QA with Attitude: Exploiting Opinion Type Analysis for Improving Question Answering in On-line Discussions and the News

Swapna Somasundaran* swapna@cs.pitt.edu
Theresa Wilson‡ twilson@inf.ed.ac.uk
Janyce Wiebe* wiebe@cs.pitt.edu
Veselin Stoyanov♯ ves@cs.cornell.edu

*Department of Computer Science, University of Pittsburgh, Pittsburgh, PA 15260, USA
‡ School of Informatics, University of Edinburgh, Edinburgh EH8 9LW, UK
♯Department of Computer Science, Cornell University, Ithaca, NY 14853, USA

Abstract

In this paper, we explore the utility of attitude types for improving question answering (QA) on both web-based discussions and news data. We present a set of attitude types developed with an eye toward QA and show that they can be reliably annotated. Using the attitude annotations, we develop automatic classifiers for recognizing two main types of attitudes: sentiment and arguing. Finally, we exploit information about the attitude types of questions and answers for improving opinion QA with promising results.

1. Introduction

Everyday in weblogs, on-line forums, and review sites, a multitude of people express their feelings and opinions. Forums on popular websites like `www.bbc.com` ("Have your say") and `http://english.aljazeera.net` ("Your Views") particularly elicit readers' opinions and viewpoints on controversial "hot topics." The news is also a source of opinions, and with the web becoming the main portal for news dissemination and news sites allowing readers to comment on articles (e.g., "Raves and Rants" in `www.wired.com`, "What is your opinion?" in `www.washingtonpost.com`, "Talkback" in `www.jpost.com`) even the news may be considered a type of social media. Over the past few years, this availability of opinions on the web has fueled new avenues of research in automatic subjectivity and sentiment analysis, including mining and summarizing product reviews (e.g., [4, 6]), classifying the sentiment of reviews (e.g., [16, 27]), and analyzing blogger mood and sentiment (e.g., [13, 2]). It has also sparked new research with applications such as information retrieval (IR) and question answering (QA), for example, using IR to retrieve opinions from weblogs [12] and QA systems to answer opinion questions [26]. In this paper, we exploit information about the attitude types of questions and answers to improve the capabilities of opinion question answering systems for news and web based discussions.

Opinion questions are questions that directly query information about what people think or feel and questions with answers that reflect a variety of perspectives on the topic. Q1 below is an example of the first type of opinion question; Q2 is an example of a question for which there are a range of answers reflecting different viewpoints.

Q1 Are you worried about climate change?

Q2 What will be the effect of reporting Iran to the Security Council?

In recent work, Stoyanov et al. [26] investigated the difference between answers to fact-based questions (e.g., *Who was the first space tourist?*) and answers to opinion questions. They showed that using subjectivity classifiers to filter out sentences containing only objectively presented information improved the ranking of answers to opinion questions. However, subjective language may be used to express any number of different types of attitudes, including, among others, beliefs, opinions, emotions, judgments, and speculations. Opinion questions, on the other hand, often target a particular type of attitude in the answers that they are seeking. For example, in the questions above, Q1 is asking about people's sentiments, while Q2 seems to be asking for people's beliefs or arguments. We hypothesize that giving a QA system more fine-grained information about the attitude types of questions and answers will help to improve the performance of opinion QA more than simply distinguishing between subjective and factual information.

To explore this hypothesis, we first developed an annotation scheme for marking attitude types. There were many choices (e.g., affective lexicons, psychological models of emotion, appraisal theory) to consider when deciding what types of attitudes to annotate. The set of attitudes we settled on was based on explorations of the data with QA in mind. Using the attitude annotation scheme, we annotate the corpus created by Stoyanov et al. [26] for use in opinion QA. This corpus, the attitude annotation scheme, and agreement studies showing that the attitudes can be reliably annotated are presented in Sections 2 and 3.

The next step in exploring our hypothesis was to develop automatic systems for recognizing sentences bearing different attitude types. For this, we focused on two main types of attitudes: sentiment and arguing. Sentiments include positive and negative opinions, emotions, and evaluations. Arguing includes beliefs, arguing about what is or is not true, and arguing for or against something. Although many researchers have worked on recognizing sentiment, automatically recognizing arguing is new in this work. As with many approaches to recognizing subjectivity and sentiment, we rely on a lex-

ICWSM'2007 Boulder, Colorado, USA

icon containing information about the subjectivity of clues. Additionally, for this work we take the novel approach of first disambiguating instances of clues from the lexicon to determine which in context are actually being used to express a sentiment or subjectivity. Our automatic systems for recognizing sentiments and arguing are presented in Section 4.

In the last part of the paper (Section 5), we explore whether the manual attitude annotations and results from the automatic attitude recognizers can be exploited for improving QA. One way in which opinion questions differ from many types of fact-based questions is that, rather than having a single best answer, opinion questions often have many relevant answers, which may reflect a variety of different viewpoints. For example, question Q2 above has a number of relevant answers, including all of the following:

- The effect will be that we will get into exactly the same situation as we did with Iraq.
- The benefits of reporting Iran to the security council will ensure a international consensus will be used to develop an appropriate solution to this problem.
- Iran, however, has chosen to call the tune, and will therefore dance before the Security Council, which is correct.

