I do not disagree: Leveraging monolingual alignment to detect disagreement in dialogue

Ajda Gokcen and **Marie-Catherine de Marneffe**
Linguistics Department
The Ohio State University
gokcen.2@osu.edu, mcdm@ling.ohio-state.edu

Abstract

A wide array of natural dialogue discourse can be found on the internet. Previous attempts to automatically determine disagreement between interlocutors in such dialogue have mostly relied on n-gram and grammatical dependency features taken from respondent text. Agreement-disagreement classifiers built upon these baseline features tend to do poorly, yet have proven difficult to improve upon. Using the Internet Argument Corpus, which comprises quote and response post pairs taken from an online debate forum with human-annotated agreement scoring, we introduce semantic environment features derived by comparing quote and response sentences which align well. We show that this method improves classifier accuracy relative to the baseline method namely in the retrieval of disagreeing pairs, which improves from 69% to 77%.

1 Introduction

To achieve robust text understanding, natural language processing systems need to automatically extract information that is expressed indirectly. Here we focus on identifying agreement and disagreement in online debate posts. Previous work on this task has used very shallow linguistic analysis: features are surface-level ones, such as n-grams, post initial unigrams, bigrams and trigrams (which aim at learning the discourse functions of discourse markers, e.g., *well*, *really*, *you know*), repeated sequential use of punctuation signs (e.g., *!!*, *?!*). When automatically detecting (dis)agreement, these features fall short, reaching around 65% accuracy on a balanced dataset (Abbott et al., 2011; Misra and Walker, 2013). Adding extra-linguistic features, such as the structure of the post threads and stance of the post's author on other subjects, boosts performance to 75% (Hasan and Ng, 2013). In this work, we leverage richer linguistic models to increase performance.

Agreement may be explicitly marked. In example (1) in Table 2, the response-initial bigram *I agree* is a strong cue of agreement that surface features can learn, but there are more complex examples that surface features cannot capture. In example (2), the response-initial word *Yes* is not indicating agreement, despite being in general a good cue for it. Instead it is necessary to capture the polarity mismatch between the first sentence in the quote and the first sentence in the response (*God doesn't take away sinful desires* vs. *Yes, God does take away sinful desires*) to infer that the response disagrees with the quote. There may also be mismatches of modality, as demonstrated in the third example (*saw* vs. *may have believed*). Here we also see an example of an explicit agreement word which is negated *(that does **not** make it **true**)* in a way that most surface features fail to capture.

Some discourse-level parsing (Joty et al., 2013) has been utilized in agreement detection, but most previous work does not take discourse structure into account: the response post is simply taken *as a whole* as the reply to the quote. To overcome this issue, we take advantage of the considerable progress in monolingual alignment (e.g., Thadani et al. 2012, Yao et al. 2013, Sultan et al. 2014) which allows us to align sentences of the quote to sentences in the response. This approach is reminiscent of the one used for Recognizing Textual Entailment (RTE, Dagan et al. 2006, Giampiccolo et al. 2007) where, given two short passages, systems identify whether the second passage follows from the first one according to the intuitions of an intelligent human reader. One common approach used in RTE was to align the two passages, and reason based on the alignment obtained.

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), pages 94–99,
Beijing, China, July 26-31, 2015.

