The Role of Polarity in Inferring Acceptance and Rejection in Dialogue

Julian J. Schlöder and **Raquel Fernández**
Institute for Logic, Language & Computation
University of Amsterdam
julian.schloeder@gmail.com, raquel.fernandez@uva.nl

Abstract

We study the role that logical polarity plays in determining the rejection or acceptance function of an utterance in dialogue. We develop a model inspired by recent work on the semantics of negation and polarity particles and test it on annotated data from two spoken dialogue corpora: the Switchboard Corpus and the AMI Meeting Corpus. Our experiments show that taking into account the relative polarity of a proposal under discussion and of its response greatly helps to distinguish rejections from acceptances in both corpora.

1 Introduction

In order to establish and maintain coherence, dialogue participants need to keep track of the information they jointly take for granted—their *common ground* (Stalnaker, 1978). As a dialogue progresses, the common ground typically evolves. New information becomes shared as the interlocutors exchange moves (such as assertions, questions, acceptances, and rejections) through the collaborative process of *grounding* (Clark and Schaefer, 1989; Clark, 1996). To keep track of the common ground, speakers must identify which information is accepted or rejected by their addressees. The basic idea is simple: If a proposal is rejected, its content does not enter the common ground, while if it is accepted, its content does become common belief.

Yet, determining whether a response to a move counts as an acceptance or a rejection is far from trivial. In many cases, the surface form of an utterance is not explicit enough to determine its acceptance or rejection force and inference is required (Horn, 1989; Lascarides and Asher, 2009; Walker, 1996). For instance, B's utterance in (1), extracted from the AMI Meeting Corpus (Carletta, 2007), exemplifies what Walker (1996) calls *implicature rejection* (the rejection arises from an inferred scalar implicature: *"normal"* implicates *"not interesting"*; see also Hirschberg (1985)).

(1) A: This is a very interesting design.
B: It's just the same as normal.

The goal of this paper is to investigate the role of *logical polarity* in distinguishing rejections from acceptances. Consider the following dialogue excerpts, again from AMI, where the same utterance form (*"Yes it is"*) acts as an acceptance in (2) and as a rejection in (3):

(2) A: But it's uh yeah it's uh original idea.
B: Yes it is.

(3) A: the shape of a banana is not it's not really handy .
B: Yes it is.

To determine whether B's utterance in either case above functions as an acceptance or a rejection, it is critical to not only look beyond the utterance itself and take into account the proposal under discussion (A's utterance), but also to specify (a) the polarity (positive *vs.* negative) of both the proposal and the response, and (b) how these polarities interact to give rise to a particular interpretation. Our aim in this paper is to develop a model of how logical polarity influences acceptance/rejection interpretation, inspired by recent work on the semantics of negation and polarity particles (Cooper and Ginzburg, 2011; Cooper and Ginzburg, 2012; Farkas and Roelofsen, 2013), and to test it on annotated data from two spoken dialogue corpora: the Switchboard Corpus (Godfrey et al., 1992) and the AMI Meeting Corpus (Carletta, 2007).

In the next section, we give an overview of related computational work on acceptance/rejection detection. In Section 3, we first briefly review recent formal semantics approaches to polarity and then present our model of logical polarity in acceptance and rejection moves. Section 4 describes

Proceedings of the SIGDIAL 2014 Conference, pages 151–160,
Philadelphia, U.S.A., 18-20 June 2014.

our experiments: We derive machine learning features from our polarity theory and test them in Switchboard and AMI datasets, achieving competitive F-scores of around 60 on the task of retrieving rejections. We conclude in Section 5 with a discussion of our results.

2 Related Computational Work

The first attempts to automatically identify acceptances and rejections (often referred to as agreements and disagreements) were carried out in the context of multiparty meetings for the purpose of dialogue summarisation tasks. Hillard et al. (2003) and Hahn et al. (2006) used the ICSI Meeting Corpus (Janin et al., 2003) to develop systems that would classify utterances into agreements, disagreements, backchannels, and 'other'. While these authors only leveraged lexical and prosodic features of the utterance to be classified (i.e., *local* features), Galley et al. (2004) showed that accuracy could be improved by taking into account contextual dependencies, in particular previous (dis)agreements between the dialogue participants, achieving an overall accuracy of 86.9%. Subsequent work built on Galley et al.'s approach showed that detecting agreement acts helped to identify public commitments to tasks (Purver et al., 2007) and other decisions made in a meeting (Fernández et al., 2008).

