Evaluating an Opinion Annotation Scheme Using a New Multi-Perspective Question and Answer Corpus

Veselin Stoyanov and **Claire Cardie**
Dept. of Comp. Science, Cornell University
Ithaca, NY 14850, USA
ves,cardie@cs.cornell.edu

Diane Litman and **Janyce Wiebe**
Dept. of Comp. Science, Univ. of Pittsburgh
Pittsburgh, PA 15260, USA
litman,wiebe@cs.pitt.edu

Abstract

In recent work, Wiebe et al. (2003; 2002) propose a semantic representation for encoding the opinions and perspectives expressed at any given point in a text. This paper evaluates the opinion annotation scheme for multi-perspective vs. fact-based question answering using a new question and answer corpus.

Introduction

In recent work, Wiebe *et al.* (2003; 2002) propose a semantic representation for encoding the opinions and perspectives expressed at any given point in a text. In addition, they develop the NRRC[1] corpus — a collection of 252 articles that are manually annotated according to this opinion representation scheme (Wiebe *et al.* 2003; Wilson & Wiebe 2003). Cardie *et al.* further hypothesize that such representations will be useful for practical natural language processing (NLP) applications like multi-perspective question answering (Cardie *et al.* 2003). In multi-perspective question answering (MPQA), for example, the goal of the NLP system is to answer opinion-oriented questions (e.g. "What is the sentiment in the Middle East towards war on Iraq?") rather than fact-based questions (e.g. "What is the primary substance used in producing chocolate?"). To be successful, such MPQA systems will presumably require the ability to recognize and organize the opinions expressed throughout one or more documents. To date, however, the proposed opinion annotation scheme has not been directly studied in this question-answering context.

The goals of this paper are two-fold. First, we present a new corpus of multi-perspective questions and answers. This Q&A corpus contains 15 opinion-oriented questions and 15 fact-oriented questions along with all text spans that constitute the answers to these questions for a subset of the documents in the above-mentioned NRRC corpus. Second, we present the results of two experiments that employ the new Q&A corpus to investigate the usefulness of the Wiebe *et al.*'s opinion annotation scheme for multi-perspective vs. fact-based question answering. We find ultimately that low-level perspective information can be useful in MPQA if used judiciously.

The paper is organized as follows. The next section provides a brief overview of Wiebe *et al.*'s opinion annotation framework and the NRRC opinion-annotated corpus. We then present the question and answer (Q&A) corpus, followed by a section that describes our evaluation using the new corpus and discusses the results.

[1]The corpus was created during a workshop on multi-perspective question answering sponsored by ARDA's Northeast Regional Research Center at Mitre.

Low-Level Perspective Information

The framework suggested by Wiebe *et al.* (2002) provides a basis for annotating opinions, beliefs, emotions, sentiment, and other private states expressed in text. *Private state* is a general term used to refer to mental and emotional states that cannot be directly observed or verified (Quirk *et al.* 1985).

There are two principal ways in which private states are expressed in language: they could be explicitly stated, or they could be expressed indirectly by the selection of words and the style of language that the speaker or writer uses. For instance, in the sentence "John is afraid that Sue might fall," "afraid" is an explicitly mentioned private state. On the other hand, the sentence "It is about time that we end Saddam's oppression," does not mention explicitly the opinion of the author, but the private state of disapproval of Saddam is expressed by the words and style of the language used: the phrases "it is about time" and "oppression" are examples of what Wiebe *et al.* call *expressive subjective elements*.

An important aspect of a private state is its *source*. The source of a private state is the experiencer of that state, that is, the person or entity whose opinion or emotion is being conveyed in the text. Trivially, the overall source is the author of the article, but the writer may write about the private states of other people, leading to multiple sources in a single text segment. For example, in the sentence "Mary believes that Sue is afraid of the dark," the private state of Sue being afraid is expressed through Mary's private state (of "believing") and Mary's private state is expressed through the implicit private state of the author of the sentence. This presents a natural *nesting of sources* in a text segment. Nesting of sources may become quite deep and complex, and expressive subjective elements may also have nested sources.