Traditionally, frameworks for evaluating QA systems use the mean reciprocal rank (MRR) and the mean rank of the first answer (MRFA). However, these metrics focus only on the ranking of the first correct answer. When there are many different and relevant answers to a question, as is often the case for opinion questions, MRR and MRFA fail to measure how well the system may actually be performing. Therefore, in this paper we propose to extend the evaluation framework for opinion QA to include two additional metrics: Average Precision and Sliding Ratio. Both of these metrics consider the ranking of all possible answers, not just the first best answer. These two metrics are described in Section 5.2.

2. Data

For the experiments in this paper, we use two corpora: the Multi-perspective Question Answering (MPQA) Opinion Corpus version 1.2[1] [29] and the Have Your Say (HYS) dataset. The MPQA Corpus contains 535 documents from the world press on a variety of topics. All documents in the collection are marked with expression-level opinion annotations. The HYS dataset is a collection of data that we compiled from blogs and from user responses to questions posted on the *BBC: Have Your Say*[2] website. *BBC: Have Your Say* poses questions on current controversial issues and invites readers to answer the questions and give their opinions.

For the attitude classification experiments, we use 284 documents from the MPQA Corpus. These documents are annotated with attitudes as described in the next section.

The QA experiments use the OpQA dataset, a subset of documents from the MPQA Corpus that were annotated with questions and answers by Stoyanov et al. [26], and the HYS dataset. The OpQA dataset contains 98 documents, all of which are included in the set of documents marked with attitude annotations. The HYS dataset contains 1,720 documents of which 1,597 are short reader responses from *BBC: Have Your Say* on 6 debate topics and 123 are weblogs on the same topics. The questions and answers marked on the datasets and used for the QA experiments are described in Section 5.

[1] Freely available from `www.cs.pitt.edu/mpqa`.

[2] `http://news.bbc.co.uk/1/hi/talking_point/default.stm`

3. Attitude annotations

Although in this paper we focus on whether two general types of attitudes, sentiment and arguing, are useful for improving QA, investigating the usefulness of other types of attitudes and more detailed attitude distinctions (e.g., positive sentiment v. negative sentiment) in QA is part of our overall research goal. Exploring the data with this in mind, we developed the following set of attitude types:

Positive Sentiment	*Positive Agreement*	*Speculation*
Negative Sentiment	*Negative Agreement*	*Other Attitude*
Positive Arguing	*Positive Intention*	
Negative Arguing	*Negative Intention*	

Positive sentiments include positive emotions, evaluations, and stances. Negative sentiments include negative emotions, evaluations, and stances. Positive arguing includes beliefs, arguing for something, and arguing that something is true or so. Negative arguing includes disbelief, arguing against something, and arguing that something not true or so. The definition of the remaining attitude types are omitted as they are not the focus of this paper.

We add the attitudes as a layer of additional annotation on top of the existing annotations in the MPQA Corpus [29]. Specifically, for each *direct subjective frame* in a sentence, an annotator creates one or more attitude annotations and links the attitude annotations back to the direct subjective frame. *Direct subjective frames* are used to mark direct references to private states (e.g., *believe* and *admiration*) and speech events expressing private states (e.g., *praise* and *deny*). A *private state* [19] is any internal mental or emotional state. Private states include beliefs, sentiments, evaluations, speculations, intentions, etc.

Each attitude annotation is represented as a frame with the following slots: *ID*, *text span*, *attitude type*, *intensity*, and *target id*. *IDs* are used to link attitude frames to direct subjective frames. The *text span* is the span of text that conveys the attitude. The *attitude type* is one of the 10 listed above. The *intensity* of an attitude is marked using the following scale: *low*, *low-medium*, *medium*, *medium*, *medium-high*, *high*. Finally, the *target id* is used as a link to the annotation capturing the target of the attitude[3].

Sentence (1) gives an example of both arguing and sentiment attitudes; the direct subjective frames are given in bold.

(1) "I **think** people **are happy** because Chavez has fallen."

For the direct subjective frame for "think," there is a positive-arguing attitude: In the context of the sentence, the speaker referred to by "I" is expressing her belief about what people feel about the fall of Chavez. For the direct subjective frame for "are happy," two attitudes are marked. A positive sentiment is marked to capture the positive sentiment of the people toward the fall of Chavez. In addition, a second attitude annotation is marked to capture that the phrase "happy that Chavez has fallen" expresses the people's negative sentiment toward Chavez himself.

The MPQA Corpus version 1.2 does contain some information about sentiments, specifically, contextual polarity judgments. Wilson et al. [30] added the contextual polarity judgments, which can be *positive*, *negative*, *both*, or *neutral*. The contextual polarity annotations are quite useful, and we exploit them later in our classification experiments. However,

[3] Although the targets of attitudes are undoubtedly important for QA, we omit further discussion of target annotations as they are beyond the scope of this current work.

Study 1	Marked	Recall	Arguing	Sentiment
A	515	91%	33%	51%
B	549	86%	38%	44%
Study 2				
A	247	95%	26%	52%
B	283	83%	23%	64%

Table 1: *Attitude annotations in the two studies*

they only provide partial information about the overall sentiments that may be expressed in a sentence. In addition, there is still a need for information about other types of attitudes that may be in the sentence, such as arguing.

3.1 Agreement studies

We conducted two inter-annotator agreement studies. In the first study, two judges independently annotated 13 documents with 325 sentences and 409 direct subjective annotations. Two months later[4], the same two judges annotated another 11 documents with 211 sentences and 207 direct subjective annotations. All intensity and contextual polarity attributes were removed from these documents before each study.