A difficulty shared by all approaches mentioned above is the skewness of the data, not only regarding (dis)agreement *vs.* other types of acts, but also agreement *vs.* disagreement. In the dialogue settings considered, acceptance/agreement is much more common than rejection/disagreement (e.g., 11.9% *vs.* 6.7% in the portion of the ICSI Meeting Corpus used by Galley et al. (2004) and 3.6% *vs.* 0.4% in the section of the AMI Meeting Corpus used by Germesin and Wilson (2009)). This can lead to reasonable overall accuracy but poor results on recognising rejections. Indeed, Germesin and Wilson (2009), who apply an approach based on Galley et al. (2004) to the AMI Meeting Corpus, achieve 98.1% accuracy, but report 0% recall for rejections/disagreements. Wang et al. (2011), who also work with AMI data, use different resampling methods to balance their dataset and then apply Conditional Random Fields (using therefore contextual information from sequences of utterances), achieving 56.9% recall and 55.9 F1 for disagreement detection.

Some recent work has moved away from spoken dialogue to address similar tasks in online discussion forums. An advantage of this kind of scenarios is that they seem to offer more opportunity for disagreement/rejection, thereby yielding more inherently balanced datasets. Abbott et al. (2011) and Misra and Walker (2013) use the Internet Argument Corpus (Walker et al., 2012), an annotated collection of posts in discussion forums with a balanced distribution of agreeing and disagreeing posts. They address a 2-way classification task—determining whether each response to a post (or to a quoted portion of a post in the case of Abbott et al. (2011)) is either an agreement or a disagreement—using a collection of features inspired by previous computational and theoretical approaches. The system developed by Misra and Walker (2013) uses only local features of the to-be-classified post, achieving an accuracy of 66% (over a 50% baseline). Abbott et al. (2011)'s best system uses features from both the quoted post and the response post, achieving an accuracy of 68.2%. However adding this contextual information does not significantly outperform a system based only on local features of the response, which yields 66.6% accuracy. Using both features from the post and the post response, Yin et al. (2012) obtain similar results: 68% accuracy on a different online corpus (the Political Forum), where the datasets are not balanced (they report a ratio of about 2 to 1 for agreement *vs.* disagreement).

All in all, this body of work has identified several linguistic features that are useful for inferring acceptances and rejections, often building on observations made by conversational analysts (Pomerantz, 1984; Brown and Levinson, 1987). Furthermore, recent work by Bousmalis et al. (2013) suggests that there are specific non-verbal behaviours associated with agreement and disagreement, such as different types of head, lip, and hand movements. However, to our knowledge, the role of logical polarity has not been investigated in any detail by computational approaches. Several systems make use of *subjective* polarity, i.e., sentiment. For instance, Galley et al. (2004) use the list of subjective adjectives compiled by Hatzivassiloglou and McKeown (1997) to assign a positive and a negative polarity value to an utterance given the number of subjective positive/negative adjectives it contains. Similarly, Misra and Walker (2013) use the MPQA Subjec-

tivity Lexicon (Wilson et al., 2005) to capture the local sentiment of an online post response given the number of words in the response with strongly subjective positive/negative polarity according to the subjectivity lexicon. Yin et al. (2012) assign a positive and a negative score to a post by aggregating the sentiment scores of those words that can be found in SentiWordNet (Baccianella et al., 2010).

Although subjective polarity may be helpful (e.g., utterances with a high positive sentiment score may be more likely to be acceptances), this is not the kind of polarity that concerns us in this paper. Note, furthermore, that local sentiment information may be superseded by logical polarity.

(4) A: But then it wouldn't sit as comfortably in your hand.
B: It would still be comfortable.

Despite the fact that B's utterance in (4)—extracted from the AMI corpus—would be assigned a positive sentiment score (given the presence of the word *"comfortable"*, classified as positive in the MPQA Subjectivity Lexicon, and the absence of negative subjective words), the utterance acts as a rejection due to logical polarity constraints, as we shall make clear in the next section.

3 Polarity in Acceptances and Rejections

In this section, we first give a brief overview of some of the main ideas put forward in recent theoretical approaches to polarity. Afterwards, we introduce our approach to logical polarity in the context of acceptance and rejection moves.