The perspective annotation framework suggested by Wiebe *et al.* (2002) includes annotations to describe expres-

Explicit private state
onlyfactive: *yes, no*
nested-source
overall-strength: *low, medium, high, extreme*
on-strength: *neutral, low, medium, high, extreme*
attitude-type: *positive, negative, both*
attitude-toward
is-implicit
minor
Expressive subjective elements
nested-source
strength: *low, medium, high, extreme*
attitude-type: *positive, negative, both*

Table 1: Attributes for the two main annotation types. For annotations that can take values from a fixed set, all values are given.

sive subjective elements as well as explicitly-mentioned private states and speech events.

Annotations for private states and speech events are comprised of what Wiebe *et al.* (2002) refer to as the *on* — the text span that constitutes the private state or speech event phrase itself — as well as the *inside* of the speech event, which is the text segment inside the scope of the private state or speech event phrase. For instance, in the sentence "Tom believes that Ken is an outstanding individual," the *on* is "believes" and the *inside* is "Ken is an outstanding individual." Similarly, in the sentence "Critics say that the new system will fail," the *on* is "say" and the *inside* is "the new system will fail."

An important aspect of each private state and speech event annotation is encoded in its *onlyfactive* attribute (Wiebe 2002). This attribute indicates whether the associated text segment is presented as factual (i.e. *onlyfactive=yes*), or indeed expresses the emotion, opinion, or other private state of the source (i.e. *onlyfactive=no*). For example, all expressions that are explicit private states such as "think" and "believe" as well as private states mixed with speech such as "praise" and "correct" by definition are *onlyfactive=no*, whereas neutral speech events such as "said" and "mentioned" may be either *onlyfactive=no* or *onlyfactive=yes*, depending on the context.

In contrast, the text span associated with expressive subject element annotations is simply that of the subjective phrase itself. The attributes that can be assigned to each of the two annotation types are summarized in Table 1.

This investigation considers both *explicit private state* and *expressive subjective element* annotations. Furthermore, the investigation makes use of the *onlyfactive* attribute of the *explicit private state* annotations as an indicator of whether the annotation should be considered factive or expressing opinion.[2] In particular, we will use the term **fact annotation** to refer to an *explicit private state* annotation with its *onlyfactive* attribute set to *yes* and **opinion annotation** to refer to either an explicit private state annotation with its *onlyfactive* attribute set to *no* or an expressive subjective element.

[2]Using other attributes of the annotation would require specific processing adapted for the MPQA task and goes beyond the scope of the current investigation.

The MPQA NRRC Corpus

Using the perspective annotation framework, Wiebe *et al.* have manually annotated a considerable number of documents (over 100 reported in Wiebe *et al.* (2003) and 252 reported in Wilson & Wiebe (2003)) to form the NRRC corpus. The annotated documents are part of a larger data collection of over 270,000 documents that appeared in the world press over an 11-month period, between June 2001 and May 2002 (Wiebe *et al.* 2003). The source of almost all of the documents in the larger collection is the U.S. foreign broadcast information service (FBIS).

Note that documents in the NRRC corpus have not been annotated with *insides* for all private states and speech events. The only private state annotations that include *insides* are those that span entire sentences.[3]

Wiebe *et al.* have performed interannotator studies to validate the annotations by assessing the consistency of human annotators. In particular, they report an interannotator agreement of 85% on direct expressions of perspective information (*explicit private states*), about 50% on indirect expressions of subjective information (*expressive subjectivity*), and up to 80% kappa agreement on the rhetorical use of perspective information (Wiebe *et al.* 2003). In a subsequent study, the average of the reported values for agreement between groups was 82% for *on* agreement and 72% for *expressive-subjective* agreement (Wilson & Wiebe 2003). Values for both studies were reported using measure *agr(a||b)* for annotator groups *a* and *b* calculated as the proportion of *a*'s annotations that were found by *b*. For every two groups *a* and *b* a value was calculated as the mean of *agr(a||b)* and *agr(b||a)*, since the measure is directional.

Wiebe *et al.* (2003) concluded that the good agreement results indicate that annotating opinions is a feasible task, and suggest ways for further improving the annotations.

