Towards a Unified Approach for Opinion Question Answering and Summarization

Elena Lloret and **Alexandra Balahur** and **Manuel Palomar** and **Andrés Montoyo**
Department of Software and Computing Systems
University of Alicante
Alicante 03690, Spain
{elloret,abalahur, mpalomar, montoyo}@dlsi.ua.es

Abstract

The aim of this paper is to present an approach to tackle the task of opinion question answering and text summarization. Following the guidelines TAC 2008 Opinion Summarization Pilot task, we propose new methods for each of the major components of the process. In particular, for the information retrieval, opinion mining and summarization stages. The performance obtained improves with respect to the state of the art by approximately 12.50%, thus concluding that the suggested approaches for these three components are adequate.

1 Introduction

Since the birth of the Social Web, users play a crucial role in the content appearing on the Internet. With this type of content increasing at an exponential rate, the field of Opinion Mining (OM) becomes essential for analyzing and classifying the sentiment found in texts.

Nevertheless, real-world applications of OM often require more than an opinion mining component. On the one hand, an application should allow a user to query about opinions in natural language. Therefore, Question Answering (QA) techniques must be applied in order to determine the information required by the user and subsequently retrieve and analyze it. On the other hand, opinion mining offers mechanisms to automatically detect and classify sentiments in texts, overcoming the issue given by the high volume of such information present on the Internet. However, in many cases, even the result of the opinion processing by an automatic system still contains large quantities of information, which are still difficult to deal with manually. For example, for questions such as “Why do people like George Clooney?” we can find thousands of answers on the Web. Therefore, finding the relevant opinions expressed on George Clooney, classifying them and filtering only the positive opinions is not helpful enough for the user. He/she will still have to sift through thousands of texts snippets, containing relevant, but also much redundant information. For that, we need to use Text Summarization (TS) techniques. TS provides a condensed version of one or several documents (i.e., a summary) which can be used as a substitute of the original ones (Spärck Jones, 2007). In this paper, we will concentrate on proposing adequate solutions to tackle the issue of opinion question answering and summarization. Specifically, we will propose methods to improve the task of question answering and summarization over opinionated data, as defined in the TAC 2008 “Opinion Summarization pilot”[1]. Given the performance improvements obtained, we conclude that the approaches we proposed for these three components are adequate.

2 Related Work

Research focused on building factoid QA systems has a long tradition, however, it is only recently that studies have started to focus on the creation and development of opinion QA systems. Example of this can be (Stoyanov et al., 2004) who took advantage of opinion summarization to support Multi-Perspective QA system, aiming at extracting opinion-oriented information of a question. (Yu and Hatzivassiloglou, 2003) separated opinions from facts and summarized them as answer to opinion questions. Apart from these studies, specialized competitions for systems dealing with opinion retrieval and QA have been organized in the past few years. The TAC 2008 Opinion Summarization Pilot track proposed a mixed setting of factoid and opinion questions.

[1] http://www.nist.gov/tac/2008/summarization/

Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, ACL-HLT 2011, pages 168–174,
24 June, 2011, Portland, Oregon, USA ©2011 Association for Computational Linguistics

It is interesting to note that most of the participating systems only adapted their factual QA systems to overcome the newly introduced difficulties related to opinion mining and polarity classification. Other relevant competition focused on the treatment of subjective data is the NTCIR MOAT (Multilingual Opinion Analysis Test Collection). The approaches taken by the participants in this task are relevant to the process of opinion retrieval, which is the first step performed by an opinion mining question answering system. For example, (Taras Zabibalov, 2008) used an almost unsupervised approach applied to two of the sub-tasks: opinionated sentence and topic relevance detection.(Qu et al., 2008) applied a sequential tagging approach at the token level and used the learned token labels in the sentence level classification task and their formal run submission was is trained on MPQA (Wiebe et al., 2005).

3 Text Analysis Conferences

In 2008, the *Opinion Summarization Pilot* task at the Text Analysis Conferences[2] (TAC) consisted in generating summaries from blogs, according to specific opinion questions provided by the TAC organizers. Given a set of blogs from the Blog06 collection[3] and a list of questions, participants had to produce a summary that answered these questions. The questions generally required determining opinion expressed on a target, each of which dealt with a single topic (e.g. George Clooney). Additionally, a set of text snippets were also provided, which contained the answers to the questions. Table 1 depicts an example of target, question, and optional snippet.

