questions and answers. We find that text spans identified as answers to opinion questions: (1) are approximately twice as long as those of fact questions, (2) are much more likely (37% vs. 9%) to represent *partial* answers rather than complete answers, (3) vary much more widely with respect to syntactic category – covering clauses, verb phrases, prepositional phrases, and noun phrases; in contrast, fact answers are overwhelming associated with noun phrases, and (4) are roughly half as likely to correspond to a single syntactic constituent type (16-38% vs. 31-53%).

Based on the disparate characteristics of opinion vs. fact answers, we argue that traditional fact-based QA approaches may have difficulty in an MPQA setting without modification. As one such modification, we propose that MPQA systems should rely on natural language processing methods to identify information about opinions. In experiments in opinion question answering using the OpQA corpus, we find that filtering potential answers using machine learning and rule-based NLP opinion filters substantially improves the performance of an end-to-end MPQA system according to both a mean reciprocal rank (MRR) measure (0.59 vs. a baseline of 0.42)

and a metric that determines the mean rank of the first correct answer (MRFA) (26.2 vs. a baseline of 61.3). Further, we find that requiring opinion answers to match the requested opinion source (e.g., does <source> approve of the Kyoto Protocol) dramatically improves the performance of the MPQA system on the hardest questions in the corpus.

The remainder of the paper is organized as follows. In the next section we summarize related work. Section 3 describes the OpQA corpus. Section 4 uses the OpQA corpus to identify potentially problematic issues for handling opinion vs. fact questions. Section 5 briefly describes an opinion annotation scheme used in the experiments. Sections 6 and 7 explore the use of opinion information in the design of MPQA systems.

2 Related Work

There is a growing interest in methods for the automatic identification and extraction of opinions, emotions, and sentiments in text. Much of the relevant research explores sentiment classification, a text categorization task in which the goal is to assign to a document either positive ("thumbs up") or negative ("thumbs down") polarity (e.g. Das and Chen (2001), Pang et al. (2002), Turney (2002), Dave et al. (2003), Pang and Lee (2004)). Other research has concentrated on analyzing opinions at, or below, the sentence level. Recent work, for example, indicates that systems can be trained to recognize opinions, their polarity, their source, and their strength to a reasonable degree of accuracy (e.g. Dave et al. (2003), Riloff and Wiebe (2003), Bethard et al. (2004), Pang and Lee (2004), Wilson et al. (2004), Yu and Hatzivassiloglou (2003), Wiebe and Riloff (2005)).

Related work in the area of corpus development includes Wiebe et al.'s (2005) opinion annotation scheme to identify *subjective expressions* — expressions used to express opinions, emotions, sentiments and other *private states* in text. Wiebe et al. have applied the annotation scheme to create the MPQA corpus consisting of 535 documents manually annotated for phrase-level expressions of opinion. In addition, the NIST-sponsored TREC evaluation has begun to develop data focusing on opinions — the 2003 Novelty Track features a task that requires systems to identify opinion-oriented documents w.r.t. a specific issue (Voorhees and Buckland, 2003).

While all of the above work begins to bridge the gap between text categorization and question answering, none of the approaches have been employed or evaluated in the context of MPQA.

3 OpQA Corpus

To support our research in MPQA, we created the OpQA corpus of opinion and fact questions and answers. Additional details on the construction of the corpus as well as results of an interannotator agreement study can be found in Stoyanov et al. (2004).

3.1 Documents and Questions

The OpQA corpus consists of 98 documents that appeared in the world press between June 2001 and May 2002. All documents were taken from the aforementioned MPQA corpus (Wilson and Wiebe, 2003)[1] and are manually annotated with phrase-level opinion information, following the annotation scheme of Wiebe et al. (2005), which is briefly summarized in Section 5. The documents cover four general (and controversial) topics: President Bush's alternative to the Kyoto protocol (*kyoto*); the US annual human rights report (*humanrights*); the 2002 coup d'etat in Venezuela (*venezuela*); and the 2002 elections in Zimbabwe and Mugabe's reelection (*mugabe*). Each topic is covered by between 19 and 33 documents that were identified automatically via IR methods.