Table 1 gives a brief description of the attitudes marked by the judges in the two studies. The first column shows the number of attitudes marked by each judge. The *Recall* column gives the percentage of those attitude annotations also marked by the other judge. The third and fourth columns show the percentage of attitudes that were of type arguing (positive or negative) and sentiment (positive or negative).

From Table 1 we see that there is a high degree of overlap in the attitudes marked by the two judges. Using the set of annotations that both judges marked, we calculated Cohen's Kappa (κ) and percent agreement for attitude types. For Study 1, judges have a κ of 0.79 (83%), and for Study 2 their κ is 0.81 (85%).

Because the unit of analysis is the sentence for both the QA system and the attitude classifiers, and the attitude classifiers are more general, identifying whether a sentence bears a sentiment or arguing attitude, we also measure sentence-level agreement for sentiment and arguing judgments. For each judge, we derive sentence-level sentiment and arguing judgments based on the lower-level attitudes that they marked in the sentence. In Study 1, we measure sentence-level agreement for the 263 sentences containing attitude annotations. In Study 2, there are 135 sentences with attitude annotations. Sentence-level sentiment agreement is κ 0.63 (83%) in Study 1 and κ 0.57 (86%) in Study 2. Arguing agreement is κ 0.68 (84%) in Study 1 and κ 0.64 (83%) in Study 2.

4. Attitude classification

In this section, we use the MPQA Corpus to train and evaluate classifiers for identifying sentence-level sentiment and arguing. The features we use are counts of different sets of instances of clues from a large lexicon, as well as bag-of-words features. Because not every instance of a clue is necessarily being used to express a sentiment or opinion, in the experiments, we also investigate the utility of first automatically disambiguating the clue instances before using them in sentence-level classification.

To disambiguate instances from the lexicon, we train two expression-level classifiers using the annotations in version 1.2 of the MPQA Corpus. The first classifier is a *sentiment-expression classifier*. The sentiment-expression classifier is trained using the contextual polarity annotations to determine if an instance from the lexicon is being used to express a sentiment. The second classifier is a *subjective-expression classifier*. The subjective-expression classifier is trained using the subjective expressions[5] marked in the MPQA Corpus, to determine whether an instance from the lexicon is subjective in context.

We use 10-fold cross validation to evaluate both the expression-level classifiers and the sentence-level attitude classifiers. The folds are created by first randomly assigning to the different folds the 4,499 sentences from the 284 documents with attitude annotation. Then the 5,788 sentences from the remaining 210 test documents are randomly assigned to folds. For the expression-level classifiers, all 10,287 sentences are used for the experiments. For the sentence-level attitude classifiers, only the subset with attitude annotations are used.

4.1 Lexicon

For the experiments in this section, we use the list of over 8,000 *subjectivity clues* made available by [30]. Subjectivity clues are words and phrases that may be used to express private states, i.e., they have subjective usages (though they may have objective usages as well). All the clues in the lexicon are single words.

Each word in the subjectivity lexicon is tagged with two pieces of information: its *reliability type* and its *prior polarity*. Words that are subjective in most contexts have the reliability type *strongly subjective* (*strongsubj*), and those that may only have certain subjective usages have the reliability type *weakly subjective* (*weaksubj*). Prior polarity captures whether a word out of context typically evokes something positive or something negative. The values for prior polarity are *positive*, *negative*, *both* and *neutral*.

For the experiments in this paper, we added one more piece of information to each clue in the lexicon: its *prior arguing polarity*. The prior arguing polarity takes on the values *positive*, *negative* or *neutral*. Arguing polarity is intended to capture whether a word out of context seems like it would be used to argue for or against something, or to argue that something is or is not true. Examples of words with positive arguing polarity are *accuse*, *must*, and *absolutely*. Examples of words with negative arguing polarity are *deny*, *impossible*, and *rather*.

4.2 Expression classifiers

The sentiment-expression classifier that we use is similar to the one presented in [30], but with a few changes to some features, which we found give slightly better results. The sentiment-classifier[6] in [30] uses 29 features to disambiguate each clue instance. The features represent information about the clue instance, information about the clue from the lexicon, information about how the clue instance relates to other word and clue instances in the surrounding context, information about the position of the clue instance in the sentence, information about counts of other clues in the current, previous, and next sentence, and information about the document

[4] During which the judges at times discussed their annotations.

[5] Wiebe et al. [29] define a subjective expressions as all expressive subjective element annotations and any direct subjective annotations with an *expression intensity* greater than neutral.

[6] Referred to as the neutral-polar classifier in [30].

accuse accusing accused accuses believe believing believed believes deny denying denied denies must should clearly cannot clear

Table 2: *High-precision arguing clues*

topic. For this work, the feature in [30] that captured context using the previous, current, and next word is replaced with two features representing the parts-of-speech of the previous and next word, respectively. Features that were previously counts were changed to be set-valued features, {0, 1, 2, 3}, where 3 represents a count of 3 or more. In addition, we added three new features that capture whether the clue instance is modifying, is being modified by, and is in a conjunction with another clue instance from the lexicon with a given polarity. The subjective-expression classifier was trained using the same features as the sentiment-expression classifier, with the exception of the three new polarity modification features.

Both expression-level classifiers were trained using BoosTexter AdaBoost.MH [24] with 5000 rounds of boosting. In 10-fold cross-validation experiments, the sentiment-expression classifier achieves an accuracy of 76.8% (sentiment recall 58.8%, precision 73.1%, and F-measure 65.1). A baseline classifier that uses just the prior polarity of the clue from the lexicon to classify each instance, has an accuracy of 52.6%. In cross-validation experiments, the subjective-expression classifier achieves an accuracy of 77.4% (subjective recall 80.1%, precision 80.2%, and F-measure 80.1). A baseline classifier that marks every strongsubj clue instance as subjective has an accuracy of 61.3%.