3.1 Formal Semantics Approaches

Polarity and in particular negation are central concepts in formal semantics and pragmatics (Horn, 1989). Recent work independently put forward within the frameworks of Type Theory with Records (Cooper and Ginzburg, 2011; 2012) and of Inquisitive Semantics (Farkas and Roelofsen, 2013) has proposed to semantically distinguish between *positive* and *negative* propositions. Such a proposal departs from the traditional view in formal semantics where propositions are taken to denote sets of possible worlds (see Partee (1989) for a survey). According to this traditional view, the meaning of (5a) would be indistinguishable from that of (5b), given that the two propositions are true in exactly the same possible worlds:

(5) a. Sue failed the exam.
b. Sue didn't pass the exam.

These utterances, however, license different types of responses. For instance, responding *"no"* to (5a) would assert that Sue did pass the exam, while the same response to (5b) would typically be understood as asserting the opposite. Leaving aside many details that distinguish the two theories, Cooper/Ginzburg and Farkas/Roelofsen propose that polarity particles—words like *"yes"* and *"no"*—are sensitive to the polarity of their antecedent: *"yes"* presupposes that a positive proposition is under discussion, while *"no"* presupposes a negative proposition. If the presupposition is met, both *"yes"* and *"no"* assert the proposition under discussion (i.e., in our terms, they act as *acceptances*); if the presupposition fails, they assert the negation of the proposition under discussion (i.e., they act as *rejections*).

This characterises the standard behaviour of polarity particles. However, the picture is slightly more complicated since, when the proposition under discussion is negative, in English *"yes"* and *"no"* can also be used to agree or disagree, respectively (contrary to the standard case):

(6) Sue didn't pass the exam.
a. No (she didn't). $\rightsquigarrow$ *standard acceptance*
Yes, she didn't.
b. Yes, she did. / #Yes. $\rightsquigarrow$ *standard rejection*
No, she did.

According to Farkas and Roelofsen (2013), this ambiguity of use makes bare forms of *"yes"*/*"no"* less likely in the non-standard cases exemplified in (6) and favours more explicit sentential forms where the presence of the verb disambiguates the intended interpretation. In this respect, however, the standard rejection in (6b) constitutes a special case: While in standard acceptances the sentential form is not required, in standard rejections it seems needed. According to these authors, in English the positive polarity particle *"yes"* has a strong preference for realising an agreement move and therefore its use as a rejection is *marked*.[1] This makes the explicit sentential form *"Yes, she did"* in (6b) more felicitous than the bare form *"Yes"*. Thus, the two types of rejections to a negative proposition we see in (6b)—with *"yes"* and

[1] The special status of English *"yes"* for rejection seems to be supported by cross-linguistic evidence. For instance, German has a special positive polarity particle *"doch"* for rejecting a negative proposition: in response to the assertion in (6), *"doch"* would be used to disagree (*"yes, she did"*) while *"ja"* would be used to agree (*"yes, she didn't"*).

Polarity of P-R	Type	Example from AMI Meeting Corpus
positive – positive	default relative acceptance	A: *And then you can buy the covers.* B: *Yes*
negative– negative	reverse relative acceptance	A: *It's not very well advertised.* B: *No, it's not.*
positive – negative	default relative rejection	A: *It's a frog.* B: *No, it's a turtle.*
negative– positive	reverse relative rejection	A: *TVs aren't capable of sending.* B: *Yes, they are.*

Table 1: Relative response types.

"no"—are expected to contain an explicit verbal constituent. We refer to this as the *markedness expectation*.

3.2 Our Model

Our aim is to exploit insights from the theories sketched above to develop a model that can be operationalised in a computational setting to test whether information regarding logical polarity can contribute to automatically distinguish acceptances from rejections.