Both fact and opinion questions for each topic were added to the OpQA corpus by a volunteer not associated with the current project. The volunteer was provided with a set of instructions for creating questions together with two documents on each topic selected at random. He created between six and eight questions on each topic, evenly split between fact and opinion. The 30 questions are given in Table 1 sorted by topic.

3.2 Answer annotations

Answer annotations were added to the corpus by two annotators according to a set of annotation instruc-

[1]The MPQA corpus is available at `http://nrrc.mitre.org/NRRC/publications.htm`. The OpQA corpus is available upon request.

	Kyoto
1 f	What is the Kyoto Protocol about?
2 f	When was the Kyoto Protocol adopted?
3 f	Who is the president of the Kiko Network?
4 f	What is the Kiko Network?
5 o	Does the president of the Kiko Network approve of the US action concerning the Kyoto Protocol?
6 o	Are the Japanese unanimous in their opinion of Bush's position on the Kyoto Protocol?
7 o	How is Bush's decision not to ratify the Kyoto Protocol looked upon by Japan and other US allies?
8 o	How do European Union countries feel about the US opposition to the Kyoto protocol?
	Human Rights
1 f	What is the murder rate in the United States?
2 f	What country issues an annual report on human rights in the United States?
3 o	How do the Chinese regard the human rights record of the United States?
4 f	Who is Andrew Welsdan?
5 o	What factors influence the way in which the US regards the human rights records of other nations?
6 o	Is the US Annual Human Rights Report received with universal approval around the world?
	Venezuela
1 f	When did Hugo Chavez become President?
2 f	Did any prominent Americans plan to visit Venezuela immediately following the 2002 coup?
3 o	Did anything surprising happen when Hugo Chavez regained power in Venezuela after he was removed by a coup?
4 o	Did most Venezuelans support the 2002 coup?
5 f	Which governmental institutions in Venezuela were dissolved by the leaders of the 2002 coup?
6 o	How did ordinary Venezuelans feel about the 2002 coup and subsequent events?
7 o	Did America support the Venezuelan foreign policy followed by Chavez?
8 f	Who is Vice-President of Venezuela?
	Mugabe
1 o	What was the American and British reaction to the reelection of Mugabe?
2 f	Where did Mugabe vote in the 2002 presidential election?
3 f	At which primary school had Mugabe been expected to vote in the 2002 presidential election?
4 f	How long has Mugabe headed his country?
5 f	Who was expecting Mugabe at Mhofu School for the 2002 election?
6 o	What is the basis for the European Union and US critical attitude and adversarial action toward Mugabe?
7 o	What did South Africa want Mugabe to do after the 2002 election?
8 o	What is Mugabe's opinion about the West's attitude and actions towards the 2002 Zimbabwe election?

Table 1: Questions in the OpQA collection by topic. *f* in column 1 indicates a fact question; *o*, an opinion question.

tions.[2] Every text segment that *contributes* to an answer to any of the 30 questions is annotated as an answer. In particular, answer annotations include segments that constitute a *partial answer*. Partial answers either (1) lack the specificity needed to constitute a full answer (e.g., "before May 2004" partially answers the question *When was the Kyoto protocol ratified?* when a specific date is known) or (2) need to be combined with at least one additional answer segment to fully answer the question (e.g., the question *Are the Japanese unanimous in their opposition of Bush's position on the Kyoto protocol?* is answered only partially by a segment expressing a single opinion). In addition, annotators mark the minimum answer spans (e.g., "a Tokyo organization," vs. "a Tokyo organization representing about 150 Japanese groups").

4 Characteristics of opinion answers

Next, we use the OpQA corpus to analyze and compare the characteristics of fact vs. opinion questions. Based on our findings, we believe that QA systems based solely on traditional QA techniques are likely to be less effective at MPQA than they are at traditional fact-based QA.

4.1 Traditional QA architectures

Despite the wide variety of approaches implied by modern QA systems, almost all systems rely on the following two steps (subsystems), which have empirically proven to be effective:

- **IR module.** The QA system invokes an IR subsystem that employs traditional text similarity measures (e.g., tf/idf) to retrieve and rank document fragments (sentences or paragraphs) w.r.t. the question (query).