Multi-Perspective Question and Answer Corpus Creation

This section describes the creation of the question and answer (Q&A) corpus used to evaluate the low-level perspective annotations in the context of opinion-oriented (*opinion*) and fact-based (*fact*) question answering.

The Q&A corpus consists of 98 documents from the opinion-annotated NRRC corpus. Each document addresses one of four general topics:

kyoto concerns President Bush's alternative to the Kyoto protocol;

mugabe concerns the 2002 elections in Zimbabwe and Mugabe's reelection;

humanrights discusses the US annual human rights report; and

venezuela describes the 2002 coup d'estat in Venezuela.

[3]These have been identified automatically and added to the corpus.

The documents were automatically selected from the bigger set of over 270,000 documents as being relevant to one of the four topics using the SMART (Salton 1971) information retrieval system. The Q&A corpus contains between 19 and 33 documents for each topic.

Fact and opinion questions for each topic were added to the Q&A corpus by a volunteer not associated with the current project. He was given two randomly selected documents on each topic along with a set of instructions for creating fact vs. opinion questions.[4] The complete set of 30 questions is shown in Table 2. The set contains an equal number of opinion (o) and fact (f) questions for each topic.

Once the documents and questions were obtained, answers for the questions in the supporting documents had to be identified. In particular, we manually added *answer* annotations for every text segment in the Q&A corpus that constituted, or contributed to, an answer to any question. The *answer* annotations include attributes to indicate the **topic** of the associated question, the **question number** within that topic, and the annotator's **confidence** that the segment actually answered the question. Annotators did not have access to the low-level perspective annotations during answer annotation.

Documents were annotated by the first two authors of the paper, with each annotator handling 61 documents.[5] Out of the 98 documents in the collection, 24 were selected at random and annotated by both annotators. The remaining 74 documents were split equally between the two annotators using a random draw. The 24 documents that were annotated by both annotators were used to study the interannotator agreement. Using Wiebe *et al.*'s (2003) *agr* measure, we determined that the agreement between the two annotators was 85% on average with values of 78% and 93% for the two annotators. The good interannotator agreement indicates that, despite the difficulties, annotating the answers is a feasible task and can be performed consistently in the presence of robust annotation instructions.

Difficulties in Corpus Creation

This section summarizes some of the difficulties encountered during creation of the Q&A corpus.

Question Creation. In spite of the question creation instructions, it appears that some questions were reverse-engineered from the available documents. These questions are answered in only one or two of documents, which presents some challenges when using the collection for evaluation. Nevertheless, the setting is not unrealistic since the situation in which questions find support in only a few documents is often present in real-world QA systems.

In addition, the classification associated with each question — fact or opinion — did not always seem appropriate. For instance, **mugabe** opinion question #6 — "What

[4]Space limitations preclude the inclusion of those instructions, which are available from `www.cs.cornell.edu/home/cardie/`.

[5]Again, space constraints preclude our inclusion of the answer annotation instructions here. They are available at: `www.cs.cornell.edu/home/cardie/`.

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 2: Questions in the Q&A collection by topic.

is the basis for the European Union and US critical attitude and adversarial action toward Mugabe?" — could arguably be classified as fact-based, since the question is in essence not asking about the European Union and US's opinion, but rather about the basis for it. Similarly, **venezuela** factual question #2 — "Did any prominent Americans plan to visit Venezuela immediately following the 2002 coup?" — could be judged as asking about the opinion of prominent Americans.

Annotating Answers. The most frequently encountered problem in answer annotation is a well-known problem from fact-based QA; namely, the difficulty of deciding what constitutes an answer to a question. The problem was further amplified by the presence of opinion questions. For instance, the question "Did most Venezuelans support the 2002 coup?" had potential answers such as "Protesters...failed to gain the support of the army" and "... thousands of citizens rallied the streets in support of Chavez." Both segments hint that most Venezuelans did not support the coup that forced Chavez to resign. Both passages, however, state it in a very indirect way. It is hard even for humans to conclude whether the above two passages constitute answers to the question.

