Annotating Agreement and Disagreement in Threaded Discussion

Jacob Andreas, Sara Rosenthal, Kathy McKeown

Columbia University
jda2129@columbia.edu, {sara,kathy}@cs.columbia.edu

Abstract

We introduce a new corpus of sentence-level agreement and disagreement annotations over LiveJournal and Wikipedia threads. This is the first agreement corpus to offer full-document annotations for threaded discussions. We provide a methodology for coding responses as well as an implemented tool with an interface that facilitates annotation of a specific response while viewing the full context of the thread. Both the results of an annotator questionnaire and high inter-annotator agreement statistics indicate that the annotations collected are of high quality.

Keywords: annotation, agreement, online discussion

1. Introduction

The identification of agreement and disagreement among participants in a discussion has been widely studied. The problem, however, suffers from a paucity of data. In order to develop systems that can recognize when a dialog participant agrees or disagrees with a previous speaker, corpora that contain gold standard annotations identifying who agrees (or disagrees) with whom are needed. At present, only a small number of corpora covering a limited range of discussion formats exist. As is usually the case, it is difficult to extend systems trained on these corpora to new tasks or differently-formatted data, particularly when discussions in the target data have a fundamentally different *structure* from that of the available training data.

In particular, almost all existing corpora cover linear discussions, typically transcriptions of meetings, in which the sentences or utterances in the discussion proceed in a single chain with a definite order. Such corpora, while good for modeling discussions that take place in real time, are ill-suited to learning patterns found in conversations taking place on Internet message boards or blogs. In online forums, discussions are typically structured as *threads*, tree-shaped structures in which multiple posts can share the same parent. In fact, it is often the case that a single post may elicit many comments, which can either respond to the initial poster or to one of the comments on the post. While there is a definite underlying temporal ordering (the sequence in which posts were created), the most important feature is a set of explicit edges between posts which explain who is responding to whom.

In this paper, we introduce a corpus of agreement and disagreement annotations on two different online sources: Wikipedia discussion forums, and LiveJournal weblogs. This corpus, which we believe to be the first of its kind, provides a source of agreement and disagreement data both for a new source (the Internet) and a new document structure (threaded discussion).

We present the annotation guidelines that were used to create this corpus, and statistics about the annotations that were collected. We also discuss the process by which the corpus was created, highlighting the annotation tool we created for this task and the results of questionnaires that were presented to the annotators.

2. Related Work

Various corpora and annotation tools for more restricted agreement/disagreement annotation tasks exist. DAMSL (Allen and Core, 1997) is a dialog act annotation scheme and tool used to annotate various kinds of communication, including agreement, for speech transcripts. The ICSI meeting corpus (Janin et al., 2003; Shriberg et al., 2004) has similarly been annotated (Galley, 2007) for agreement and other dialog acts. As discussed, these linear transcripts do not generalize well to threaded documents. Perhaps the most similar work to ours is that of Abbott et al (2011); they identify disagreement in political blogs on ⟨quote,response⟩ pairs using only lexical features. In contrast, our annotation tool explores several forms of agreement and disagreement and asks the annotator to take into account the context of the phrases by providing the entire document (which was an optional feature in their annotation). In other related work, Bender et al (2011) annotate Wikipedia discussion forums for positive and negative alignment moves which express agreement and disagreement respectively between the source and target. Their annotation includes praise, doubt, and sarcasm in addition to explicit agreement and disagreement. They did not have an annotation tool, but simply had the annotators annotate the documents directly. Our annotation tool could be easily modified for their approach.

This project is directly motivated by the lack of adequate data for training an agreement module in a larger project aimed at identifying influential participants and subgroup formation in online message boards; we expect that agreement classifiers trained on this data will be useful for a wide variety of higher-level discourse analysis tasks like ours. There is already a great deal of work on the problem of labeling individual utterances in a (linear) meeting transcript; we note in particular the work of (Galley et al., 2004), which focused on identifying adjacency pairs, and a similar paper by (Hillard et al., 2003), which studied the same task using a reduced feature set. More recent work on agreement/disagreement detection includes (Hahn et al., 2006; Germesin and Wilson, 2009; Wang et al., 2011). We hope that this corpus will enable a generalization of (Galley et al., 2004)'s work to other document structures.