- **Linguistic filters.** QA systems employ a set of filters and text processing components to discard some document fragments. The following filters have empirically proven to be effective and are used universally:

  Semantic filters prefer an answer segment that matches the semantic class(es) associated with the question type (e.g., *date* or *time* for *when* questions; *person* or *organization* for *who* questions).

  Syntactic filters are also configured on the type of question. The most common and effective syntactic filters select a specific constituent (e.g., noun phrase) according to the question type (e.g., *who* question).

QA systems typically interleave the above two subsystems with a variety of different processing steps of both the question and the answer. The goal of the processing is to identify text fragments that contain an answer to the question. Typical QA systems do not perform any further text processing; they return the text fragment as it occurred in the text. [3]

4.2 Corpus-based analysis of opinion answers

We hypothesize that QA systems that conform to this traditional architecture will have difficulty handling opinion questions without non-trivial modification. In support of this hypothesis, we provide statistics from the OpQA corpus to illustrate some of the characteristics that distinguish answers to opinion vs. fact questions, and discuss their implications for a traditional QA system architecture.

Answer length. We see in Table 2 that the average length of opinion answers in the OpQA corpus

[2] The annotation instructions are available at `http://www.cs.cornell.edu/ ves/ Publications/publications.htm`.

[3] This architecture is seen mainly in QA systems designed for TREC's "factoid" and "list" QA tracks. Systems competing in the relatively new "definition" or "other" tracks have begun to introduce new approaches. However, most such systems still rely on the IR step and return the text fragment as it occurred in the text.

	Number of answers	Length	Number of partials
fact	124	5.12	12 (9.68%)
opinion	415	9.24	154 (37.11%)

Table 2: Number of answers, average answer length (in tokens), and number of partial answers for fact/opinion questions.

is 9.24 tokens, almost double that of fact answers. Unfortunately, longer answers could present problems for some traditional QA systems. In particular, some of the more sophisticated algorithms that perform **additional processing** steps such as logical verifiers (Moldovan et al., 2002) may be less accurate or computationally infeasible for longer answers. More importantly, longer answers are likely to span more than a single syntactic constituent, rendering the syntactic filters, and very likely the semantic filters, less effective.

Partial answers. Table 2 also shows that over 37% of the opinion answers were marked as partial vs. 9.68% of the fact answers. The implications of partial answers for the traditional QA architecture are substantial: an MPQA system will require an **answer generator** to (1) distinguish between partial and full answers; (2) recognize redundant partial answers; (3) identify which subset of the partial answers, if any, constitutes a full answer; (4) determine whether additional documents need to be examined to find a complete answer; and (5) asemble the final answer from partial pieces of information.

Syntactic constituent of the answer. As discussed in Section 4.1, traditional QA systems rely heavily on the predicted syntactic and semantic class of the answer. Based on answer lengths, we speculated that opinion answers are unlikely to span a single constituent and/or semantic class. This speculation is confirmed by examining the phrase type associated with OpQA answers using Abney's (1996) CASS partial parser.[4] For each question, we count the number of times an answer segment for the question (in the manual annotations) matches each constituent type. We consider four constituent types – noun phrase (n), verb phrase (v), prepositional phrase (p), and clause (c) – and three matching criteria:

[4]The parser is available from `http://www.vinartus.net/spa/`.

Fact						Opinion					
Question	# of answers	Matching Criteria ex	up	up/dn	syn type	Question	# of answers	Matching Criteria ex	up	up/dn	syn type
H 1	1	0	0	0		H 3	15	5	5	5	c
H 2	4	2	2	2	n	H 5	24	5	5	10	n
H 4	1	0	0	0		H 6	123	17	23	52	n
K 1	48	13	14	24	n	K 5	3	0	0	1	
K 2	38	13	13	19	n	K 6	34	6	5	12	c
K 3	1	1	1	1	c n	K 7	55	9	8	19	c
K 4	2	1	1	1	n	K 8	25	4	4	10	v
M 2	3	0	0	1		M 1	74	10	12	29	v
M 3	1	0	0	1		M 6	12	3	5	7	n
M 4	10	2	2	5	n	M 7	1	0	0	0	
M 5	3	1	1	2	c	M 8	3	0	0	1	
V 1	4	3	3	4	n	V 3	1	1	0	1	c
V 2	1	1	1	1	n	V 4	13	2	2	2	c
V 5	3	0	1	1		V 6	9	2	2	5	c n
V 8	4	2	4	4	n	V 7	23	3	1	5	
Coverage	124	39 31%	43 35%	66 53%		Coverage	415	67 16%	70 17%	159 38%	

Table 3: Syntactic Constituent Type for Answers in the OpQA Corpus

1. The **ex**act match criterion is satisfied only by answer segments whose spans exactly correspond to a constituent in the CASS output.