	Quote	Response	Score
1	CCW LAWS ARE FOR TRACKING GUN OWNERS WHO EXERCISE THIER RIGHTS!!!	I agree. What is the point? Felons with firearms do not bother with CCW licenses.	2.5
2	God doesn't take away sinful desires. You've never had sinful desires? I know I have. People assume that when you become a Christian some manner of shield gets put up around you and shields you from "worldly" things. I believe that's wrong, I actually believe that life as a Christian is very hard. We often pawn it off as the end of our troubles to "convert" people. I don't believe it.	Yes, God does take away sinful desires. (If you ask Him.) I'm not saying that it doesn't take any work on your part, though. When you have a sinful desire, you allow a thought to become more than just a stray idea. You foster and encourage the thought and it becomes a desire. God takes away the desires, helps you deal with your "stray thoughts", and shows you how to keep them from becoming desires.	-1.7
3	Your idea about science is a philosophy of science. [...] *The Apostles saw Jesus walk on water.* There was no 'measure' by your version of science, but what they saw remains true.	Many people once believed that the earth is flat: perhaps some still do. [...] *The apostles may have believed that Jesus walked on water: that does NOT make it true.*	-2
4	*does life end here?*	*end where?* ambiguously phrased. if "here" = "death", then yes! by definition, yes!	-1.4
5	*Is even 'channel' sufficiently ateleological a verb?*	Yes. It describes an action without ascribing its form to its end result, outcome, whatever but strictly to a cause's force's in action. [...] *But since it is understood that mechanical forces can also 'channel', unintentional, out of simple mechanics, the word channel cannot be called teleological.* In the same way, 'sorting' can be considered non-teleological, hence mechanical, and thus suited to your glossary, because things can be sorted by mechanical forces alone.	2.8

Table 1: QR pairs from the Internet Argument Corpus.

Here, similarly, once we have identified sentences in the response which align well with sentences in the quote, it becomes easier to extract deep semantic features such as polarity and modality mismatch between sentences as well as embeddings under modality, negation, or attitude verbs. For instance, in example (2) in Table 1, the first sentence in the quote gets aligned with high probability to the first sentence in the response, which enables us to identify the polarity mismatch (*doesn't* vs. *does*). In example (3), the italicized sentences are the most well-aligned, enabling us to identify that the response's author embeds under modality the event of Jesus walking on water and thus does not take it as a fact, whereas the quote's author does take it as a fact.

Our experiments demonstrate that our linguistic model based on alignment significantly outperforms a baseline bag-of-words model in the recall of disagreeing quote-response (QR) pairs. Such linguistic models will transfer more easily to any debate dialogue, independent of the structural information of post threads and author's stance which might not always be recoverable.

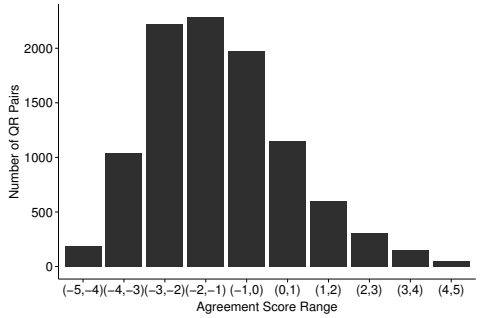
	Full Data Set	Balanced Training Set
Disagree	5741	779
Neutral	3125	0
Agree	1113	779
Total	9980	1158

Table 2: Category counts in the training set.

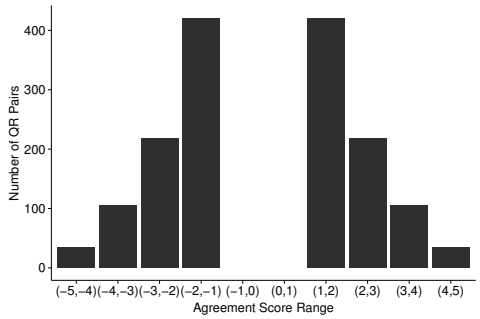
2 Data

We used the Internet Argument Corpus (IAC), a corpus of quote-response pairs annotated for agreement via Mechanical Turk (Walker et al., 2012). Agreement scores span from -5 (strong disagreement) and +5 (strong agreement). The distribution is shown in Figure 2. Because the original data skews toward disagreement, following Abbott et al. (2011), we created a balanced set, discarding "neutral" pairs between -1 and +1. We split the data into training, development and test sets. [1] Table 2 shows the category counts in the training set.

[1]We could not obtain the training-development-test split from Abbott et al. (2011). Our split is available at `www.ling.ohio-state.edu/~mcdm/data/2015/Balanced_IAC.zip`.



(a) Full dataset.



(b) Balanced training set.

Figure 1: Agreement score distribution of the dataset, before and after balancing. -5 is high disagreement, +5 is high agreement.