c_1 There seems to be a much better list at the National Cancer Institute than the one we've got. It ties much better to the actual publication (the same 11 sections, in the same order). I'd like to replace that section in this article. Any objections?
c_2 Not a problem. Perhaps we can also insert the relative incidence as published in this month's wiki Blood journal
c_3 I've made the update. I've included template links to a source that supports looking up information by ICD-O code.
c_4 Can Arcadian tell me why he/she included the leukemia classification to this lymphoma page? It is not even listed in the Wikipedia leukemia page! I vote for dividing the WHO classification into 4 parts in 4 distinct pages: leukemia, lymphoma, histocytic and mastocytic neoplasms. Let me know what you think before I delete them.
c_5 Emmanuelm, aren't you the person who added those other categories on 6 July 2005?
c_6 Arcadian, I added only the lymphoma portion of the WHO classification. You added the leukemias on Dec 29th. Would you mind moving the leukemia portion to the leukemia page
c_7 Oh, and please note that I would be very comfortable with a "cross-coverage" of lymphocytic leukemias in both pages. My comment is really about myeloid, histiocytic and mast cell neoplasms who share no real relationship with lymphomas.
c_8 To simplify the discussion, I have restored that section to your version. You may make any further edits, and I will have no objection.

Table 1: Examples of agreement and disagreement in a Wikipedia discussion forum. Direct Response: $c_2 \rightarrow c_1$, $c_6 \rightarrow c_5$, $c_8 \rightarrow c_7, c_6$

c_1 I want this jacket. Because 100% silk is so practical, especially in a household with cats. but it's so, I don't know – raggedy looking! That's awesome!
c_2 That jacket is gorgeous. Impractical, way too expensive for the look, and pretty much gorgeous. guh.
c_3 I knoooooow, and you're not helping. :)
c_4 Monday! WHEE! It is a bit raggedy looking. I think it's because of the ties.
c_5 Wow, that jacket looks really nice... I wish I could afford it!

Table 2: Examples of a agreement in a LiveJournal weblog. Direct Response: $c_2 \rightarrow c_1$, $c_5 \rightarrow c_1$, Direct Paraphrase: $c_4 \rightarrow c_1$, Indirect Paraphrase: $c_5 \rightarrow c_2$

3. Annotation Guidelines

A thread consists of a set of posts organized in a tree. We use standard terminology to refer to the structure of this tree (so every post has a single "parent" to which it replies, and all nodes descend from a single "root"). Each post is marked with a timestamp and an author, a string (its "body").

Agreement annotation is performed on pairs of sentences $\{s, t\}$, where each sentence is a substring of the body of a post. s is referred to as the "antecedent sentence", and t as the "reaction sentence." The antecedent sentence and reaction sentence occur in different posts written by different authors. Annotations are implicitly directed from reaction to antecedent; the reaction is always the sentence from the post with the later timestamp. Annotations between pairs of posts with the same author are forbidden. Each pair is also annotated with a type.

Type

Each sentence pair can be of either type agreement or disagreement. Two sentences are in agreement when they provide evidence that their authors believe the same fact or opinion, and in disagreement otherwise.

Mode

Mode indicates the manner in which agreement or disagreement is expressed. Broadly, a pair of posts are in a "direct" relationship if one is an ancestor of the other, and indirect otherwise; they are in a "response" relationship if one explicitly acknowledges a claim made in the other, and "paraphrase" otherwise. More specifically:

Direct response: The reaction author explicitly states that they are in agreement or disagreement, e.g. by saying "I agree" or "No, that's not true." An agreement/disagreement is only a direct response if it is a direct reply to its closest ancestor, i.e. its parent. For example, in Table 2, the reaction sentence *"I knooooow."* in c_3 is a direct response to the antecedent sentence *"That jacket is gorgeous."* in c_2. In Table 1, the reaction sentence *"Arcadian, I added only the lymphoma portion of the WHO classification."* in c_6 is a direct disagreement to the sentence *"Emmanuelm, arent you the person who added those other categories on 6 July 2005?"* in c_5.

Direct paraphrase: The reaction author restates a claim made in an ancestor post. An agreement/disagreement is only a direct paraphrase if it is a direct rewording of its closest ancestor, i.e. its parent. For example, in Table 2, the sentence *"It is a bit raggedy looking."* in c_4 is a direct paraphrase of the sentence *"but it's so, I don't know – raggedy looking!"* of its parent, c_1.