Target:	George Clooney
Questions:	Why do people like George Clooney? Why do people dislike George Clooney?
Snippets:	1050 BLOG06-20060209-006-0013539097 he's a great actor.

Table 1: Example of target, question, and snippet.

Following the results obtained in the evaluation at TAC 2008 (Balahur et al., 2008), we propose an opinion question answering and summarization (OQA&S) approach, which is described in detail in the following sections.

4 An Opinion Question Answering and Summarization Approach

In order to improve the results of the OQA&S system presented at TAC, we propose new methods for each of the major components of the system: information retrieval, opinion mining and text summarization.

4.1 Opinion Question Answering and Summarization Components

- **Information Retrieval**

  JAVA Information Retrieval system (JIRS) is a IR system especially suited for QA tasks (Gómez, 2007). Its purpose is to find fragments of text (passages) with more probability of containing the answer to a user question made in natural language instead of finding relevant documents for a query. To that end, JIRS uses the own question structure and tries to find an equal or similar expression in the documents. The more similar the structure between the question and the passage is, the higher the passage relevance.

  JIRS is able to find question structures in a large document collection quickly and efficiently using different n-gram models. Subsequently, each passage is assessed depending on the extracted n-grams, the weight of these n-grams, and the relative distance between them. Finally, it is worth noting that the number of passages in JIRS is configurable, and in this research we are going to experiment with passages of length 1 and 3.

- **Opinion Mining**

  The first step we took in our approach was to determine the opinionated sentences, assign each of them a polarity (positive or negative) and a numerical value corresponding to the polarity strength (the higher the negative score, the more negative the sentence and vice versa). In our first approximation (OMaprox1), we employed a simple, yet efficient method, presented in Balahur et al. (Balahur et al., 2009). As lexicons for affect detection, we used WordNet Affect (Strapparava and Valitutti, 2004), SentiWordNet (Esuli and Sebas-

[2] www.nist.gov/tac/

[3] http://ir.dcs.gla.ac.uk/test_collections/access_to_data.html

tiani, 2006), and MicroWNOp (Cerini et al., 2007). Each of the resources we employed were mapped to four categories, which were given different scores: positive (1), negative (-1), high positive (4) and high negative (-4). First, the score of each of the blog posts was computed as the sum of the values of the words that were identified. Subsequently, we performed sentence splitting[4] and classified the sentences we thus obtained according to their polarity, by adding the individual scores of the affective words identified.

In the second approach (OMaprox2), we first filter out the sentences that are associated to the topic discussed, using LSA. Further on, we score the sentences identified as relating to the topic of the blog post, in the same manner as in the previous approach. The aim of this approach is to select for further processing only the sentences which contain opinions on the post topic. In order to filter these sentences in, we first create a small corpus of blog posts on each of the topics included in our collection[5]. For each of the corpora obtained, we apply LSA, using the Infomap NLP Software[6]. Subsequently, we compute the 100 most associated words with two of the terms that are most associated with each of the topics and the 100 most associated words with the topic word. The approach was proven to be successful in (Balahur et al., 2010).

- **Text Summarization**

  The text summarization approach used in this paper was presented in (Lloret and Palomar, 2009). In order to generate a summary, the suggested approach first carries out a basic preprocessing stage comprising HTML parsing, sentence segmentation, tokenization, and stemming. Once the input document or documents have been pre-processed, a relevance detection stage, which is the core part of the approach, is applied. The objective of this step is to identify potential relevant sentences in the document by means of three techniques: textual entailment, term frequency and the code quantity principle (Givón, 1990). Then, each potential relevant sentence is given a score which is computed on the basis of the aforementioned techniques. Finally, all sentences are ordered according to their scores, and the highest ranked ones (which mean those sentences contain more important information) are selected and extracted up to the desired length, thus building the final summary. It is worth stressing upon the fact that in an attempt to maintain the coherence of the original documents, sentences are shown in the same order they appear in the original documents.