2. The **up** criterion considers an answer to match a CASS constituent if the constituent completely contains the answer and no more than three additional (non-answer) tokens.

3. The **up/dn** criterion considers an answer to match a CASS constituent if it matches according to the **up** criterion or if the answer completely contains the constituent and no more than three additional tokens.

The counts for the analysis of answer segment syntactic type for fact vs. opinion questions are summarized in Table 3. Results for the 15 fact questions are shown in the left half of the table, and for the 15 opinion questions in the right half. The leftmost column in each half provides the question topic and number, and the second column indicates the total number of answer segments annotated for the question. The next three columns show, for each of the **ex**, **up**, and **up/dn** matching criteria, respectively, the number of annotated answer segments that match the majority syntactic type among answer segments for that question/criterion pair. Using a traditional QA architecture, the MPQA system might filter answers based on this majority type. The *syn type* column indicates the majority syntactic type using the exact match criterion; two values in the column indicate a tie for majority syntactic type, and an empty syntactic type indicates that no answer exactly matched any of the four constituent types. With only a few exceptions, the **up** and **up/dn** matching criteria agreed in majority syntactic type.

Results in Table 3 show a significant disparity between fact and opinion questions. For fact ques-

tions, the syntactic type filter would keep 31%, 35%, or 53% of the correct answers, depending on the matching criterion. For opinion questions, there is unfortunately a two-fold reduction in the percentage of correct answers that would remain after filtering — only 16%, 17% or 38%, depending on the matching criterion. More importantly, the majority syntactic type among answers for fact questions is almost always a noun phrase, while no single constituent type emerges as a useful syntactic filter for opinion questions (see the **syn phrase** columns in Table 3). Finally, because semantic class information is generally tied to a particular syntactic category, the effectiveness of traditional semantic filters in the MPQA setting is unclear.

In summary, identifying answers to questions in an MPQA setting within a traditional QA architecture will be difficult. First, the implicit and explicit assumptions inherent in standard linguistic filters are consistent with the characteristics of fact- rather than opinion-oriented QA. In addition, the presence of relatively long answers and partial answers will require a much more complex **answer generator** than is typically present in current QA systems.

In Sections 6 and 7, we propose initial steps towards modifying the traditional QA architecture for use in MPQA. In particular, we propose and evaluate two types of **opinion filters** for MPQA: **subjectivity filters** and **opinion source filters**. Both types of linguistic filters rely on phrase-level and sentence-level opinion information, which has been manually annotated for our corpus; the next section briefly describes the opinion annotation scheme.

5 Manual Opinion Annotations

Documents in our OpQA corpus come from the larger MPQA corpus, which contains manual opinion annotations. The annotation framework is described in detail in (Wiebe et al., 2005). Here we give a high-level overview.

The annotation framework provides a basis for *subjective expressions*: expressions used to express opinions, emotions, and sentiments. The framework allows for the annotation of both directly expressed private states (e.g., *afraid* in the sentence "John is afraid that Sue might fall,") and opinions expressed by the choice of words and style of language (e.g., *it is about time* and *oppression* in the sentence "It is about time that we end Saddam's oppression"). In addition, the annotations include several attributes, including the *intensity* (with possible values *low*, *medium*, *high*, and *extreme*) and the *source* of the private state. The *source* of a private state is the person or entity who holds or experiences it.

6 Subjectivity Filters for MPQA Systems

This section describes three **subjectivity filters** based on the above opinion annotation scheme. Below (in Section 6.3), the filters are used to remove fact sentences from consideration when answering opinion questions, and the OpQA corpus is used to evaluate their effectiveness.

