An Annotated Corpus of Quoted Opinions in News Articles

Tim O'Keefe James R. Curran Peter Ashwell Irena Koprinska

ə-lab, School of Information Technologies
University of Sydney
NSW 2006, Australia
{tokeefe,james,pash4408,irena}@it.usyd.edu.au

Abstract

Quotes are used in news articles as evidence of a person's opinion, and thus are a useful target for opinion mining. However, labelling each quote with a polarity score directed at a textually-anchored target can ignore the broader issue that the speaker is commenting on. We address this by instead labelling quotes as *supporting* or *opposing* a clear expression of a point of view on a topic, called a *position statement*. Using this we construct a corpus covering 7 topics with 2,228 quotes.

1 Introduction

News articles are a useful target for opinion mining as they discuss salient opinions by newsworthy people. Rather than asserting what a person's opinion is, journalists typically provide evidence by using reported speech, and in particular, direct quotes. We focus on direct quotes as expressions of opinion, as they can be accurately extracted and attributed to a speaker (O'Keefe et al., 2012).

Characterising the opinions in quotes remains challenging. In sentiment analysis over product reviews, polarity labels are commonly used because the target, the product, is clearly identified. However, for quotes on topics of debate, the target and meaning of polarity labels is less clear. For example, labelling a quote about abortion as simply positive or negative is uninformative, as a speaker can use either positive or negative language to support or oppose either side of the debate.

Previous work (Wiebe et al., 2005; Balahur et al., 2010) has addressed this by giving each expression of opinion a textually-anchored target. While this makes sense for named entities, it does not apply as obviously for topics, such as abortion, that may not be directly mentioned. Our solution is to instead define *position statements*, which are clear statements of a viewpoint or position on a particular topic. Quotes related to this topic can then be labelled as *supporting*, *neutral*, or *opposing* the position statement. This disambiguates the meaning of the polarity labels, and allows us to determine the side of the debate that the speaker is on. Table 1 shows the topics and position statements used in this work, and some example quotes from the republic topic are given below. Note that the first example includes no explicit mention of the monarchy or the republic.

Abortion: Women should have the right to choose an abortion.
Carbon tax: Australia should introduce a tax on carbon or an emissions trading scheme to combat global warming.
Immigration: Immigration into Australia should be maintained or increased because its benefits outweigh any negatives.
Reconciliation: The Australian government should formally apologise to the Aboriginal people for past injustices.
Republic: Australia should cease to be a monarchy with the Queen as head of state and become a republic with an Australian head of state.
Same-sex marriage: Same-sex couples should have the right to attain the legal state of marriage as it is for heterosexual couples.
Work choices: Australia should introduce WorkChoices to give employers more control over wages and conditions.

Table 1: Topics and their position statements.

Positive: "I now believe that the time has come... for us to have a truly Australian constitutional head of state."
Neutral: "The establishment of an Australian republic is essentially a symbolic change, with the main arguments, for and against, turning on national identity..."
Negative: "I personally think that the monarchy is a tradition which we want to keep."

With this formulation we define an annotation scheme and build a corpus covering 7 topics, with 100 documents per topic. This corpus includes 3,428 quotes, of which 1,183 were marked invalid, leaving 2,228 that were marked as *supporting*, *neutral*, or *opposing* the relevant topic statement. All quotes in our corpus were annotated by three annotators, with Fleiss' κ values of between 0.43 and 0.45, which is moderate.

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 516–520,
Sofia, Bulgaria, August 4-9 2013.

2 Background

Early work in sentiment analysis (Turney, 2002; Pang et al., 2002; Dave et al., 2003; Blitzer et al., 2007) focused on product and movie reviews, where the text under analysis discusses a single product or movie. In these cases, labels like *positive* and *negative* are appropriate as they align well with the overall communicative goal of the text.

Later work established *aspect-oriented opinion mining* (Hu and Liu, 2004), where the aim is to find features or *aspects* of products that are discussed in a review. The reviewer's position on each aspect can then be classified as *positive* or *negative*, which results in a more fine-grained classification that can be combined to form an *opinion summary*. These approaches assume that each document has a single source (the document's author), whose communicative goal is to evaluate a well-defined target, such as a product or a movie. However this does not hold in news articles, where the goal of the journalist is to present the viewpoints of potentially many people.