We focus on proposal-response pairs (P-R), where R either accepts or rejects P. We propose to assign both the proposal and the response a logical polarity: either positive or negative. Furthermore, we differentiate *absolute* (polarity independent) from *relative* (polarity independent) responses. A response type R is absolute if its acceptance/rejection function does not depend on the polarity of P, and it is relative if it does. Formally, we say that a proposal P is rejected by a response R if $P \wedge R$ is inconsistent. This gives us the following four possible responses to P:

- $R \equiv \top$: absolute acceptance
- $R \equiv \bot$: absolute rejection
- $R \equiv P$: relative acceptance
- $R \equiv \neg P$: relative rejection

Our focus of attention is on relative responses. Given a P-R pair with a relative response, we infer an acceptance if the polarities of P and R align, and a rejection if the polarities differ. This gives us four possible relative responses, shown in Table 1. In the default cases, where P is positive, positive responses act as acceptances and negative responses as rejections—exactly as absolute response types would act. When P is negative (i.e., $P \equiv \neg P'$), we are faced with what we call *reverse* relative responses: Negative polarity responses act as acceptances and positive polarity responses as rejections. An acceptance can have the form $R \equiv P \equiv \neg P'$ while a rejection can have the form $R \equiv \neg P \equiv P'$ (with R being positive, i.e., with the double negation $\neg\neg P'$ eliminated). We call these cases reverse responses because their polarity signature is precisely the negation of the respective default cases (*cf.* Table 1).

The next obvious question to address is how the polarity of proposals and responses can be determined. Clearly, this will differ across languages. For the case of English, we shall assume that polarity is linked to the presence of particular particles and grammatical indicators. In particular, we consider the words in Table 2 to be positive and negative polarity markers.[2] Amongst negative polarity markers, we distinguish between negative polarity particles and negation indicators.

positive	particles:	*yes, yeah, yep*
negative	particles:	*no, nope, nah*
	negation:	*not, -n't, never, nothing, nobody, nowhere*

Table 2: Polarity markers.

All markers in Table 2 are key cues of polarity. However, they do not straightforwardly determine the polarity of a contribution. Firstly, there are cases where the presence of a marker does not have the expected effect on polarity. For instance, a negative tag question (*"isn't it?"*) at the end of an utterance does not mark that utterance as negative. Also, the polarity effect of a marker can be invalidated if it is followed by the contrast connective *"but"*. For instance, in the following AMI examples, *"but"* cancels out the effect of the negative polarity particle *"no"* in (7), making B's utterance positive, and the effect of the positive polarity particle *"yeah"* in (8), making B's utterance negative (in conjunction with the verbal negation in this case):

(7) *Reverse rejection: negative–positive*
A: Yes, but some televisions don't support it.
B: No, *but* then they would also support that button, because it's the same thing.

(8) *Default rejection: positive–negative*
A: Yeah, uh materials like wood that
B: Yeah, *but* wood is not a not a material you which you build a a remote control of .

[2] We do not claim that this list is exhaustive.

Secondly, it is important to take into account that a large amount of acceptances and rejections do not include any marker of polarity at all. For instance, in our datasets extracted from the AMI and Switchboard (SWB) corpora (which we will describe in detail in Section 4.1), 49% and 70% of acceptances in AMI and SWB, respectively, do not contain any explicit polarity marker; and similarly for 40% (AMI) and 15% (SWB) of rejections. In part this is due to the fact that in English (as in most languages) there is no morphologically-realised positive counterpart of verbal negation.

Given the observations above, we adopt the heuristics in Figure 1 to assign a polarity to P and R. Since this heuristics is intended to be applicable to dialogue corpora, we forgo the use of deep semantic analysis, which is difficult to achieve when dealing with naturally occurring spoken language.[3]

P-**polarity:** A proposal P has negative polarity if it contains a negation indicator (excluding tag questions); otherwise, P has positive polarity.

R-**polarity:** We define a precedence order on polarity markers: negative polarity particles take precedence over positive polarity particles, which in turn take precedence over negation indicators.

- If a response R contains a negative polarity particle (not followed by *"but"*), its polarity is negative.
- Else, if R contains a positive polarity marker (not followed by *"but"*), its polarity is positive.
- Else, if R contains a negation indicator, its polarity is negative.
- Otherwise, R has positive polarity.

Figure 1: Heuristics for polarity determination.

Drawing on the notion of *markedness expectation* we introduced at the end of Section 3.1, we hypothesise that the lack of explicit positive polarity markers will be compensated for by the presence of sentential similarity patterns between proposals and responses. It follows from our description of relative responses (see Table 1) that they will either semantically mirror the proposal (acceptances) or negate it (rejections). In the absence of an explicit positive polarity marker in the proposal or the response, therefore, we expect to find some form of sentential parallelism, potentially in both cases—when P-R polarities align, as in (9), and when they differ, as in (10):[4]

(9) A: It's still it's still working,
B: It is.