4.2 Experimental Framework

The objective of this section is to describe the corpus used and the experiments performed with the data provided in TAC 2008 *Opinion Summarization Pilot*[7] task. The approaches analyzed comprise:

- **OQA&S:** The three components explained in the previous section (information retrieval, opinion mining and summarization) were bound together in order to produce summaries that include the answer to opinionated questions. First, the most relevant passages of length 1 and 3 are retrieved by the IR module, as in the aforementioned approach, and then the subjective information is found and classified within them using the OM approaches described in the previous section. Further on, we incorporate the TS module, to select and extract the most relevant opinionated facts from the pool of subjective information identified by the OM module. We generate opinion-oriented summaries of compression rates ranging from 10% to 50%. In the end, four different approaches result from the integration of the three components: ***IRp1-OMaprox1-TS***; ***IRp1-OMaprox2-TS***; ***IRp3-OMaprox1-TS***; and ***IRp3-OMaprox2-TS***.

Moreover, apart from these approaches, two baselines were also defined. On the one hand, we sug-

[4] http://alias-i.com/lingpipe/

[5] These small corpora (30 posts for each of the topics) are gathered using the search on topic words on http://www.blogniscient.com/ and crawling the resulting pages.

[6] http://infomap-nlp.sourceforge.net/

[7] http://www.nist.gov/tac/data/past-blog06/2008/OpSummQA08.html#OpSumm

gest a baseline using the list of snippets provided by the TAC organization (**QA-snippets**). This baseline produces a summary by joining all the answers in the snippets that related to the same topic On the other hand, we took as a second baseline the approach from our participation in TAC 2008 (**DLSIUAES**), without not taking into account any information retrieval or question answering system to retrieve the fragments of information which may be relevant to the query. In contrast, this was performed by computing the cosine similarity[8] between each sentence in the blog and the query. After all the potential relevant sentences for the query were identified, they were classified in terms of subjectivity and polarity, and the most relevant ones were selected for the final summary.

4.3 Evaluation Methodology

Since we used the corpus provided at the *Opinion Summarization Pilot* task, and we followed similar guidelines, we should evaluate our OQA&S approach in the same way as participant systems were assessed. However, the evaluation methodology proposed differs slightly from the one carried out in the competition. The reason why we took such decision was due to the fact that the evaluation carried out in TAC had some limitations, and therefore was not suitable for our purposes. In this manner, our evaluation is also based on the gold-standard nuggets provided by TAC, but in addition we proposed an extended version of them, by adding other pieces of information that are also relevant to the topics.

In this section, all the issues concerning the evaluation are explained. These comprise the original evaluation method used in the Opinion Summarization Pilot task at TAC (Section 4.3.1) , its drawbacks (Section 4.3.2), and the extended version for the evaluation method we propose (Section 4.3.3). Further on, the results obtained together with a wide discussion, as well as its comparison with the baselines and the TAC participants is provided in Section 4.4.

4.3.1 Nugget-based Evaluation at TAC

Within the *Opinion Summarization Pilot* task, each summary was evaluated according to its content using the Pyramid method (Nenkova et al., 2007). A list of nuggets was provided and the assessors used such list of nuggets to count the number of nuggets a summary contained. Depending on the number of nuggets the summary included and the importance of each one given by their weight, the values for recall, precision and F-measure were obtained. An example of several nuggets corresponding to different topics can be seen in Table 2, where the weight for each one is also shown in brackets.

Topic	Nugget (weight)
Carmax	CARMAX prices are firm, the price is the price (0.9)
Jiffy Lube	They should have torque wrenches (0.2)
Talk show hosts	Funny (0.78)

Table 2: Example of evaluation nuggets and associated weights.

4.3.2 Limitations of the Nugget Evaluation

The evaluation method suggested at TAC requires a lot of human effort when it comes to identify the relevant fragments of information (nuggets) and compute how many of them a summary contains, resulting in a very costly and time-consuming task. This is a general problem associated to the evaluation of summaries, which makes the task of summarization evaluation especially hard and difficult.

But, apart from this, when an exhaustive examination of the nuggets used in TAC is done, some other problems arised which are worth mentioning. The average number of nuggets for each topic is 27, and this would mean, that longer summaries will be highly penalized, because it will contain more useless information according to the nuggets. After analyzing in detail all the provided nuggets, we mainly classified the possible problems into six groups, which are:

1. **Some of the nuggets were expressed differently from how they appeared in the original blogs.** Since most of the summarization systems are extractive, this fact forced that humans had to evaluate the summaries, otherwise it would be very difficult to account for the presence of such nugget in the summary, if they are not using the same vocabulary as the original blogs.