Several studies (Wiebe et al., 2005; Wilson et al., 2005; Kim and Hovy, 2006; Godbole et al., 2007) have looked at sentiment in news text, with some (Balahur and Steinberger, 2009; Balahur et al., 2009, 2010) focusing on quotes. In all of these studies the authors have textually-anchored the target of the sentiment. While this makes sense for targets that can be resolved back to named entities, it does not apply as obviously when the quote is arguing for a particular viewpoint in a debate, as the topic may not be mentioned explicitly and polarity labels may not align to sides of the debate.

Work on debate summarisation and subgroup detection (Somasundaran and Wiebe, 2010; Abu-Jbara et al., 2012; Hassan et al., 2012) has often used data from online debate forums, particularly those forums where users are asked to select whether they support or oppose a given proposition before they can participate. This is similar to our aim with news text, where instead of a textually-anchored target, we have a proposition, against which we can evaluate quotes.

3 Position Statements

Our goal in this study is to determine which side of a debate a given quote supports. Assigning polarity labels to a textually-anchored target does not work here for several reasons. Quotes may not mention the debate topic, there may be many relevant textually-anchored targets for a single topic, and polarity labels do not necessarily align with sides of a debate.

		No cont.		Context	
Topic	Quotes	AA	κ	AA	κ
Abortion	343	.77	.57	.73	.53
Carbon tax	278	.71	.42	.57	.34
Immigration	249	.58	.18	.58	.25
Reconcil.	513	.66	.37	.68	.44
Republic	347	.68	.51	.71	.58
Same-sex m.	246	.72	.51	.71	.55
Work choices	269	.72	.45	.65	.44
Total	2,245	.69	.43	.66	.45

Table 2: Average Agreement (AA) and Fleiss' κ over the valid quotes

We instead define *position statements*, which clearly state the position that one side of the debate is arguing for. We can then characterise opinions as *supporting*, *neutral* towards, or *opposing* this particular position. Position statements should not argue for a particular position, rather they should simply state what the position is. Table 1 shows the position statements that we use in this work.

4 Annotation

For our task we expect a set of news articles on a given topic as input, where the direct quotes in the articles have been extracted and attributed to speakers. A position statement will have been defined, that states a point of view on the topic, and a small subset of quotes will have been labelled as *supporting*, *neutral*, or *opposing* the given statement. A system performing this task would then label the remaining quotes as *supporting*, *neutral*, or *opposing*, and return them to the user.

A major contribution of this work is that we construct a fully labelled corpus, which can be used to evaluate systems that perform the task described above. To build this corpus we employed three annotators, one of whom is an author, while the other two were hired using the outsourcing website Freelancer[1]. Our data is drawn from the Sydney Morning Herald[2] archive, which ranges from 1986 until 2009, and it covers seven topics that were subject to debate within Australian news media during that time. For each topic we used

[1] http://www.freelancer.com

[2] http://www.smh.com.au

		No cont.		Context	
Topic	Quotes	AA	κ	AA	κ
Abortion	343	.78	.52	.74	.46
Carbon tax	278	.72	.39	.59	.19
Immigration	249	.58	.08	.58	.14
Reconcil.	513	.66	.31	.69	.36
Republic	347	.69	.39	.72	.41
Same-sex m.	246	.73	.43	.73	.40
Work choices	269	.73	.40	.67	.32
Total	2,245	.70	.36	.68	.32

Table 3: Average Agreement (AA) and Fleiss' κ when the labels are neutral versus non-neutral

Apache Solr[3] to find the top 100 documents that matched a manually-constructed search query. All documents were tokenised and POS-tagged and the named entities were found using the system from Hachey et al. (2013). Finally, the quotes were extracted and attributed to speakers using the system from O'Keefe et al. (2012).

For the first part of the task, annotators were asked to label each quote without considering any context. In other words they were asked to only use the text of the quote itself as evidence for an opinion, not the speaker's prior opinions or the text of the document. They were then asked to label the quote a second time, while considering the text surrounding the quote, although they were still asked to ignore the prior opinions of the speaker. For each of these choices annotators were given a five-point scale ranging from *strong or clear opposition* to *strong or clear support*, where *support* or *opposition* is relative to the position statement.