3 Features

In this section, we detail the features of our model. We use the maximum entropy model as implemented in the Stanford CoreNLP toolset (Manning and Klein, 2003). Many of the features make use of the typed dependencies from the CoreNLP toolset (de Marneffe et al., 2006). For comparison, the baseline features attempt to replicate Abbott et al. (2011).

3.1 Baseline Features from Abbott et al. 2011

N-Grams. All unigrams, bigrams, and trigrams were taken from each response.

Discourse Markers. In lieu of tracking discourse markers such as *oh* and *so really*, Abbott et al. (2011) tracked response-initial unigrams, bigrams, and trigrams.

Typed Dependencies and MPQA. In addition to all dependencies from the response being used as features, dependencies were supplemented with MPQA sentiment values (Wilson et al., 2005). A dependency like (*agree,I*) would also yield the sentiment-dependency feature (*positive,I*), whereas (*wrong, you*) would also yield (*negative,you*).

Punctuation. The presence of special punctuation such as repeated exclamation points *(!!)*, question marks *(??)*, and interrobang strings *(?!)* were tracked as binary features.

3.2 Alignment+ Features

Our features utilize focal sentences: not only well-aligned sentences from the quote and response, but also the first sentence of the response in general. Tracking certain features in initial and aligned sentences proved more informative than doing the same without discerning location.

Alignment scoring comes from running the Jacana aligner (Yao et al., 2013) pairwise on every sentence from each QR pair. Pairs of quote and response sentences with alignment scores above a threshold tuned on the development set are then analyzed for feature extraction. The sentence pair with the maximum alignment score for each post pair is also analyzed regardless of its meeting the threshold.

Post Length. Following Misra and Walker (2013), we track various length features such as word count, sentence count, and average sentence length, including differentials of these measures between quote and response. Short responses (relative to both word-wise and sentence-wise counts) tend to correlate with agreement, while longer responses tend to correlate with disagreement.

Emoticons. Emoticons are a popular way of communicating sentiment in internet text. Many emoticons in the corpus are in forum-specific code, such as *emoticon_rolleyes*. We also detect a wider array of common emoticons as regular expressions beginning with colons, semicolons, or equals signs, such as *:-D*, *;)*, and *=)*.

Speech Acts. To account for phenomena such as commands (e.g., *please read carefully*, *try again*, and *define evil*) and the rhetorical use of multiple questions in a row, we use punctuation, dependencies, and phrase-level analysis to automatically detect and count interrogative and imperative sentences. A phrase-structure tree headed by SQ or a sentence-final question mark means a sentence is considered interrogative; if a sentence's root is labeled VB and has no subject relation, it is deemed an imperative. The features in the classifier are counts of the instances of interrogatives and imperatives in the response.

	Accuracy	Agreement			Disagreement		
		P	R	F1	P	R	F1
Baseline	71.85	70.64	74.77	72.65	73.21	68.92	71.00
Alignment+	75.45	76.04	74.32	75.17	74.89	76.58	75.73

Table 3: Accuracy, precision (P), recall (R) and F1 scores for both categories (agreement and disagreement) on the test set.

Personal Pronouns. The presence of first, second, and third person pronouns in the response are each tracked as binary features. The inclusion of personal pronouns in a post tends to indicate a more emotional or personal argument, especially second person pronouns.

Explicit Truth Values. Rather than simply relying on n-gram-based tracking of explicit statements of agreement, we include as features polar (positive or negative) and modal (modal or non-modal) context of instances of the words *agree*, *disagree*, *true*, *false*, *right*, and *wrong* found in the response, parallel to the agreement and denial tracking in Misra and Walker (2013). Polar context is determined by the presence or absence of negation modifiers (e.g., *not*, *never*) in the dependencies; modal context is determined by the presence of modal auxiliaries (e.g., *might*, *could*) and adverbs (e.g., *possibly*).

Sentiment Scoring. Expanding on the use of MPQA sentiment values, we use the *positive/negative/neutral* and *strong/weak* classifications of the words in the MPQA lexicon to calculate sentiment scores of the posts and focal sentences (well-aligned sentences from the quote and response as well as the first sentence of the response). The scoring assigns a value to each MPQA word in the quote or response: the *positive/negative* label of a word means a positive or negative score and the *strong/weak* label determines the weight: whether the word is worth +/-2 or +/-1. The sum of these values is computed as the sentiment score. A score is generated for both the response and quote in their entireties as well as for focal sentences.