A related issue is that opinionated documents often express answers to the questions only very indirectly, by using word selection and style of language (*expressive subjectivity*), which is often hard to judge. An indication of the difficulties associated with judging the subjectivity expressed indirectly is contained in the interannotator studies reported by Wiebe *et al.* (2003), which showed that annotators agree less often on *expressive subjectivity* (50% of the time) than on direct expressions of opinions (80% of the time).

An additional problem is that opinion questions often ask about the opinions of certain collective entities, such as countries, governments, and popular opinions. It was hard for human annotators to judge what can be considered an expression of the opinion of collective entities (e.g. what sources represent "ordinary Venezuelans" or "the Japanese" or "Japan"?), and often the conjecture required a significant amount of background information (e.g. knowing what countries are "EU" countries or "U.S. allies").

Evaluation of Perspective Annotations for MPQA

We designed two different experiments to evaluate the usefulness of the perspective annotations in the context of fact- and especially opinion-based QA. The first experiment, *answer probability*,

1. visits each answering text segment (as denoted by the manual answer annotations),
2. categorizes it as either OPINION or FACT based on the associated perspective annotations (using one of the criteria described below), and
3. counts how many FACT/OPINION segments answer fact/opinion questions.

That is, we compute the probabilities *P(*FACT/OPINION *answer* | *fact/opinion question)* for all combinations of fact and opinion questions and answers.

The second experiment, *answer rank*, implements the first step of most contemporary QA systems: given a question from the Q&A corpus as the query, it performs sentence-based information retrieval (IR) on all documents in the collection. We then study the effect of considering only retrieved sentences classified as FACT vs. OPINION (using the criteria below) for fact and opinion questions, respectively, on the performance of the information retrieval (IR) component.

For both experiments, we consider multiple criteria to determine whether a text segment (or sentence) should be considered FACT or OPINION based on the underlying perspective annotations. First, we use two *association criteria* to determine which perspective annotations should be considered associated with an arbitrary text segment.

- For the *overlap* criterion, a perspective annotation is considered associated with the segment if its span includes any part of the segment.
- For the *cover* criterion, a perspective annotation is considered associated with the segment if its span contains the entire text segment.[6]

Once we determine the set of perspective annotations associated with a text segment, we use four *classification criteria* to categorize the segment as one of FACT or OPINION:

***most nested* (*m nested*):** a segment is considered OPINION if the **most nested** annotation from the set of associated perspective annotations is an opinion; the segment is considered FACT otherwise. Note that nested sources can have nested perspective annotations. Overlapping non-nested annotations are not possible if the annotation instructions are followed (Wiebe 2002).

***all*:** a segment is considered OPINION if **all** associated perspective annotations are opinion; FACT otherwise.

***any*:** a segment is considered OPINION if **any** of the associated perspective annotations is opinion; FACT otherwise.

***most*:** a segment is considered OPINION if **the number** of associated perspective annotations that are opinions is **greater than** the number of associated perspective annotations that are fact. A segment is considered FACT otherwise.

The above criteria exhibit a bias towards opinion annotations. Criteria were designed in such a way because we expected opinion annotations to be more discriminative. For instance, if a fact annotation is embedded inside an opinion annotation, the fact expressed in the internal annotation will be expressed from the perspective of the outer source.

Results: Answer Probability

As mentioned above, this experiment counts the number of answer segments classified as FACT and OPINION, respec-

[6] As mentioned earlier, the only *insides* annotated in the Q&A corpus are those that ccover entire sentences. This affects both criteria, but especially *cover*, since it is only these sentence-length *inside* annotations that will ever be considered associated with an answer segment that spans more than a single *on*.