Annotators were also asked to mark instances where either the speaker or quote span was incorrectly identified, although they were asked to continue annotating the quote as though it were correct. They were also asked to mark quotes that were invalid due to either the quote being off-topic, or the item not being a quote (e.g. book titles, scare quotes, etc.).

5 Corpus results

In order to achieve the least amount of noise in our corpus, we opted to discard quotes that any annotator had marked as invalid. From the original set of 3,428 quotes, 1,183 (35%) were removed, which leaves 2,245 (65%). From the original corpus, 23% were marked off-topic, which shows that in order to label opinions in news, a system would first have to identify the topic-relevant parts of the text. The annotators further indicated that 16% were not quotes, and there were a small number of cases (<1%) where the quote span was incorrect. Annotators were able to select multiple reasons for a quote being invalid.

Table 2 shows both Fleiss' κ and the raw agreement averaged between annotators for each topic. We collapsed the two supporting labels together, as well as the two opposing labels, such that we end up with a classification of *opposes* vs. *neutral* vs. *supports*. The no context and context cases scored 0.69 and 0.66 in raw agreement, while the κ values were 0.43 and 0.45, which is moderate.

Intuitively we expect that the confusion is largely between neutral and the two polar labels. To examine this we merged all the non-neutral labels into one group and calculated the agreement between the non-neutral group and the neutral label, as shown in Table 3. For the non-neutral vs. neutral agreement we find that despite stability in raw agreement, Fleiss' κ drops substantially, to 0.36 (no context) and 0.32 (context).

For comparison we remove all neutral annotations and focus on disagreement between the polar labels. For this we cannot use Fleiss' κ, as it requires a fixed number of annotations per quote, however we can average the pairwise κ values between annotators, which results in values of 0.93 (no context) and 0.92 (context). Though they are not directly comparable, the magnitude of the difference between the numbers (0.36 and 0.32 vs. 0.93 and 0.92) indicates that deciding when an opinion provides sufficient evidence of support or opposition is the main challenge facing annotators.

To adjudicate the decisions annotators made, we opted to take a majority vote for cases of two or three-way agreement, while discarding cases where annotators did not agree (1% of quotes). The final distribution of labels in the corpus is shown in Table 4. For both the no context and context cases the largest class was neutral with 61% and 46% of the corpus respectively. The drop in neutrality between the no context and context cases shows that the interpretation of a quote can change based on the context it is placed in.

6 Discussion

In refining our annotation scheme we noted several factors that make annotation difficult.

[3] http://lucene.apache.org/solr/

	No context				Context			
Topic	Quotes	Opp.	Neut.	Supp.	Quotes	Opp.	Neut.	Supp.
Abortion	343	.13	.63	.25	340	.16	.52	.32
Carbon tax	273	.09	.70	.21	273	.14	.44	.42
Immigration	247	.09	.72	.19	245	.12	.64	.23
Reconciliation	509	.05	.57	.38	503	.07	.42	.50
Republic	345	.24	.48	.28	342	.32	.37	.32
Same-sex marriage	246	.16	.55	.28	243	.25	.38	.37
Work choices	265	.14	.72	.14	266	.26	.50	.24
Total	2,228	.12	.61	.26	2,212	.18	.46	.36

Table 4: Label distribution for the final corpus.

Opinion relevance When discussing a topic, journalists will often delve into the related aspects and opinions that people hold. This introduces a challenge as annotators need to decide whether a particular quote is on-topic enough to be labelled. For instance, these quotes by the same speaker were in an article on the carbon tax:

1) "Whether it's a stealth tax, the emissions trading scheme, whether it's an upfront. . . tax like a carbon tax, there will not be any new taxes as part of the Coalition's policy"

2) "I don't think it's something that we should rush into. But certainly I'm happy to see a debate about the nuclear option."

In the first quote the speaker is voicing opposition to a tax on carbon, which is easy to annotate with our scheme. However in the second quote, the speaker is discussing nuclear power in relation to a carbon tax, which is much more difficult, as it is unclear whether is is *off-topic* or *neutral*.