4.3 Sentence-level attitude classification

The sentence-level sentiment and arguing classifiers are binary classifiers. The sentiment classifiers judge only whether a sentence contains a sentiment, and the arguing classifiers similarly determine only whether a sentence is arguing.

For each sentence, the gold-standard sentiment and arguing classes are determined based on the attitude annotations in the sentence and their intensities. If a sentence contains an attitude annotation that is a *Positive Sentiment* or a *Negative Sentiment* with an intensity greater than *low*, then it is a sentiment sentence. The gold-standard class for arguing is determined in the same way.

We experiment both with classifiers trained using SVM^{light} (SVM) [7] and with rule-based (RB) classifiers. For the SVM classifiers, we use bag-of-words features and features that are counts of instances of different sets of clues from the lexicon. The following are the different classifiers:

Sentiment Classifiers: *SVM-BL*, *SVM-cluelex*, *SVM-clueauto*, *RB-cluelex*, *RB-clueauto*

Arguing Classifiers: *SVM-BL*, *SVM-cluelex*, *SVM-clueauto*, *RB-clue*

SVM-BL is a baseline classifier that uses as features all the words in a sentence (bag-of-words).

SVM-cluelex and SVM-clueauto classifiers use bag-of-words features plus four features representing counts of different sets of clues. For the sentiment classifiers, these four features are *strongsubj-sentiment count*, *strongsubj-neutral count*, *weaksubj-sentiment count*, *weaksubj-neutral count*. For the SVM-cluelex sentiment classifier, whether a clue instance is strongsubj/weaksubj or sentiment/neutral is determined based on the reliability class and prior polarity of the clue in the lexicon. Sentiment clues are those with positive, negative, or both prior polarity. For the SVM-clueauto sentiment classifier, whether an instance is sentiment or neutral is determined based on the output of the sentiment-expression classifier.

For arguing, the SVM-cluelex and SVM-clueauto classifiers use bag-of-words plus the features: *strongsubj-arguing count*, *strongsubj-neutral count*, *weaksubj-arguing count*, *weaksubj-neutral count*. For the SVM-cluelex arguing classifier, whether a clue instance is strongsubj/weaksubj or arguing/neutral is determined based on the reliability class and prior arguing polarity of the clue in the lexicon. For the SVM-clueauto arguing classifier, this information is also obtained from the lexicon, but only clue instances that are subjective according the subjective-expression classifier are used.

The rule-based classifiers are RB-cluelex, RB-clueauto, and RB-clue. RB-cluelex is a sentiment classifier that marks a sentence as a sentiment sentence if it contains one or more strongsubj, sentiment clue instances from the lexicon. RB-clueauto marks a sentence as a sentiment sentence if it contains one or more sentiment clue instances as identified by the output of the sentiment-expression classifier. RB-clue is an arguing classifier that marks a sentence as arguing if it contains one or more of the high-precision, arguing clues listed in Table 2.

Table 3 gives the results for the different sentiment and arguing classifiers. All results are averages over 10-fold cross-validation experiments. The results in bold significantly improve over the over the baseline (SVM-BL) for the given attitude classifier, as measured using a two-sided t-test ($p < 0.05$). For both sentiment and arguing, the best classifier is SVM-clueauto as measured by accuracy. This is the classifier that uses the output from the expression-level sentiment classifier (for sentiment) or the expression-level subjectivity classifier (for arguing) to disambiguate clue instances from the lexicon. Interestingly, the rule-based classifier RB-clueauto, which uses only output from the expression-level sentiment classifier, performs almost as well as the SVM-clueauto sentiment classifier in terms of accuracy, and slightly better in terms of sentiment F-measure.

The results in Table 3 also show that disambiguating clue instances from the lexicon is helpful for sentence-level attitude classification. Comparing the SVM-clueauto sentiment classifier to the SVM-cluelex sentiment classifier, the SVM-clueauto classifier gives a significantly higher accuracy ($p < 0.05$). The same is true for the RB-clueauto classifier as compared to the RB-cluelex classifier. For arguing, the SVM-clueauto classifier also performs better than the SVM-cluelex classifier, although the improvements are not significant.

5. Question answering experiments

To test the impact of attitude-type analysis on QA, we use a simple QA system, similar to the one used by Stoyanov et al. [26]. This QA system retrieves answer sentences based on keyword matching, which forms our baseline. The answers are then re-ranked based on the results of attitude-type analysis. For comparison, we also re-rank answers based on the more general, subjective/fact distinction used in [26].