2. **Some nuggets for the same topic express the**

[8] http://www.d.umn.edu/ tpederse/text-similarity.html

same idea, despite not being identical. In these cases, we are counting a single piece of information in the summary twice, if the idea that nuggets expressed is included.

3. Moreover, **the meaning of one nugget can be deduced from another's**, which is also related to the problem stated before.

4. **Some of the nuggets are not very clear in meaning** (e.g. *"hot"*, *"fun"*). This would mean that a summary might include such terms in a different context, thus, obtaining incorrectly that it is revelant when might be out of context.

5. **A sentence in the original blog can be covered by several nuggets**. For instance, both nuggets *"it is an honest book"* and *"it is a great book"* correspond to the same sentence *"It was such a great book-honest and hard to read (content not language difficulty)"*. In this case, it is not clear how to proceed with the evaluation; whether to count both nuggets or just one of them.

6. **Some information which is also relevant for the topic is not present in any nugget**. For instance: *"I go to Starbucks because they generally provide me better service"*. Although it is relevant with respect to the topic and it appears in a number of summaries, it would be not counted because it has not been chosen as a nugget.

4.3.3 Extended Nugget-based Evaluation

Since we are interested in testing a wide range of approaches involving IR, OM and TS, sticking to the rules to the original TAC evaluation would mean that a lot of time as well as human effort will be required, as well as not accounting for important information that summaries may contain in addition to the one expressed by the nuggets. Therefore, taking as a basis the nuggets provided at TAC, we set out a modified version of them.

The underlying idea behind this is to create an extended set of nuggets that serve as a reference for assessing the content of the summaries. In this manner, we will map each original nugget with the set of sentences in the original blogs that are most similar to it, thus generating a gold-standard summary for each topic. For creating this extended gold-standard nuggets we compute the cosine similarity[9] between every nugget and all the sentences in the blog related to the same topic. We empirically established a similarity threshold of 0.5, meaning that if a sentence was equal or above such similarity value, it will be considered also relevant. One main disadvantage of such a lower threshold value is that we can consider relevant sentences that share the same vocabulary but in fact they are not relevant to the summary. In order to avoid this, once we had identified all the most similar sentences to each nugget, we carried out a manual analysis to discard cases like this. Having created the extended set of nuggets, we grouped all of them pertaining to the same topic, and considered it a gold-standard summary. Now, the average number of nuggets per topic is 53, which we have increased by twice the number of original nuggets provided at TAC.

Further on, our summaries are compared against this new gold-standard using ROUGE (Lin, 2004). This tool computes the number of different kinds of overlap n-grams between an automatic summary and a human-made summary. For our evaluation, we compute ROUGE-1 (unigrams), ROUGE-2 (bigrams), ROUGE-SU4 (it measures the overlap of skip-bigrams between a candidate summary and a set of reference summaries with a maximum skip distance of 4), and ROUGE-L (Longest Common Subsequence between two texts). The results and discussion are next provided.

4.4 Results and Discussion

This section contains the results obtained for our OQA&S approach and all the sub-approaches tested. IRp*N* refers to the length of the passage employed in the information retrieval approach, whereas OMaprox*N* indicates the approach used for the opinion mining component. Firstly, we show and analyze the results of our different approaches, and then we compared the best performing one with the baselines and the average *Opinion Summarization Pilot* task participants results in TAC.

Table 3 shows the precision (Pre), recall (Rec) and F-measure results of ROUGE-1 (R-1) for all the approaches we experimented with.