Criterion	Answer type	Question type			
		f	% of ttl	o	% of ttl
overlap m nested	f	84	70.00%	40	9.64%
	o	36	30.00%	375	90.36%
cover m nested	f	94	78.33%	238	57.35%
	o	26	21.67%	177	42.65%
overlap any	f	84	70.00%	34	8.19%
	o	36	30.00%	381	91.81%
cover any	f	94	78.33%	238	57.35%
	o	26	21.67%	177	42.65%
overlap all	f	94	78.33%	307	73.98%
	o	26	21.67%	108	26.02%
cover all	f	94	78.33%	301	72.53%
	o	26	21.67%	114	27.47%
overlap most	f	93	77.50%	223	53.73%
	o	27	22.50%	192	46.27%
cover most	f	94	78.33%	305	73.49%
	o	26	21.67%	110	26.51%

Table 3: Number of fact/opinion questions answered in fact/opinion segments based on each of the 6 criteria (*f* stands for fact and *o* for opinion).

tively, that answer each question. We hypothesize that opinion questions will be answered more often in answer segments classified as OPINION, and that fact questions will be answered more often in text segments classified as FACT. For this experiment we consider every text segment annotated as an answer and examine the perspective annotations associated with the text segment.

The results of this experiment are summarized in Table 3. Table 3 has eight rows, one for each combination of association (total of two) and classification (total of four) criteria. For each of the eight criteria, Table 3 shows the total number of fact and opinion questions answered in text segments classified as FACT and OPINION. Overall, there were 120 answers annotated for fact questions and 415 answers annotated for opinion questions. The first row of the table, for example, indicates that 84 of the answers to fact questions were classified as FACT using the *overlap m nested* criterion. This represents 70% of all fact questions. Similarly, 375 of the answers to opinion questions (90.35% of the total) were classified as OPINION using the text same *overlap m nested* criterion.

Several interesting observations can be made from Table 3. First, for each of the eight criteria, the percentage of fact questions answered in FACT text segments is significantly greater than the percentage of fact questions answered in OPINION segments (e.g. 70.00% vs. 30.00% for *overlap m nested*). Furthermore, for two of the eight criteria, namely *overlap m nested* and *overlap any*, the percentage of opinion questions answered in OPINION segments is greater than the percentage of opinion question answered in FACT segments (e.g. 90.36% vs. 9.64% for *overlap m nested*). Additionally, for five of the eight criteria, excluding *overlap all*, *cover all*, and *cover most*, *P(*FACT *answer* | *fact question)* is significantly greater than *P(*FACT *answer* | *opinion question)* (and symmetrically for opinion answers) (e.g. 70.00% vs. 9.64% for *overlap m nested*).

The most discriminative runs for fact questions appear to be *cover*, with any of the four classification criteria. Using any of the *cover* criteria, 78.33% of the fact questions are answered in FACT segments and only 21.67% are answered in OPINION segments. As for opinion questions, the most accurate criterion is *overlap any*, for which 91.81% of the opinion questions are answered in OPINION segments and only 8.19% in FACT segments. Considering the characteristics of the data, the above results can be expected, since *cover* is more likely to classify segments as FACT than OPINION, with *cover all* being the most restrictive criterion in terms of classifying segments as OPINION. At the same time, *overlap any* is the most liberal criterion, in that it is likely to classify the most segments as OPINION. Two of the four *overlap* criteria, namely *overlap m nested* and *overlap any* appear to exhibit a good balance between classifying answers to fact questions as FACT and at the same time classifying opinion question answers as OPINION. These two criteria show the two best performances on opinion questions, while diverging from the best performance on fact questions only slightly. The best predictor for the classification of the answer, however, appears to be a combined measure that relies on *overlap any* for opinion questions and on any of four *cover* criteria for the fact questions. For such a combined criterion, 78.33% of the answers to fact question appear in segments classified as FACT and 91.81% of the answers to opinion questions appear in segments classified as OPINION.

A somewhat surprising fact is that all four variations of the *cover* criterion exhibit identical performance. This is due to the fact that in most cases the only perspective annotation segments that cover answer text segments spanning more than a single *on* are perspective annotations that span the entire sentence, as described in the experimental setup section.

Results: Answer Rank

The second experiment is designed to resemble the operation of a traditional QA system. More precisely, we attempt to determine whether information from the perspective annotations can assist in the IR phase of traditional QA approaches. The hypothesis is that perspective annotations can be useful in ranking the retrieved text segments. More precisely, we hypothesize that low-level perspective information can be used to promote the correct answer segments in the ranking.