As previously mentioned in Section 2, we use two datasets for these these experiments: the OpQA dataset and the HYS dataset. The OpQA dataset is annotated with answers to 15 opinion questions. Of these questions, four have fewer than five answers in the data. These were excluded from the experiments. The HYS data is annotated with answers to 13

Sentiment	Acc	Sent Rec	Sent Prec	Sent F	¬Sent Rec	¬Sent Prec	¬Sent F
(1) SVM-BL	75.2	57.2	75.7	65.2	87.4	75.0	80.7
(2) SVM-cluelex	**78.0**	**64.8**	77.1	**70.4**	86.9	**78.5**	82.5
(3) SVM-clueauto	**80.0**	**66.8**	**80.6**	**73.1**	89.0	**79.8**	**84.1**
(4) RB-cluelex	75.2	**68.5**	69.6	**69.1**	79.7	**78.9**	79.3
(5) RB-clueauto	**79.2**	**71.5**	75.8	**73.6**	84.5	**81.4**	**82.9**
Arguing	Acc	Arg Rec	Arg Prec	Arg F	¬Arg Rec	¬Arg Prec	¬Arg F
(1) SVM-BL	78.4	42.8	71.6	53.6	93.0	80.0	86.0
(2) SVM-cluelex	80.2	**49.2**	73.9	**59.0**	92.9	81.8	87.0
(3) SVM-clueauto	**80.7**	**51.4**	74.3	**60.8**	92.8	82.4	87.3
(4) RB-clue	76.0	21.7	**83.1**	34.4	**98.1**	75.5	85.3

Table 3: *Sentence-level attitude classification results*

opinion questions. Stoyanov et al. annotated the answers in the OpQA data [26]. We annotated answers in the HYS data using the same scheme and instructions. In the opinion QA task, there are multiple answers and some answers are more relevant than others. The annotators recorded the relevance of an answer by assigning it a number from 1 (less relevant) to 5 (most relevant).

For use in the experiments, we annotated the attitude types of the the 11 OpQA and 13 HYS questions. These questions and their attitude types are listed below (S and A denote Sentiment type and Arguing type respectively).

OpQA Questions
(1) Are the Japanese unanimous in their opinion of Bush's position on the Kyoto protocol: *S*
(2) How is Bush's decision not to ratify the Kyoto protocol looked upon by Japan and other US allies: *S*
(3) How do European Union countries feel about the US opposition to the Kyoto protocol: *S*
(4) How do the Chinese regard the human rights record of the United States: *S*
(5) What factors influence the way in which the US regards the human rights records of other nations: *A*
(6) Is the US annual human rights report received with universal approval around the world: *S*
(7) Did most Venezuelans support the 2002 coup: *S*
(8) How did ordinary Venezuelans feel about the 2002 coup and subsequent events: *S*
(9) Did America support the Venezuelan foreign policy followed by Chavez: *S*
(10) What was the American and British reaction to the reelection of Mugabe: *S*
(11) What is the basis for the European Union and US critical attitude and adversarial action toward Mugabe: *A*

HYS Questions
(1) Are the protests over the Muhammad cartoons justified: *A*
(2) What should the response be to the protests over the Muhammad cartoons: *A*
(3) Should the Muhammad cartoons have been published: *A*
(4) Are you worried about climate change: *S*
(5) Is enough being done to tackle climate change: *A*
(6) Should the Guantanamo Bay detention center be closed: *A*
(7) Do you agree with the UN stance on Guantanamo Bay closure: *A*
(8) Will the Palestinians be able to form a working government: *A*
(9) How should Israel deal with a Hamas-led Palestinian government: *A*
(10) Will the Hamas led Palestinian government negotiate with Israel: *A*
(11) What will be the effect of reporting Iran to the UN Security Council: *A*
(12) Should Iran be referred to the UN Security Council: *A*
(13) Who is to blame for the poor response to Katrina: *S,A*

5.1 QA system and answer re-ranking

Given a dataset and a question, the baseline QA system uses keyword matching to identify sentences that are potential answers to the question. The returned answers are then ranked based on confidence scores assigned by the baseline system. We perform several experiments, re-ranking the answers returned by the QA system by determining new scores using the results of the sentence-level attitude-type classifiers, subjective-sentence classifiers, and when possible, manual subjectivity and attitude-type annotations.

To re-rank the answers returned by the system, the baseline system confidence scores (*sc1*) are normalized to range between 0 and 1. Each answer sentence also receives an opinion/subjectivity score (*sc2*) based on the classifier used in the given experiment. The score *sc2* takes the value 0 or 1. $sc2 = 1$ if the question and answer type match; otherwise, $sc2 = 0$. According to this requirement, a system making a binary subjective/objective distinction matches subjective questions with subjective answers only. Similarly a system making an attitude type distinction has a non-zero score for *sc2* only when sentiment and arguing questions are matched with answers of the respective categories.

The *combined-score* for each answer sentence is a function of both the QA system confidence and the opinion score: $\textit{combined-score} = (\theta \times sc1) + ((1 - \theta) \times sc2)$ where θ (determined experimentally) is the fraction of the weight given to *sc1*. The optimal value for θ was found to be 0.7 for sentiment questions and 0.89 for arguing questions. A higher value of theta indicates keyword matching is more prominent in the *combined-score*.

Once the *combined-score* has been calculated for each answer, the answers are re-ranked based on this score. For our evaluations, the baseline system is the original output of the QA system before subjectivity/attitude re-ranking.

We experiment with the effect of re-ranking based on the output of six different sentence-level classifiers. Three of the classifiers are subjective-sentence classifiers. The *MAN-subj* classifier classifies a sentence as subjective or objective based on the manual MPQA annotations. The second classifier, *RB-subj*, is a high-precision rule-based subjective-sentence classifier [28]. The third classifier, *NB-subj* is a naive bayes subjective-sentence classifier [28]. These subjectivity classifiers are the ones used by Stoyanov et al. [26].