Indirect response: The reaction is a direct response to a claim, but the post does not descend from the source. This often occurs when the author pressed the "reply"

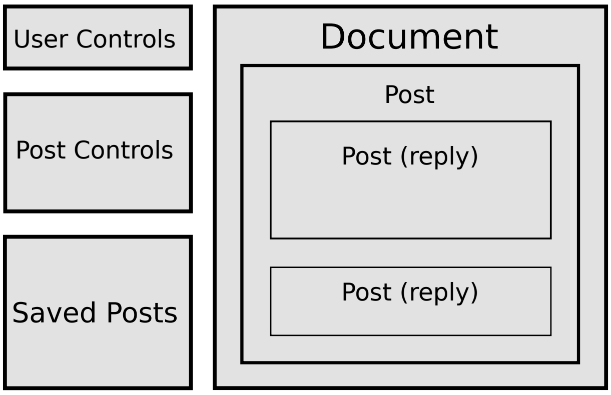


Figure 1: Schematic of the annotation tool: The left side shows the controls used for navigation and the right displays the current thread.

button on a post other than the one they were attempting to respond to (this would be the case if, for example, c_3 descended from c_5 instead of c_2 above). Or, perhaps it is intended to answer more than one previous post. The reaction of an indirect response should be the single sentence written closest in time to its antecedent.

Indirect paraphrase: The reaction restates a claim made in a post that is earlier in time, but not an ancestor of the post. The reaction of an indirect paraphrase annotation be the single sentence written closest in time to its antecedent. For example, in Table 2, c_5 is an indirect paraphrase of c_2.

4. The Annotation Process

In recent years there has been an increasingly popular trend to use Amazon's Mechanical Turk to label data. Studies have shown (Snow et al., 2008) that Mechanical Turk users are able to produce high quality data that is comparable to expert annotators for simple labeling tasks that can be completed in a few seconds such as affective text analysis and word sense disambiguation. However, others have shown (Callison-Burch and Dredze, 2010) that the annotations are considerably less reliable for tasks requiring a substantial amount of reading or involving complicated annotation schemes. Our annotation task was difficult for several reasons; observation of the entire thread was necessary to annotate each edge and the annotations themselves were fairly involved. Therefore, we decided to rely on two trained annotators rather than a large number of untrained annotators.

Both annotators were undergraduates, and neither had any previous experience with NLP annotation tasks. They were trained to use the web-based annotation tool described in the following section for approximately one hour; they then annotated the remainder of the corpus on their own.

4.1. The Annotation Tool

The web-based annotation tool (Fig. 1 and 2) is used to provide a simple and easy way to annotate threads. The interface consists of two parts; the left hand-side which contains controls to navigate through threads and add agreements, and the right-hand side which displays the current thread.

The document is displayed using its thread structure indicated by both indentation and boxes nesting children under their parents as shown in Figure 1. To clearly differentiate between possible sentences, each sentence is displayed on its own line.

Annotators begin by selecting the individual sentences from the antecedent and reaction which provide evidence of agreement or disagreement, and then mark type and mode using the post controls on the left-hand side. (While the response/paraphrase annotation must be encoded by hand, the direct/indirect distinction is inferred automatically from the document structure.) Each post pair that is added appears as a saved post on the left-hand side of the tool directly below the post controls. Saved posts can be removed if they were mistakenly added. Figure 2 shows the annotation tool in use.

The system automatically prevents users from annotating the forbidden cases mentioned in Section 3., such as the antecedent and reaction sentences being written by the same author. It also automatically determines the antecedent and reaction of the annotation based on the timestamps of the two posts involved.

The annotation tool outputs a CSV file (Table 3) encoding the annotator ID, the document ID and the post structure as a JSON array with entries for each annotated arc.