For this experiment, we divide each document into a set of text segments at sentence borders. We then run an IR algorithm (the standard tf.idf retrieval implemented in the Lemur IR kit, available from `http://www.cs.cmu.edu/ lemur/`) on the set of all sentences from all documents in the Q&A collection, treating each question, in turn, as the query. We then refine the ranked list of sentences returned by Lemur for each particular question. We optionally applying one of two filters, each of which removes OPINION answers for fact questions and vice versa. The two filters constitute the two best performing criteria from the *answer probability* experiment for opinion and fact questions, *overlap any* criterion to classify a retrieved answer and *cover all*, respectively. From the mod-

Topic	Q#	Rank of first answer			
		unfilt	overlap	cover	mixed
Kyoto	1 f	10	4	6	6
	2 f	1	1	1	1
	3 f	3		3	3
	4 f	2	2	2	2
	5 o	1	1		1
	6 o	5	4	2	4
	7 o	1	1	1	1
	8 o	1	1	2	1
Hum Rights	1 f	1	1	1	1
	2 f	1	14	1	1
	3 o	1	1	10	1
	4 f	1		1	1
	5 o	10	7	24	7
	6 o	1	1	1	1
Vene-zuela	1 f	50	9	32	32
	2 f	13		9	9
	3 o	106	93	44	93
	4 o	3	3	7	3
	5 f	2		1	1
	6 o	1	1	1	1
	7 o	3	3	2	3
	8 f	1	1	1	1
Mugabe	1 o	2	2	39	2
	2 f	64	89	55	55
	3 f	2		2	2
	4 f	16	15	16	16
	5 f	1	117	1	1
	6 o	7	6	111	6
	7 o	447	356		356
	8 o	331	260		260
MRR:		0.52	0.39	0.45	0.55
MRFA:		36.27	39.72	13.92	29.07
fact questions only:					
MRR:		0.54	0.27	0.58	0.58
MRFA:		11.2	25.3	8.8	8.8
opinion questions only:					
MRR:		0.51	0.52	0.32	0.52
MRFA:		61.33	49.33	20.33	49.33

Table 4: Results for IR module evaluation. An *f* after the question number indicates a fact question and *o* indicates opinion.

ified ranked list of answers, we determine the rank of the first retrieved sentence that correctly answers the question. A sentence is considered a correct answer if any part of it is annotated as answer to the question in the Q&A corpus.

After the ranking from the IR system are refined we obtain for each question the rank of the first sentence containing a correct answer to the question (1) without using the perspective annotations (*unfilt* ranking), and (2) using one of the two filters. If our hypothesis is supported, we would expect to see a higher ranking for the first correct answer for each question in runs that make use of the perspective-based filters.

Table 4 summarizes the results from the answer rank experiment. It shows the rank of the first answering sentence for every question in the collection. Table 4 has four columns, one for the baseline *unfiltered* results, one for each of the *overlap any* and *cover any* perspective-based filters, and one for a filter that combines the two filters (*mixed*). The *mixed* filter combines the *overlap* and *cover* filters, using *overlap* to filter answer sentences for opinion questions, and *cover* to filter answers for fact questions. The construction of the *mixed* filter was motivated by observing from the data in Table 3 that *overlap any* discriminates well answers to opinion questions, while *cover any* discriminates well answers to fact questions.

Table 4 computes two cumulative measures as well, the Mean Reciprocal Rank (MRR) of the first correct answer, which is a standard evaluation measure in QA, and the mean rank of the first correct answer (MRFA). MRR is computed as the average of the reciprocals of the ranks of the first correct answer (i.e. if the first correct answer to a question is ranked 4, the contribution of the question to the mean will be 1/4). The two cumulative measures are computed across all of the questions and also for fact and opinion questions separately for each of the four rankings.

We see from Table 4 that in the ranking using the *overlap* filter the first OPINION answer for each of the 15 opinion questions in the collection is at least as highly ranked as in the *unfiltered* ranking. As a result, the MRR for *overlap* is higher than the MRR for *unfiltered* for opinion questions. Similarly, in the *cover* ranking the first FACT answer for each of the 15 fact questions in the collection is at least as highly ranked as in the *unfiltered* ranking. Thus, the MRR for *cover* for fact questions is higher than MRR for *unfiltered* for fact questions. At the same time, for five of the fact questions, *overlap* filters all answering segments, returning no sentence answering the question. Similarly, *cover* fails to return answering sentences for three of the opinion questions.