The remaining three classifiers are attitude-type classifiers. *MAN-att* is an attitude-type classifier based on the manual attitude annotations. *RB-att* is a rule-based attitude-type classifier. For sentiment classification, this is the *RB-clueauto* classifier from Section 4.3. For arguing classification, this is the high-precision (but low recall) classifier *RB-clue*. Finally,

	AvePrec	mSR	MRR	MRFA
Baseline	0.107	0.442	0.652	2.111
MAN-subj	0.119 p<0.01	0.472 p<0.07	0.676	1.889
MAN-att	0.128 p<0.01	0.488 p<0.05	0.685	1.778
RB-subj	0.121	0.489	0.759	1.556
NB-subj	0.110	0.442	0.657	2.000
RB-att	0.125 p<0.01	0.482 p<0.05	0.694	1.778
SVM-att	0.128 p<0.01	0.493 p<0.02	0.694	1.778

Table 4: *QA performance on sentiment questions in the OpQA corpus. t-tests were carried out for AvePrec and mSR. For those that are significant over baseline, p-values are given*

the *SVM-att* classifier is the *SVM-clueauto* classifier from Section 4.3 for both sentiment and arguing.

5.2 Evaluation metrics

As mentioned earlier, our QA task has multiple correct answers. Hence, in addition to the popular QA evaluation metrics like MRR and MRFA that evaluate only the first correct answer, we also evaluate the performance of the QA system using Average Precision (AvgPrec) and modified Sliding Ratio (mSR). Other researchers (e.g., [15]) have also reported the inadequacy of MRR as a performance metric for QA systems that retrieve more than one correct answer.

Average Precision: We use the definition of Average Precision from [3]: "The mean of the precision scores obtained after each relevant document is retrieved, using zero as the precision for relevant documents that are not retrieved." Suppose there are two QA systems that retrieve n answers, of which exactly m answers are relevant. Using the above definition, the system that assigns a better rank (on average) to the m correct answers gets a better Average Precision score.

Sliding Ratio: We use the modified sliding ratio (mSR) as defined by [23, 9]. This metric is more sensitive to the ranking of answers than the original sliding ratio proposed by [17], and it takes into account the ordering of the answers based on multi-grade relevance. As mentioned previously, our answer annotations capture 5 levels of relevance. This multi-grade relevance is captured by the mSR.

5.3 Evaluation results

In this section, we present the results of our re-ranking experiments. For each experiment, we re-rank the top 100 answers returned by the baseline system. Re-ranking of the top 400 answers yields similar results and conclusions.

Table 4 gives the results for the sentiment questions in the OpQA dataset. In general, re-ranking based on sentiment produces a larger improvement over the baseline than re-ranking based on subjectivity alone, for all evaluation metrics. The improvements attained by incorporating sentiment information are statistically significant over the baseline for both AvePrec and mSR. These results suggest that the sentiment category distinction is useful for opinion QA.

Table 5 shows the performance of re-ranking based on output of the different classifiers for arguing questions in the OpQA corpus. Re-ranking based on the manual annotations performs better than the baseline for all metrics. Furthermore, re-ranking based on the manual arguing annotations outperforms re-ranking based on the subjectivity annotations across all metrics. This indicates that making the arguing distinction is useful for opinion QA. We did not perform significance tests for these results because there are only two

	AvePrec	mSR	MRR	MRFA
Baseline	0.045	0.170	0.100	7.5
MAN-subj	0.057	0.219	0.125	5.5
MAN-att	0.072	0.296	0.125	5.5
RB-subj	0.048	0.227	0.125	5.5
NB-subj	0.057	0.216	0.100	6.5
RB-att	0.062	0.336	0.125	8.5
SVM-att	0.033	0.183	0.125	9.5

Table 5: *QA performance on arguing questions in OpQA corpus*

	AvePrec	mSR	MRR	MRFA
Baseline	0.059	0.342	0.50	13.8
RB-subj	0.059	0.335	0.46	13.6
NB-subj	0.059	0.338	0.46	13.3
RB-att	0.059	0.345	0.50	13.9
SVM-att	0.061 p<0.1	0.357 p<0.1	0.50	16.5

Table 6: *QA performance on arguing questions in the HYS dataset*

questions. As for re-ranking based on the automatic arguing classifiers, only *RB-att* shows improvements over the baseline.

Table 6 shows the performance of the re-ranking experiments on the HYS dataset for the 12 arguing questions[7]. Unlike the results for the OpQA dataset, re-ranking based on subjectivity alone does not help the performance for any of the metrics. On the other hand re-ranking the answers based on the SVM classifier gives an improvement over the baseline for Average Precision as well as mSR. However, none of the conclusions in the arguing category are statistically significant. These results suggest that the arguing category is inherently difficult. We discuss this in detail in the Section 5.4. Evaluation results for the two HYS sentiment questions were similar to those for the sentiment question on the OpQA dataset; they are omitted due to space restrictions.

Using the sentiment questions from the OpQA dataset and the arguing questions from the HYS dataset, we conducted some further analysis, investigating the effect of different values of θ on QA performance. All QA systems in Table 4 with $p < 0.01$ for Average Precision (i.e., MAN-subj, MAN-att, RB-att, and SVM-att) were rerun for the same (sentiment) questions with $\theta = 0.5$. The resulting QA performance, evaluated using AvePrec, mSR, MRR and MRFA, was similar to the that shown in Table 4, i.e., each of the systems performed better than the baseline and also maintained their relative ranking with respect to each other. This shows that for sentiment questions, any particular value of θ does not favor one sentiment/ subjectivity system over the other. On the other hand, for the arguing questions, all systems performed worse than the baseline for this value of θ. The best value of θ for arguing questions, as we have seen before, is large. We believe this is because the arguing category is complex resulting in low classifier accuracy. Consequently all the systems rely heavily on keyword matching to get a good performance.