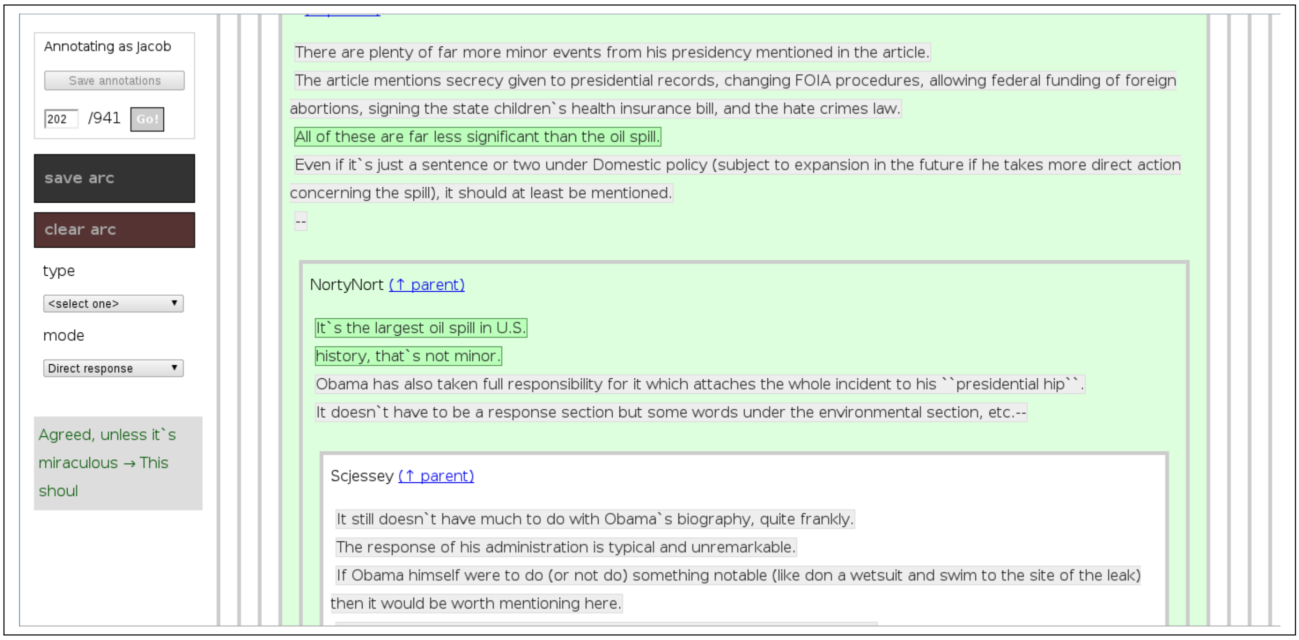
Field	Value
Document ID	5
Annotator	John Doe
Antecedent ID	13
Reaction ID	11
Antecedent	It does seem heavily censored
Reaction	Um, I can‘t help but notice that this article seems heavily censored.
Type	agreement
Mode	indirect paraphrase

Table 3: Sample annotator output

5. User Studies

After completing their portion of the task, annotators were asked to fill out a brief questionnaire describing their experience (Fig. 3). They reported that the annotation tool was “easy to use” and “effective”, that the annotation task was “interesting”, and that there were “no real challenges” in annotating. They reported that between the two genres, the LiveJournal entries were both conceptually easier to annotate and required less time, primarily because the posts were shorter in length. Annotators reported that LiveJournal entries required an average 2 to 10 minutes to annotate, while Wikipedia discussions required 10 to 20 minutes. Annotators were divided in their opinions on whether agreements or disagreements, and direct or indirect were easier to identify indicating that it is a matter of personal preference.

These responses have given us confidence that the annotation tool succeeded in its purpose (of simplifying the data



Annotating as Jacob

Save annotations

202 /941 Go!

save arc

clear arc

type

<select one>

mode

Direct response

Agreed, unless it`s miraculous → This shoul

There are plenty of far more minor events from his presidency mentioned in the article.

The article mentions secrecy given to presidential records, changing FOIA procedures, allowing federal funding of foreign abortions, signing the state children`s health insurance bill, and the hate crimes law.

All of these are far less significant than the oil spill.

Even if it`s just a sentence or two under Domestic policy (subject to expansion in the future if he takes more direct action concerning the spill), it should at least be mentioned.

--

NortyNort (↑ parent)

It`s the largest oil spill in U.S.

history, that`s not minor.

Obama has also taken full responsibility for it which attaches the whole incident to his ``presidential hip``.

It doesn`t have to be a response section but some words under the environmental section, etc.--

Scjessey (↑ parent)

It still doesn`t have much to do with Obama`s biography, quite frankly.

The response of his administration is typical and unremarkable.

If Obama himself were to do (or not do) something notable (like don a wetsuit and swim to the site of the leak) then it would be worth mentioning here.

Figure 2: Screenshot of the annotation tool in use.

1. Did you find the tool easy to use?
2. What challenges did you encounter when using the tool?
3. Were the LiveJournal or Wikipedia discussions easier to annotate?
4. Were the LiveJournal or Wikipedia discussions faster to annotate?
5. Did you have any previous experience with annotation?
6. What was the learning curve associated with the task?
7. On average, how long did it take you to complete a single LiveJournal discussion? A Wikipedia discussion?
8. Was it easier to find agreements or disagreements?
9. Was it easier to find direct or indirect agreements/disagreements?
10. What is your general opinion about the task?
11. Is there anything else that you would like to let us know about with regard to this annotation task?

Figure 3: Annotator questionnaire

collection process for this corpus), and that it will be easy to further expand the corpus if we require additional data.