6.1 Manual Subjectivity Filter

Much previous research on automatic extraction of opinion information performed classifications at the sentence level. Therefore, we define sentence-level opinion classifications in terms of the phrase-level annotations. For our gold standard of manual opinion classifications (dubbed MANUAL for the rest of the paper) we will follow Riloff and Wiebe's (2003) convention (also used by Wiebe and Riloff (2005)) and consider a sentence to be *opinion* if it contains at least one opinion of intensity *medium* or higher, and to be *fact* otherwise.

6.2 Two Automatic Subjectivity Filters

As discussed in section 2, several research efforts have attempted to perform automatic opinion classification on the clause and sentence level. We investigate whether such information can be useful for MPQA by using the automatic sentence level opinion classifiers of Riloff and Wiebe (2003) and Wiebe and Riloff (2005).

Riloff and Wiebe (2003) use a bootstrapping algorithm to perform a sentence-based opinion classification on the MPQA corpus. They use a set of high precision subjectivity and objectivity clues to identify subjective and objective sentences. This data is then used in an algorithm similar to AutoSlog-TS (Riloff, 1996) to automatically identify a set of extraction patterns. The acquired patterns are then used iteratively to identify a larger set of subjective and objective sentences. In our experiments we use

		precision	recall	F
MPQA corpus	RULEBASED	90.4	34.2	46.6
	NAIVE BAYES	79.4	70.6	74.7

Table 4: Precision, recall, and F-measure for the two classifiers.

the classifier that was created by the reimplementation of this bootstrapping process in Wiebe and Riloff (2005). We will use RULEBASED to denote the opinion information output by this classifier.

In addition, Wiebe and Riloff used the RULEBASED classifier to produce a labeled data set for training. They trained a Naive Bayes subjectivity classifier on the labeled set. We will use NAIVE BAYES to refer to Wiebe and Riloff's naive Bayes classifier.[5] Table 4 shows the performance of the two classifiers on the MPQA corpus as reported by Wiebe and Riloff.

6.3 Experiments

We performed two types of experiments using the subjectivity filters.

6.3.1 Answer rank experiments

Our hypothesis motivating the first type of experiment is that subjectivity filters can improve the answer identification phase of an MPQA system. We implement the IR subsystem of a traditional QA system, and apply the subjectivity filters to the IR results. Specifically, for each opinion question in the corpus [6] , we do the following:

1. Split all documents in our corpus into sentences.
2. Run an information retrieval algorithm[7] on the set of all sentences using the question as the query to obtain a *ranked list* of sentences.
3. Apply a subjectivity filter to the *ranked list* to remove all fact sentences from the *ranked list*.

We test each of the MANUAL, RULEBASED, and NAIVE BAYES subjectivity filters. We compare the rank of the first answer to each question in the *ranked list* before the filter is applied, with the rank of the first answer to the question in the *ranked list* after the filter is applied.

[5]Specifically, the one they label *Naive Bayes 1*.

[6]We do not evaluate the opinion filters on the 15 fact questions. Since opinion sentences are defined as containing at least one opinion of intensity medium or higher, opinion sentences can contain factual information and sentence-level opinion filters are not likely to be effective for fact-based QA.

[7]We use the Lemur toolkit's standard tf.idf implementation available from `http://www.lemurproject.org/`.

Topic	Qnum	Baseline	Manual	NaiveBayes	Rulebased
Kyoto	5	1	1	1	1
	6	5	4	4	3
	7	1	1	1	1
	8	1	1	1	1
Human Rights	3	1	1	1	1
	5	10	6	7	5
	6	1	1	1	1
Venezuela	3	106	81	92	35
	4	3	2	3	1
	6	1	1	1	1
	7	3	3	3	2
Mugabe	1	2	2	2	2
	6	7	5	5	4
	7	447	291	317	153
	8	331	205	217	182
MRR :		0.4911	0.5189	0.5078	0.5856
MRFA:		61.3333	40.3333	43.7333	26.2

Table 5: Results for the subjectivity filters.