In order to test whether the distinction between attitude types is indeed helpful, we measured the performance (Average Precision) of the QA system when questions of one attitude type were matched with answers of a different attitude type, or a broader category. Table 7 shows the results of these experiments. In the table, "Attitude" is the broader

[7] The HYS dataset does not contain subjectivity or attitude annotations, so there are no results for the MAN-subj or MAN-att classifiers for these questions.

	Sentiment Questions $\theta = 0.5, 0.8$	Arguing Questions $\theta = 0.8$
	Matched with answer types	
Rank-1	Sentiment*	Arguing*
Rank-2	Attitude*	Attitude*
Rank-3	Arguing	Sentiment

Table 7: *Relative performance of QA systems (Average Precision) when questions of one category were matched with answers of the same or different attitude, or a broader category. * indicates QA performance was above baseline*

category that denotes "either sentiment or arguing." For this experiment, we used the output of the best attitude classifier (SVM-att) for both sentiment and arguing.

The systems with attitude type mismatch (sentiment questions matched with answers classified as arguing and vice versa) exhibit performance below the baseline. The highest ranked system is the one in which sentiment questions are matched with sentiment-type answers and arguing questions are matched with arguing-type answers. Not surprisingly, the broad "Attitude" category, which includes answers of both attitude types, is ranked in the middle. These results reinforce our hypothesis that matching the attitude types of questions and answers helps QA performance.

5.4 Discussion

Automatic detection of the arguing attitude type is difficult in general and particularly complicated for QA. We discuss some of the complexities of this category below:

One can argue against something by arguing for, or suggesting something opposite. In an answer to the question, *Are the protest against the Muhammad cartoons justified?*, a writer states, "Those insulted by these cartoons are free to ignore them." In this statement, the writer uses the expression, "are free to ignore them," to suggest an alternative to protesting, and in this way argues against the protests.

Rhetoric is used to argue. People often use rhetoric, sarcasm, and humor to make their point. For example, consider the sentence, "Can anybody tell me why the rules against drawing Mohammed should apply to people who don't believe in Islam?", where the writer is arguing that it is alright for a non-believer to have drawn (and hence published) the picture of Mohammed. However, the argument is enveloped in a rhetorical question.

Arguing answers may be indirect and implied, hence difficult to connect to the question. Consider the following answer to the question, *Should the Muhammad cartoons have been published?*: "Because we have the legal right to freely express our opinions does not justify using these rights to deliberately hurt others." This sentence, conveys, quite easily to the reader, the writer's belief that it was wrong to publish the Muhammad cartoons. However, this meaning is conveyed only in the light of the discourse when the topic under discussion is made clear. Inference is needed to understand that the phrase "legal right to freely express our opinions" is referring to the newspaper's freedom to publish. Further, the speaker uses the expression "deliberately hurt others" to convey his belief that the publication was uncalled for. It requires world knowledge to infer that the cartoons could have hurt someone's feelings.

These complexities of the arguing type give us insight into the low performance of the arguing classifiers and the QA system for arguing questions. However, the analysis of Table 7 shows that making distinction between attitude types is indeed important. Even though the performance of the (SVM) arguing classifier was not at par with the performance of the (SVM) sentiment classifier, matching the arguing type questions with answers classified as arguing yielded better QA performance.

6. Related work

In recent years, complex QA systems, such as HITIQA [25], have ventured to answer analytical exploratory questions like *What has been Russia's reaction to U.S. bombing of Kosovo?* Such questions characteristically do not have one correct answer. The QA system interacts with the user to expand or reduce the answer space. Knowledge of sentiment and arguing (Russia's sentiment versus what Russia argues should be done) would enable a system to ask the user to choose from "Do you want to know about how the people of Russia feel about the US bombing of Kosovo" or "Do you want to know what the people think Russia should do about the US bombing of Kosovo." Lita et al. [11] call for using a sentiment dimension for definitional QA. They state that answers could be enhanced by adding information on how entities are regarded by different sources. Yu and Hatzivassiloglou [31] and Stoyanov et al. [26] motivate using subjectivity analysis to improve QA. Stoyanov et al. show that subjectivity filtering improves MRR and MRFA for opinion QA. In this work, we use re-ranking instead of filtering, and we consider a finer-grained distinction, whether giving a QA system information about sentiment and arguing attitude types is useful.

Re-ranking of pre-selected answers in QA has been used by a number of researchers (e.g., [1, 20, 21]) to improve the performance of QA Systems. These systems re-score the pre-selected answers based on additional processing and heuristics. Previous work on QA re-ranking [15] has reported that a direct match between questions and answers is an important component in the answer score. This is evidenced in our system too as our best reported system results are obtained when we assign a large weight to the keyword match (*sc1*).