6. Corpus

Documents in the corpus came from two sources: Wikipedia and LiveJournal. Wikipedia is, a free, collaboratively-edited encyclopedia which records conversations among editors about page content, and LiveJournal is, a journaling website which allows threaded discussion about each posting. In order to ensure that the corpus contains documents in which substantial conversation takes place, both of these sources were initially filtered for threads where

		Indirect	Direct	Tot.
LiveJournal	agreement	143	236	379
	disagreement	14	66	80
	total	157	302	459
Wikipedia	agreement	30	111	141
	disagreement	38	172	210
	total	68	283	351

Table 4: The number or direct and indirect responses in the corpus

$$\frac{\text{\# of posts}}{\text{\# of participants}} > 1.5$$

At the time of publication, one annotator had labeled 92 documents and the other 109. In total, 118 unique documents were labeled; the 83 documents annotated in common were used to determine an inter-annotator agreement statistic. Restricted just to the three-class agreement/disagreement/none labeling task on all edges, their annotations had Cohen's $\kappa = 0.73$, indicating substantial agreement. Considering the more granular five-class labeling task that distinguishes between agreement-response, agreement-paraphrase, disagreement-response and disagreement-paraphrase, we have $\kappa = 0.66$, also indicating substantial agreement. Table 4 shows the breakdown of direct vs. indirect responses in the corpus.

In general, we observed that a greater percentage of the posts in LiveJournal entries participated in agreement/disagreement relations while Wikipedia articles tended to have a higher disagreement / agreement ratio.

7. Conclusion

We have introduced a new corpus of agreement and disagreement annotations over threaded online discussions.

In addition to the resulting corpus, we have also provided a methodology for labeling online discussions which makes use of thread structure and distinguishes whether agreement/disagreement was directly stated or conveyed by means of a paraphrase of the original post. The corpus was collected using an easy-to-use annotation tool by a pair of trained annotators. Inter-annotator agreement showed substantial agreement for both the three-class and five-class labeling tasks, with Kappa of .67 and above. The annotation process is ongoing, and we plan to release the complete corpus at a later time.

8. Acknowledgements

This research was funded by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), through the U.S. Army Research Lab. All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of IARPA, the ODNI or the U.S. Government.

9. References

Rob Abbott, Marilyn Walker, Pranav Anand, Jean E. Fox Tree, Robeson Bowmani, and Joseph King. 2011. How can you say such things?!?: Recognizing disagreement in informal political argument. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 2–11, Portland, Oregon, June. Association for Computational Linguistics.

James Allen and Mark Core. 1997. Draft of DAMSL: Dialog act markup in several layers. Unpublished manuscript.

Emily M. Bender, Jonathan T. Morgan, Meghan Oxley, Mark Zachry, Brian Hutchinson, Alex Marin, Bin Zhang, and Mari Ostendorf. 2011. Annotating social acts: authority claims and alignment moves in wikipedia talk pages. In *Proceedings of the Workshop on Languages in Social Media*, LSM '11, pages 48–57, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chris Callison-Burch and Mark Dredze. 2010. Creating speech and language data with amazon's mechanical turk. *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazons Mechanical Turk Los Angeles*, (June):1–12.

Michel Galley, McKeown, Kathleen, Julia Hirschberg, and Elizabeth Shriberg. 2004. Identifying agreement and disagreement in conversational speech: Use of bayesian networks to model pragmatic dependencies. In *Proceedings of the ACL*.

Michel Galley. 2007. *Incorporating discourse and syntactic dependencies into probabilistic models for summarization of multiparty speech*. Ph.D. thesis, Columbia University, New York, NY, USA. Adviser-Mckeown,, Kathleen R.

Sebastian Germesin and Theresa Wilson. 2009. Agreement detection in multiparty conversation. In *ICMI-MLMI '09*, November.

Sangyun Hahn, Richard Ladner, and Mari Ostendorf. 2006. Agreement/disagreement classification: exploiting unlabeled data using contrast classifiers. In *In HLT/NAACL 2006*.

Dustin Hillard, Mari Ostendorf, and Elizabeth Shriberg. 2003. Detection of agreement vs. disagreement in meetings: Training with unlabeled data. In *Proceedings of the HLT-NAACL Conference*.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters. 2003. The icsi meeting corpus. pages 364–367.

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. The icsi meeting recorder dialog act (mrda) corpus. In Michael Strube and Candy Sidner, editors, *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, pages 97–100, Cambridge, Massachusetts, USA, April 30 - May 1. Association for Computational Linguistics.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of EMNLP*, pages 254–263.

Wen Wang, Sibel Yaman, Kristin Precoda, Colleen Richey, and Geoffrey Raymond. 2011. Detection of Agreement and Disagreement in Broadcast Conversations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 374–378, Portland, Oregon, USA, June. Association for Computational Linguistics.