Since *overlap* always outperforms *unfiltered* for opinion questions and *cover* always outperforms *unfiltered* for fact questions, it is not surprising that *mixed* performs at least as well as *unfiltered* on every question in the collection. As a result, *mixed* exhibits an overall MRR of .55 as opposed to *unfiltered*'s MRR of .52. The mean rank of the first correct answer for *mixed* is 29.07 as opposed to 36.27 for *unfiltered*.

Discussion

Results of the first experiment support the hypothesis that low-level perspective information can be useful for multi-perspective question answering. The discriminative abilities of the criteria show that perspective information can be a reliable predictor of whether a given segment of a document answers an opinion/fact question. More specifically, an MPQA system might use the low-level perspective information in one of two ways: the system can combine the two top-performing criteria on fact and opinion questions, or can use one of the two highly performing *overlap* criteria, *overlap all* and *overlap any*. The low-level perspective information may be used to re-rank potential answers by using the knowledge that the probability that a fact answer appears in an OPINION segment, and vice versa, is very low.

An interesting observation constitutes the performance of the eight criteria on questions that were identified as problematic in their fact/opinion classification during corpus creation. Such questions are discussed in the corpus creation

section. The performance of all eight criteria on the problematic questions was worse than the performance on the rest of the questions in the collection. For instance, one of the question given as example in the corpus creation section, "What is the basis for the European Union and US critical attitude and adversarial action toward Mugabe?" (*mugabe*, question #6), is answered at least as often from FACT text segments as from OPINION segments for all of the eight criteria, despite being classified as opinion. An MPQA system that can classify questions as fact or opinion and assign a confidence to the assignment might be able to recognize such situations and rely less on the low-level perspective information for "borderline" questions.

The second experiment provides further evidence in support of the hypothesis that low-level perspective information can be useful in MPQA. An IR subsystem has been an important part of almost all existing effective QA systems (Cardie *et al.* 2000; Moldovan *et al.* 1999; 2002; Pasca & Harabagiu 2000; Harabagiu *et al.* 2001; Voorhees & Tice 1999; Voorhees 2000; 2001; 2002). Our results suggest that, if used properly, low-level perspective information can improve the ranking of potential answer segments returned by the IR subsystem. Our experiments show that the most effective criterion that can be used for re-ranking is *mixed*. Using filters, however, can sometimes cause all answering segments for a particular question to be discarded.

Based on the results of *answer ranking*, we can conclude that while being good predictor for re-ranking of the results from the IR subsystem, low-level perspective information should not be used as an absolute indicator of the relevance of a potential answer segment. In particular, low-level perspective information helps improve the ranking, but in doing so at least some answering summaries are discarded, which can prove costly if the system uses a limited set of supporting documents. The number of discarded entities is smaller for *mixed*, which provides the most conservative estimation.

In summary, both the *answer probability* and the *answer rank* experiments shows that low-level perspective information can be a generally useful predictor of whether a text segment answers a question given the type of the question. It is unrealistic, however, to use the FACT/OPINION segment classification as an absolute indicator of whether the segment can answer fact/opinion questions. Completely disregarding potential answer segments of the incorrect type can cause an MPQA system to eliminate all answer to a question in the supporting collection. This is less of a concern for systems that rely on a larger supporting set of documents (i.e. the World Wide Web), but a valid limitation to systems built to use restricted support document sets.

Conclusions and Future Work

The current investigation addressed two main tasks: constructing a data collection for MPQA and evaluating the hypothesis that low-level perspective information can be useful for MPQA. Both tasks provided insights into potential difficulties of the task of MPQA and the usefulness of the low-level perspective information.

As a result of the first task, a small data collection for MPQA was constructed. The current collection consists of 98 manually annotated documents and a total of 30 questions divided into four topics. As part of future work, the collection can be improved using questions from a real-world question logs.