Obfuscation and self-contradiction While journalists usually quote someone to provide evidence of the person's opinion, there are some cases where they include quotes to show that the person is inconsistent. The following quotes by the same speaker were included in an article to illustrate that the speaker's position was inconsistent:

1) "My point is that. . . the most potent argument in favour of the republic, is that why should we have a Briton as the Queen – who, of course, in reality is also the Queen of Australia – but a Briton as the head of State of Australia"

2) "The Coalition supports the Constitution not because we support the. . . notion of the monarchy, but because we support the way our present Constitution works"

The above example also indicates a level of obfuscation that is reasonably common for politicians. Neither of the quotes actually expresses a clear statement of how the speaker feels about a potential republic. The first quote is an opinion about the strongest argument in favour of a republic, without necessarily making that argument, while the second quote states a party line, with a caveat that might indicate personal disagreement.

Annotator bias This task is prone to be influenced by an annotator's biases, including their political or cultural background, their opinion about the topic or speaker, or their level of knowledge about the topic.

7 Conclusion

In this work we examined the problem of annotating opinions in news articles. We proposed to exploit quotes, as they are used by journalists to provide evidence of an opinion, and are easy to extract and attribute to speakers. Our key contribution is that rather than requiring a textually-anchored target for each quote, we instead label quotes as *supporting*, *neutral*, or *opposing* a position statement, which states a particular viewpoint on a topic. This allowed us to resolve ambiguities that arise when considering a polarity label towards a topic. We next defined an annotation scheme and built a corpus, which covers 7 topics, with 100 documents per topic, and a total of 2,228 annotated quotes. Future work will include building a system able to perform the task we have defined, as well as extending this work to include indirect quotes.

Acknowledgements

O'Keefe was supported by a University of Sydney Merit scholarship and a Capital Markets CRC top-up scholarship. This work was supported by ARC Discovery grant DP1097291 and the Capital Markets CRC Computable News project.

References

Amjad Abu-Jbara, Mona Diab, Pradeep Dasigi, and Dragomir Radev. 2012. Subgroup detection in ideological discussions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 399–409.

Alexandra Balahur and Ralf Steinberger. 2009. Rethinking sentiment analysis in the news: From theory to practice and back. *Proceedings of the First Workshop on Opinion Mining and Sentiment Analysis*.

Alexandra Balahur, Ralf Steinberger, Mijail Kabadjov, Vanni Zavarella, Erik Van Der Goot, Matina Halkia, Bruno Pouliquen, and Jenya Belyaeva. 2010. Sentiment analysis in the news. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, pages 2216–2220.

Alexandra Balahur, Ralf Steinberger, Erik Van Der Goot, Bruno Pouliquen, and Mijail Kabadjov. 2009. Opinion mining on newspaper quotations. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, pages 523–526.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447.

Kushal Dave, Steve Lawrence, and David Pennock. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, pages 519–528.

Namrata Godbole, Manjunath Srinivasaiah, and Steven Skiena. 2007. Large-scale sentiment analysis for news and blogs. In *Proceedings of the International Conference on Weblogs and Social Media*.

Ben Hachey, Will Radford, Joel Nothman, Matthew Honnibal, and James R. Curran. 2013. Evaluating entity linking with Wikipedia. *Artificial Intelligence*.

Ahmed Hassan, Amjad Abu-Jbara, and Dragomir Radev. 2012. Detecting subgroups in online discussions by modeling positive and negative relations among participants. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 59–70.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177.

Soo-Min Kim and Eduard Hovy. 2006. Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, pages 1–8.

Tim O'Keefe, Silvia Pareti, James R. Curran, Irena Koprinska, and Matthew Honnibal. 2012. A sequence labelling approach to quote attribution. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 790–799.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 79–86.

Swapna Somasundaran and Janyce Wiebe. 2010. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 116–124.

Peter Turney. 2002. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 417–424.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2/3):165–210.

Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005. OpinionFinder: a system for subjectivity analysis. In *Proceedings of HLT/EMNLP Interactive Demonstrations*.