Results. Results for the opinion filters are compared to a simple baseline, which performs the information retrieval step with no filtering. Table 5 gives the results on the 15 opinion questions for the baseline and each of the three *subjectivity filters*. The table shows two cumulative measures – the mean reciprocal rank (MRR) across the top five answers in the *ranked list* [8] and the mean rank of the first answer (MRFA). [9]

Table 5 shows that all three *subjectivity filters* outperform the baseline: for all three filters, the first answer in the filtered results for all 15 questions is ranked at least as high as in the baseline. As a result, the three subjectivity filters outperform the baseline in both MRR and MRFA. Surprisingly, the best performing subjectivity filter is RULEBASED, surpassing the gold standard MANUAL, both in MRR (0.59 vs. 0.52) and MRFA (40.3 vs. 26.2). Presumably, the improvement in performance comes from the fact that RULEBASED identifies subjective sentences with the highest precision (and lowest recall). Thus, the RULEBASED subjectivity filter discards non-subjective sentences most aggressively.

6.3.2 Answer probability experiments

The second experiment, *answer probability*, begins to explore whether opinion information can be

[8]The MRR is computed as the average of $1/r$, where r is the rank of the first answer.

[9]MRR has been accepted as the standard performance measure in QA, since MRFA can be strongly affected by outlier questions. However, the MRR score is dominated by the results in the high end of the ranking. Thus, MRFA may be more appropriate for our experiments because the filters are an intermediate step in the processing, the results of which other MPQA components may improve.

			sentence	
			fact	opinion
question	Manual	fact	56 (46.67%)	64 (53.33%)
		opinion	42 (10.14%)	372 (89.86%)
	Naive Bayes	fact	49 (40.83%)	71 (59.17%)
		opinion	57 (13.77%)	357 (86.23%)
	Rulebased	fact	96 (80.00%)	24 (20.00%)
		opinion	184 (44.44%)	230 (55.56%)

Table 6: Answer probability results.

used in an **answer generator**. This experiment considers correspondences between (1) the classes (i.e., opinion or fact) assigned by the subjectivity filters to the sentences containing answers, and (2) the classes of the questions the answers are responses to (according to the OpQA annotations). That is, we compute the probabilities (where *ans* = answer):

P(*ans* is in a *C1* sentence | *ans* is the answer to a *C2* question) for all four combinations of *C1*=opinion, fact and *C2*=opinion, fact.

Results. Results for the answer probability experiment are given in Table 6. The rows correspond to the classes of the questions the answers respond to, and the columns correspond to the classes assigned by the subjectivity filters to the sentences containing the answers. The first two rows, for instance, give the results for the MANUAL criterion. MANUAL placed 56 of the answers to fact questions in fact sentences (46.67% of all answers to fact questions) and 64 (53.33%) of the answers to fact questions in opinion sentences. Similarly, MANUAL placed 42 (10.14%) of the answers to opinion questions in fact sentences, and 372 (89.86%) of the answers to opinion questions in opinion sentences.

The answer probability experiment sheds some light on the subjectivity filter experiments. All three subjectivity filters place a larger percentage of answers to opinion questions in opinion sentences than they place in fact sentences. However, the different filters exhibit different degrees of discrimination. Answers to opinion questions are almost always placed in opinion sentences by MANUAL (89.86%) and NAIVE BAYES (86.23%). While that aspect of their performance is excellent, MANUAL and NAIVE BAYES place more answers to fact questions in opinion rather than fact sentences (though the percentages are in the 50s). This is to be expected, because MANUAL and NAIVE BAYES are more conservative and err on the side of classifying sentences as opinions: for MANUAL, the presence of any subjective expression makes the entire sentence opinion, even if parts of the sentence are factual; NAIVE BAYES shows high recall but lower precision in recognizing opinion sentences (see Table 4). Conversely, RULEBASED places 80% of the fact answers in fact sentences and only 56% of the opinion answers in opinion sentences. Again, the lower number of assignments to opinion sentences is to be expected, given the high precision and low recall of the classifier. But the net result is that, for RULEBASED, the off-diagonals are all less than 50%: it places more answers to fact questions in fact rather than opinion sentences (80%), and more answers to opinion questions in opinion rather than fact sentences (56%). This is consistent with its superior performance in the subjectivity filtering experiment.