Discourse Markers. Initial 1, 2 and 3-grams are tracked relative to focal sentences. This picks up on discourse markers (such as *well* and *but*) without having to explicitly code for each marker we want to track.

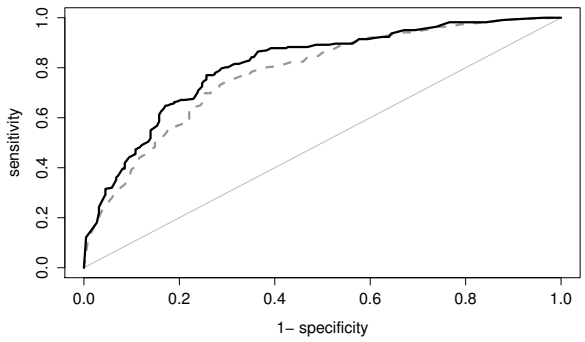


Figure 2: ROC curves. The gray dotted line represents the baseline feature set, while the solid black line represents the alignment+ feature set.

Punctuation. As in the baseline, the presence of special punctuation like *!!* and *?!* are used as binary features.

Factuality Comparison. Given aligned words from well-aligned sentences in the quote and response (e.g., *God doesn't* ***take*** *away sinful desires* and *Yes, God does* ***take*** *away sinful desires*), we analyze the polarity, modality, and any subsequent contradiction of both the quote and response instances. As with the analysis of explicit truth value words, polarity and modality are determined according to the presence or absence of negation and modal modifiers (auxiliaries and adverbs) in the dependencies. Contradictions are tracked as phrases marked with known contradictory adverbs and conjunctions (e.g., *however...*, *but...*). An aligned word pair is analyzed if it involves content words, or if the words serve as the root of their sentence's dependency structure regardless of part of speech. The features generated indicate the part of speech of the word in the quote and whether there is (1) a polarity match/clash, (2) a modality match/clash, or (3) any contradiction phrases immediately following the word or sentence in the quote or response.

4 Results and Discussion

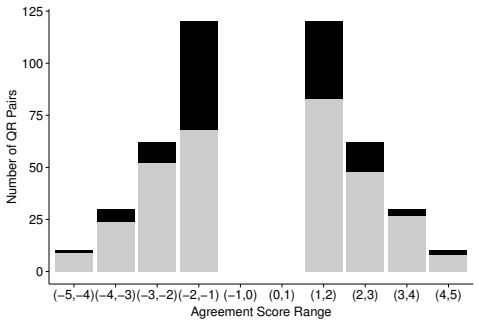
Table 3 compares the results obtained with the baseline features and the alignment+ features. The alignment+ features lead to an overall improvement, but a statistically significant improvement ($p < 0.05$, McNemar's test) is only achieved for classifying disagreeing pairs. The baseline model underclassifies for disagreement and overclassifies for agreement, but the alignment+ model does well on both. As most cases of high alignment do, indeed, correspond with disagreement, these features are better in picking up on disagreement in general. The ROC curve in Figure 3 shows that the alignment+ classifier consistently has a higher sensitivity (true-positive) rate than the baseline.

Figure 4 shows for both feature sets (baseline and alignment+) the correct (gray bar) and incorrect (black bar) classifications on the test set, by agreement score. The agreement score is predictive of the correctness of the system (confirmed by a logistic regression predicting system accuracy given strength of agreement score, $p < 0.001$): the stronger the (dis)agreement score, the more accurate the system is. The alignment+ features help classify accurately the less strong (dis)agreements.