During the collection construction phase some of the potential difficulties associated with the tasks of MPQA where identified. The main problems identified consist of the problem of deciding what constitutes answer, the presence of indirect answers (*expressive subjectivity*), the difficulty of judging what constitutes an opinion of a collective entity, and the fact that most answers to opinion questions are not stated explicitly in the text, but have to be deduced.

The investigation showed that low-level perspective information can be an effective predictor of whether a text segment contains an answer to a question, given the type of the question. The results, however, suggest that low-level perspective information should not be used as an absolute indicator of whether a segment answers a particular question, especially in the setting where each question is expected to be answered in a limited number of documents.

References

Cardie, C.; Ng, V.; Pierce, D.; and Buckley, C. 2000. Examining the role of statistical and linguistic knowledge sources in a general-knowledge question-answering system. In *Proceedings of the Sixth Applied Natural Language Processing Conference*, 180–187.

Cardie, C.; Wiebe, J.; Wilson, T.; and Litman, D. 2003. Combining low-level and summary representations of opinions for multi-perspective question answering. *Working Notes of the 2003 AAAI Spring Symposium on New Directions in Question Answering*.

Harabagiu, S.; Moldovan, D.; Pasca, M.; Surdeanu, M.; Mihalcea, R.; Girju, R.; Rus, V.; Lacatusu, F.; Morarescu, P.; and Bunescu, R. 2001. Answering complex, list and context questions with lcc's question-answering server. In Voorhees, E., and Harman, D. K., eds., *Proceedings of the Tenth Text Retrieval Conference (TREC 2001)*. 355–362.

Moldovan, D.; Harabagiu, S.; Pasca, M.; Mihalcea, R.; Girju, R.; Goodrum, R.; and Rus, V. 1999. Lasso: A tool for surfing the answer net. In Voorhees, E., and Harman, D. K., eds., *Proceedings of the Eighth Text Retrieval Conference (TREC-8)*.

Moldovan, D.; Harabagiu, S.; Girju, R.; Morarescu, P.; Lacatusu, F.; Novischi, A.; Badulescu, A.; and Bolohan, O. 2002. Lcc tools for question answering. In Voorhees, E., and Buckland, L. P., eds., *Proceedings of the The Eleventh Text REtrieval Conference (TREC 2002)*. 79–89.

Pasca, M., and Harabagiu, S. 2000. High performance question/answering. In *Proceedings of the 38th annual meeting of the association for computational linguistics (ACL-2000)*, 563–570.

Quirk, R.; Greenbaum, S.; Leech, G.; and Svartvik, J. 1985. *A comprehensive grammar of the English language*. New York: Longman.

Salton, G., ed. 1971. *The SMART Retrieval System—*

Experiments in Automatic Document Processing. Englewood Cliffs, NJ: Prentice Hall Inc.

Voorhees, E., and Tice, D. 1999. The TREC-8 question answering track evaluation. In Voorhees, E., and Harman, D. K., eds., *Proceedings of the Eighth Text Retrieval Conference (TREC-8)*. 83–105.

Voorhees, E. 2000. The TREC-9 question answering track evaluation. In Voorhees, E., and Harman, D. K., eds., *Proceedings of The Ninth Text REtrieval Conference (TREC-9)*. 71–81.

Voorhees, E. 2001. Overview of the TREC 2001 question answering track. In Voorhees, E., and Harman, D. K., eds., *Proceedings of the Tenth Text Retrieval Conference (TREC 2001)*. 42–52.

Voorhees, E. 2002. Overview of the TREC 2002 question answering track. In Voorhees, E., and Buckland, L. P., eds., *Proceedings of the The Eleventh Text REtrieval Conference (TREC 2002)*. 53–75.

Wiebe, J.; Breck, E.; Buckley, C.; Cardie, C.; Davis, P.; Fraser, B.; Litman, D.; Pierce, D.; Riloff, E.; Wilson, T.; Day, D.; and Maybury, M. 2003. Recognizing and organizing opinions expressed in the world press. *Working Notes of the 2003 AAAI Spring Symposium on New Directions in Question Answering*.

Wiebe, J. 2002. Instructions for annotating opinions in newspaper articles. Department of Computer Science TR-02-101, University of Pittsburgh, Pittsburgh, PA.

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