In addition to explaining the performance of the subjectivity filters, the answer rank experiment shows that the automatic opinion classifiers can be used directly in an **answer generator** module. The two automatic classifiers rely on evidence in the sentence to predict the class (the information extraction patterns used by RULEBASED and the features used by NAIVE BAYES). In ongoing work we investigate ways to use this evidence to extract and summarize the opinions expressed in text, which is a task similar to that of an **answer generator** module.

7 Opinion Source Filters for MPQA Systems

In addition to subjectivity filters, we also define an opinion *source filter* based on the manual opinion annotations. This filter removes all sentences that do not have an opinion annotation with a source that matches the source of the question[10]. For this filter we only used the MANUAL source annotations since we did not have access to automatically extracted source information. We employ the same Answer Rank experiment as in 6.3.1, substituting the source filter for a subjectivity filter.

Results. Results for the source filter are mixed. The filter outperforms the baseline on some questions and performs worst on others. As a result the MRR for the source filter is worse than the base-

[10] We manually identified the sources of each of the 15 opinion questions.

line (0.4633 vs. 0.4911). However, the source filter exhibits by far the best results using the MRFA measure, a value of 11.267. The performance improvement is due to the filter's ability to recognize the answers to the hardest questions, for which the other filters have the most trouble (questions *mugabe* 7 and 8). For these questions, the rank of the first answer improves from 153 to 21, and from 182 to 11, respectively. With the exception of question *venezuela* 3, which does not contain a clear source (and is problematic altogether because there is only a single answer in the corpus and the question's qualification as opinion is not clear) the *source filter* always ranked an answer within the first 25 answers. Thus, *source filters* can be especially useful in systems that rely on the presence of an answer within the first few ranked answer segments and then invoke more sophisticated analysis in the **additional processing** phase.

8 Conclusions

We began by giving a high-level overview of the OpQA corpus. Using the corpus, we compared the characteristics of answers to fact and opinion questions. Based on the different characteristics, we surmise that traditional QA approaches may not be as effective for MPQA as they have been for fact-based QA. Finally, we investigated the use of machine learning and rule-based opinion filters and showed that they can be used to guide MPQA systems.

Acknowledgments We would like to thank Diane Litman for her work eliciting the questions for the OpQA corpus, and the anonymous reviewers for their helpful comments.This work was supported by the Advanced Research and Development Activity (ARDA), by NSF Grants IIS-0208028 and IIS-0208798, by the Xerox Foundation, and by a NSF Graduate Research Fellowship to the first author.

References

Steven Abney. 1996. Partial parsing via finite-state cascades. In *Proceedings of the ESSLLI '96 Robust Parsing Workshop*.

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*.

S. Das and M. Chen. 2001. Yahoo for amazon: Extracting market sentiment from stock message boards. In *Proceedings of the 8th Asia Pacific Finance Association Annual Conference*.

Kushal Dave, Steve Lawrence, and David Pennock. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *International World Wide Web Conference*, pages 519–528.

D. Moldovan, S. Harabagiu, R. Girju, P. Morarescu, F. Lacatusu, A. Novischi, A. Badulescu, and O. Bolohan. 2002. LCC tools for question answering. In *Proceedings of TREC 2002*.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the ACL*, pages 271–278.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of EMNLP*.

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

Ellen Riloff. 1996. Automatically generating extraction patterns from untagged text. *Proceedings of AAAI*.

V. Stoyanov, C. Cardie, J. Wiebe, and D. Litman. 2004. Evaluating an opinion annotation scheme using a new Multi-Perspective Question and Answer corpus. In *2004 AAAI Spring Symposium on Exploring Attitude and Affect in Text*.

Peter Turney. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the ACL*, pages 417–424.

E. Voorhees and L. Buckland. 2003. Overview of the TREC 2003 Question Answering Track. In *Proceedings of TREC 12*.

Ellen Voorhees. 2001. Overview of the TREC 2001 Question Answering Track. In *Proceedings of TREC 10*.

Ellen Voorhees. 2002. Overview of the 2002 Question Answering Track. In *Proceedings of TREC 11*.

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

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 1(2).

Theresa Wilson and Janyce Wiebe. 2003. Annotating opinions in the world press. *4th SIGdial Workshop on Discourse and Dialogue (SIGdial-03)*.

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*.