Generally speaking, the results obtained show better figures for precision than for recall, and there-

[9] The cosine similarity was computed using Pedersen's Text Similarity Package: http://www.d.umn.edu/ tpederse/text-similarity.html

Approach		**Summary length**				
Name	**R-1**	**10%**	**20%**	**30%**	**40%**	**50%**
IRp1 -OMaprox1-TS	Pre	24.29	26.17	29.73	30.82	32.54
	Rec	14.45	18.58	22.32	23.63	26.32
	$F_{\beta=1}$	16.53	20.65	24.58	25.75	28.12
IRp1 -OMaprox2-TS	Pre	24.29	26.17	29.73	30.82	32.54
	Rec	16.90	20.02	23.36	24.15	26.77
	$F_{\beta=1}$	19.45	22.13	25.36	25.94	28.40
IRp3 -OMaprox1-TS	Pre	27.27	30.18	30.91	30.05	30.19
	Rec	20.56	24.76	28.25	31.67	34.47
	$F_{\beta=1}$	22.65	26.23	27.98	29.18	29.74
IRp3 -OMaprox2-TS	Pre	30.16	32.11	32.35	32.41	32.11
	Rec	20.64	24.03	27.25	29.78	32.68
	$F_{\beta=1}$	23.28	25.64	27.42	28.44	29.21

Table 3: Results of our OQA&S approaches

Approach		**Performance (ROUGE)**			
Name	**%**	**R-1**	**R-2**	**R-L**	**R-SU4**
IRp3-OMaprox2 -TS (50%)	Pre	32.11	7.34	29.00	11.37
	Rec	32.68	8.31	33.24	12.76
	$F_{\beta=1}$	29.21	7.22	28.60	11.13
QA-snippets	Pre	17.97	8.76	17.65	9.98
	Rec	71.24	31.30	70.10	37.44
	$F_{\beta=1}$	24.73	11.58	24.29	13.45
DLSIUAES	Pre	20.54	7.00	19.46	9.29
	Rec	57.66	18.98	54.61	25.77
	$F_{\beta=1}$	27.04	9.10	25.59	12.22
Average TAC participants	Pre	23.74	8.35	22.72	10.81
	Rec	56.65	19.37	54.56	25.40
	$F_{\beta=1}$	27.45	9.64	26.33	12.46
Average TAC participants'	Pre	20.42	6.06	19.55	8.62
	Rec	56.45	17.3	54.40	24.11
	$F_{\beta=1}$	24.31	7.25	23.31	10.29

Table 4: Comparison with other systems

fore the F-measure value, which combines both values, will be affected. Good precision values means that the information our approaches select is the correct one, despite not including all the relevant information.

Our best performing approach in general is the one which uses a length passage of 3 and, as far as OM is concerned, when topic-sentiment analysis is carried out (*IRp3-OMaprox2-TS*). This shows that the approach dealing with topic-sentiment analysis in opinion mining is more suitable than the one which does not consider topic relevance. Taking a look at some individual results, we next try to elucidate the reasons why our approach performs better at some approaches and not so good at others. Concerning the IR module, it is important to mention that a passage length of 1 always obtains poorer results that when it is increased to 3, meaning that the longer the passage, the better.

Regarding the best summary length, we observed that in general terms, the more content we allow for the summary, the better. In other words, compression rates of 50% get higher results than 20% or 10%. However, there are cases in which shorter summaries (10% and 20%) obtains better results than longer ones (e.g. *IRp3-OMaprox2-TS* vs. *IRp3-OMaprox1-TS*).

Although the results theirselves are not very high (around 30%), they are in line with the state-of-the-art, as can be seen in Table 4, where our best performing approach is compared with respect to other approaches.

Although the compression rate which obtains best results is not very high (50%), indeed the final summaries have an average length of 2,333 non-white space characters. This is really low compared to the length that TAC organization allowed for the Opinion Summarization Pilot task, which was 7,000 non-white space characters per question, and most of the times there were two questions for each topic. Whereas the results of TAC participants are much better for the recall value than ours, if we take a look at the precision, our approach outperforms them according to this value in all of the cases. The longer a summary is, the more chances it has to contain information related to the topic. However, not all this information may be relevant, as it is shown in the results for the precision values, which decrease considerably compared to the recall ones. In contrast, due to the fact that our approach is missing some relevant information because we use a rather short passage length (3 sentences), we do not obtain such high values for the recall, but we obtain good precision results, which indicate that the information that we keep is important.

Moreover, comparing those results with the ones obtained by our approach, it is worth mentioning that *IRp3-OMaprox2-TS* outperforms the F-measure value for all the ROUGE metrics with respect to *Average TAC participants'*. More in detail, when the ROUGE scores are averaged, *IRp3-OMaprox2-TS* improves by 12.50% the *Average TAC participants'* for the F-measure value.