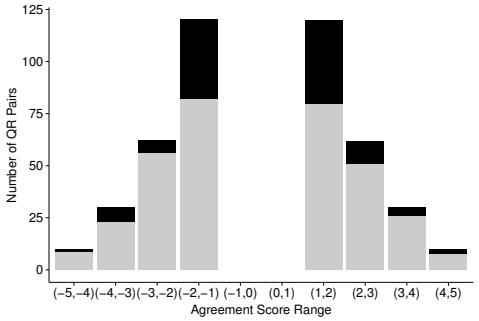
Examples (4) and (5) in Table 1 are incorrectly classified by the baseline but correctly by the alignment+ classifier. In (4), the strongest feature in the baseline is the unigram *yes*, but the alignment+ features compare *does life end here?* to *end where?*, and the fact that the aligned sentence in the response is a question suggests disagreement. Example (5) shows that superficial features like a response-initial *yes* are not always enough, even when the pair is indeed in agreement. Here the alignment+ model aligns the italic sentences (*Is even 'channel' sufficiently ateleological a verb?* and *[...]the word channel cannot be called teleological*), finding them to be in agreement and thus getting the correct classification.

5 Conclusion

The incorporation of alignment-based features shows promise in improving agreement classification. Further ablation testing is needed to determine the full extent to which alignment features contribute, and not only better whole-post features on their own. However, given that many pairs do not have sentences which align at all, alignment features cannot classify on their own without some more basic features to fill in the gaps.



(a) Baseline feature set classifications.



(b) Alignment+ feature set classifications.

Figure 3: Correct and incorrect classifications on the test set given the corpus agreement scores, for both feature sets. The gray area represents correct classifications, while the black area represents incorrect classifications.

Following previous work, we focused on pairs judged as being in strong (dis)agreement. How do systems fare when uncertain cases are present in the training data? This has not been investigated. One aspect of language interpretation, however, is its inherent uncertainty. In future work, we will use the full IAC corpus, and instead of drawing a binary distinction between strong agreements and disagreements, have a three-way classification where unclear instances are also categorized.

Acknowledgments

We thank Christopher Potts and the members of the Clippers group at The Ohio State University for productive discussions about this work, as well as our anonymous reviewers for their helpful comments. We also thank Xuchen Yao for his tremendous help with the Jacana aligner. This research is supported in part by the NSF under Grant No. 1464252. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the NSF.

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 Languages in Social Media*, pages 2–11.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In J. Quinonero-Candela, I. Dagan, B. Magnini, and F. d'Alché-Buc, editors, *Machine Learning Challenges, Lecture Notes in Computer Science*, volume 3944, pages 177–190. Springer-Verlag.

Marie-Catherine de Marneffe, Bill McCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*.

Danilo Giampiccolo, Ido Dagan, Bernardo Magnini, and Bill Dolan. 2007. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 1–9.

Kazi Saidul Hasan and Vincent Ng. 2013. Extra-linguistic constraints on stance recognition in ideological debates. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 816–821.

Shafiq Joty, Giuseppe Carenini, Raymond T. Ng, and Yashar Mehdad. 2013. Combining intra-and multi-sentential rhetorical parsing for document-level discourse analysis. In *ACL (1)*, pages 486–496.

Christopher D. Manning and Dan Klein. 2003. Optimization, maxent models, and conditional estimation without magic. In *Tutorial at HLT-NAACL 2003 and ACL 2003*. http://nlp.stanford.edu/software/classifier.shtml.

Amita Misra and Marilyn A. Walker. 2013. Topic independent identification of agreement and disagreement in social media dialogue. In *Conference of the Special Interest Group on Discourse and Dialogue*, pages 41–50.

Md Arafat Sultan, Steven Bethard, and Tamara Sumner. 2014. Back to basics for monolingual alignment: Exploiting word similarity and contextual evidence. *Transactions of the Association for Computational Linguistics*, 2:219–230.

Kapil Thadani, Scott Martin, and Michael White. 2012. A joint phrasal and dependency model for paraphrase alignment. In *Proceedings of COLING 2012*, pages 1229–1238. The COLING 2012 Organizing Committee.

Marilyn A. Walker, Jean E. Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *Proceedings of the 8th Language Resources and Evaluation Conference*, pages 812–817.

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 *Demonstration Description in Conference on Empirical Methods in Natural Language Processing*, pages 34–35.

Xuchen Yao, Benjamin Van Durme, Chris Callison-Burch, and Peter Clark. 2013. A lightweight and high performance monolingual word aligner. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 702–707.