The research most closely related to our work on recognizing sentences bearing sentiments and arguing/beliefs is the work on sentence-level subjectivity and sentiment analysis (e.g., [22, 31, 14, 8, 6, 18, 5]). Riloff and Wiebe [22], Yu and Hatzivassiloglou [31], and Kudo and Matsumoto [10] train classifiers to discriminate between subjective and objective sentences. Our work differs from theirs in that we seek to recognize sentences bearing different types of subjective attitudes. Yu and Hatzivassiloglou [31], Nasukawa and Yi [14], Kim and Hovy [8], Hu and Liu [6], Kudo and Matsumoto [10], Popescu and Etzioni [18], and Gamon et al. [5] identify sentences expressing positive and negative sentiments. We also seek to identify sentences where sentiments are expressed; however, this work does not focus on further discriminating positive and negative sentences. We anticipate that for the task of question answering, identifying positive and negative sentiments and arguing will be important for later stages of processing, for example, creating clusters of positive and negative answers to present to the user of the QA system. Yu and Hatzivassiloglou [31], Nasukawa and Yi [14], Kim and Hovy [8], and Hu and Liu [6] classify sentiment sentences by aggregating information about words from a lexicon. We also rely on lexicon information for our sentence classification, but we take the novel approach of first disambiguating the instances

from the lexicon to determine which in context are actually being used to express sentiments or subjectivity. To the best of our knowledge, this is the first work to automatically identify sentences bearing arguing attitudes.

7. Conclusions

In this paper, we explored the utility of attitude types for improving opinion question answering (QA) on both web-based discussions and news data. We presented a set of attitude types developed with an eye toward QA and showed that they can be reliably annotated. Using the attitude annotations, we developed automatic classifiers for recognizing when a sentence is expressing either of two main types of attitudes: sentiment or arguing. The best classifiers performed significantly better than the baselines. These experiments also showed that disambiguating instances of subjectivity clues is useful for sentence-level attitude-type classification. In our question answering experiments, we used information about the attitude type of questions and answers, provided by the manual attitude annotations and the automatic sentence classifiers, to re-rank answers retrieved by the QA system. By trying to match the attitude type of questions and answers, we achieved better performance for our QA system than the baseline system or a system re-ranking answers based on a more general subjective/objective distinction.

References

[1] S. Abney, M. Collins, and A. Singhal. Answer extraction. In *ANLP-2000*, 2000.

[2] K. Balog, G. Mishne, and M. de Rijke. Why are they excited? identifying and explaining spikes in blog mood levels. In *11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL 2006)*, 2006.

[3] C. Buckley and E. M. Voorhees. Evaluating evaluation measure stability. In *SIGIR-2000*, 2000.

[4] K. Dave, S. Lawrence, and D. M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *WWW-2003*, 2003.

[5] M. Gamon, A. Aue, S. Corston-Oliver, and E. Ringger. Pulse: Mining customer opinions from free text. In *Intelligent Data Analysis*, 2005.

[6] M. Hu and B. Liu. Mining and summarizing customer reviews. In *KDD-2004*, 2004.

[7] T. Joachims. Text categorization with support vector machines: learning with many relevant features. In *ECML-98*, 1998.

[8] S.-M. Kim and E. Hovy. Determining the sentiment of opinions. In *Coling 2004*, 2004.

[9] K. Kishida. Property of average precision and its generalization: An examination of evaluation indicator for information retrieval experiments. Nii technical report (nii-2005-014e), NII, 2005.

[10] T. Kudo and Y. Matsumoto. A boosting algorithm for classification of semi-structured text. In *EMNLP-2004*, 2004.

[11] L. V. Lita, A. H. Schlaikjer, W. Hong, and E. Nyberg. Qualitative dimensions in question answering: Extending the definitional qa task. In *AAAI-2005*, 2005.

[12] G. Mishne. Multiple ranking strategies for opinion retrieval in blogs. In *TREC 2006*, 2006.

[13] G. Mishne and N. Glance. Predicting movie sales from blogger sentiment. In *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW 2006)*, 2006.

[14] T. Nasukawa and J. Yi. Sentiment analysis: Capturing favorability using natural language processing. In *K-CAP 2003*, 2003.

[15] J. Otterbacher, G. Erkan, and D. R. Radev. Using random walks for question-focused sentence retrieval. In *HLT/EMNLP-2005*, 2005.

[16] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In *EMNLP-2002*, 2002.

[17] S. M. Pollock. Measures for the comparison of information retrieval systems. *American Documentation*, 19(4), 1968.

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

[19] R. Quirk, S. Greenbaum, G. Leech, and J. Svartvik. *A Comprehensive Grammar of the English Language.* Longman, New York, 1985.

[20] D. Radev, W. Fan, H. Qi, H. Wu, and A. Grewal. Probabilistic question answering on the web. In *WWW-2002*, 2002.

[21] G. Ramakrishnan, S. Chakrabarti, D. Paranjpe, and P. Bhattacharyya. Is question answering an acquired skill? In *WWW-2004*, 2004.

[22] E. Riloff and J. Wiebe. Learning extraction patterns for subjective expressions. In *EMNLP-2003*, 2003.

[23] Y. Sagara. Performance measures for ranked output retrieval systems. *Journal of Japan society of Information and knowledge*, 12(2), 2002.

[24] R. E. Schapire and Y. Singer. BoosTexter: A boosting-based system for text categorization. *Machine Learning*, 39(2/3):135–168, 2000.

[25] S. Small, T. Strzalkowski, T. Liu, S. Ryan, and R. Salkin. Hitiqa: Towards analytical question answering. In *Coling-2004*, 2004.

[26] V. Stoyanov, C. Cardie, and J. Wiebe. Multi-perspective question answering using the opqa corpus. In *HLT/EMNLP 2005*, 2005.

[27] P. Turney. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *ACL-2002*, 2002.

[28] J. Wiebe and E. Riloff. Creating subjective and objective sentence classifiers from unannotated texts. In *CICLing-2005*, 2005.

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

[30] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT/EMNLP 2005*, 2005.

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