5 Conclusion and Future Work

In this paper, we tackled the process of OQA&S. In particular, we analyzed specific methods within each component of this process, i.e., information retrieval, opinion mining and text summarization. These components are crucial in this task, since our final goal was to provide users with the correct information containing the answer of a question. However, contrary to most research work in question answering, we focus on opinionated questions rather than factual, increasing the difficulty of the task.

Our analysis comprises different configurations and approaches: i) varying the length for retrieving the passages of the documents in the retrieval information stage; ii) studying a method that take into consideration topic-sentiment analysis for detecting and classifying opinions in the retrieved passages and comparing it to another that does not; and iii) generating summaries of different compression rates (10% to 50%). The results obtained showed that the proposed methods are appropriate to tackle the OQA&S task, improving state of the art approaches by 12.50% approximately.

In the future, we plan to continue investigating suitable approaches for each of the proposed components. Our final goal is to build an integrated and complete approach.

Acknowledgments

This research work has been funded by the Spanish Government through the research program FPI (BES-2007-16268) associated to the project TEXT-MESS (TIN2006-1526-C06-01). Moreover, it has been also partially funded by projects TEXT-MESS 2.0 (TIN2009-13391-C04), and PROMETEO (PROMETEO/2009/199) from the Spanish and the Valencian Government, respectively.

References

A. Balahur, E. Lloret, O. Ferrández, A. Montoyo, M. Palomar, and R. Muñoz. 2008. The DLSIUAES team's participation in the tac 2008 tracks. In *Proceedings of the Text Analysis Conference*.

Alexandra Balahur, Ralf Steinberger, Erik van der Goot, Bruno Pouliquen, and Mijai Kabadjov. 2009. Opinion mining from newspaper quotations. In *Proceedings of the Workshop on Intelligent Analysis and Processing of Web News Content*.

A. Balahur, M. Kabadjov, and J. Steinberger. 2010. Exploiting higher-level semantic information for the opinion-oriented summarization of blogs. In *Proceedings of CICLing'2010*.

S. Cerini, V. Compagnoni, A. Demontis, M. Formentelli, and G. Gandini. 2007. Micro-WNOp: A gold standard for the evaluation of automatically compiled lexical resources for opinion mining. In *Language resources and linguistic theory: Typology, second language acquisition, English linguistics*.

A. Esuli and F. Sebastiani. 2006. SentiWordNet: A publicly available resource for opinion mining. In *Proceedings of LREC*.

Talmy Givón, 1990. *Syntax: A functional-typological introduction, II.* John Benjamins.

José M. Gómez. 2007. *Recuperación de Pasajes Multilingüe para la Búsqueda de Respuestas*. Ph.D. thesis.

Chin-Yew Lin. 2004. ROUGE: a Package for Automatic Evaluation of Summaries. In *Proceedings of ACL Text Summarization Workshop*, pages 74–81.

Elena Lloret and Manuel Palomar. 2009. A gradual combination of features for building automatic summarisation systems. In *Proceedings of TSD*, pages 16–23.

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

Lizhen Qu, Cigdem Toprak, Niklas jakob, and iryna Gurevych. 2008. Sentence level subjectivity and sentiment analysis experiments in ntcir-7 moat challenge. In *Proceedings of NTCIR-7 Workshop meeting*.

Karen Spärck Jones. 2007. Automatic summarising: The State of the Art. *Information Processing & Management*, 43(6):1449–1481.

V. Stoyanov, C. Cardie, D. Litman, and J. Wiebe. 2004. Evaluating an opinion annotation scheme using a new multi-perspective question and answer corpus. In *AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*.

C. Strapparava and A. Valitutti. 2004. WordNet-Affect: an affective extension of wordnet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1083–1086.

John Carroll Taras Zabibalov. 2008. Almost-unsupervised cross-language opinion analysis at ntcis-7. In *Proceedings of NTCIR-7 Workshop meeting*.

J. Wiebe, T. Wilson, and C. Cardie. 2005. Annotating expressions of opinions and emotions in language. In *Language Resources and Evaluation*, volume 39.

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