(10) A: It's a fat cat.
B: It is not a fat cat.

According to the markedness expectation, this type of parallelism is expected in reverse relative responses even when polarity particles are present as in (11) from Switchboard and in the reverse response examples in Table 1. Hence, we conjecture that parallelism will be present with higher frequency in the reverse cases.

(11) A: They wouldn't be able to own a house.
B: Yes, they would.

4 Experiments

In order to automatically test the extent to which logical polarity plays a role in determining the function of naturally occurring acceptances and rejections, we conduct machine learning experiments on dialogue corpus data. We first explain how we create our dataset, then describe how we devise features that encode polarity information, and finally report the results obtained.

4.1 Datasets

We test our model on two different corpora: The Switchboard Corpus (SWB) (Godfrey et al., 1992) and the AMI Meeting Corpus (Carletta, 2007). SWD is a collection of around 2400 recorded and transcribed telephone conversations between two dialogue participants. The speakers are provided with a topic and then converse freely. In contrast, AMI contains transcriptions from around 100 hours of recorded multiparty conversations amongst four dialogue participants who interact face-to-face in a meeting setting. The speakers converse freely, but they play roles (such as industrial designer or project manager) in a fictitious design team whose goal is to design a remote control. Therefore the dialogue is mildly task-oriented. Both corpora have been annotated with dialogue acts (DAs), albeit with slightly different DA annotation schemes: SWD is annotated with the SWBD-DAMSL tagset (Jurafsky et al., 1997), while AMI uses a coarser-grained tagset but includes relations between some DAs (loosely called *adjacency pair* annotations).[5]

[3] Amongst other things, this means we do not account for the scope of negation.

[4] Both examples are extracted from the AMI corpus.

[5] The AMI DA annotation manual is available at `http://mmm.idiap.ch/private/ami/annotation/dialogue_acts_manual_1.0.pdf`

	acceptances	rejections	total P-R
SWB	4534 (97%)	145 (3%)	4679
AMI	7405 (91%)	697 (9%)	8102

Table 3: Class distribution in our datasets.

We use the DA annotations to extract a dataset of proposal-response (P-R) pairs for each corpus as follows. To construct the SWB dataset, we extract all utterances u annotated as `Agree/Accept` or `Reject` that are turn-initial and that are immediately preceded by a turn whose last utterance u' is annotated as `Statement-non-opinion`, `Statement-opinion` or `Summarize/Reformulate`. To construct the AMI dataset, we extract all utterances u annotated as `Assessment` that are turn initial and that are linked with the relations `Support/Positive Assessment` or `Objection/Negative Assessment` to an earlier utterance u' that is not annotated as `Elicit Inform` or `Elicit Assessment` (i.e., that is not a question). In both cases, P corresponds to u' and R to the first five words of u. We consider R an acceptance if u is annotated as `Agree/Accept` in SWB or as `Support/Positive Assessment` in AMI, and a rejection if it is annotated as `Reject` in SWB or as `Objection/Negative Assessment` in AMI.

We take the first five words of a turn-initial utterance to be the most relevant ones for conveying acceptance or rejection. This is motivated by the fact that dialogue participants typically provide evidence of understanding—and, by extension, of agreement or disagreement—at the earliest opportunity in order to avoid misunderstandings on what they take to be common ground (Pomerantz, 1984; Clark, 1996). However, when extracting our P-R pairs we retain the entire utterance u (of which R is a prefix) in order to be able to take its length into account in the automatic classification experiment, as explained in the next section.

Finally, we observe that in the two corpora all the P-R pairs where R is just a single *"yeah"* are acceptances. Thus, in the terminology we introduced in Section 3.2, bare *"yeah"* seems to be an *absolute* response type, whose acceptance function is independent of the polarity of P (in contrast to the relative response types in Table 1). Since identification of these acceptances is trivial, we discard them from our datasets. The final distribution of acceptances and rejections in each of the datasets is shown in Table 3. As can be seen, the data is highly skewed, with less than 10% of P-R pairs corresponding to rejections.

4.2 Features

We derive different types of features to test our model. We are not interested in using large amounts of unmotivated features, but rather in exploiting a small set of meaningful domain- and setting-independent features that can help us to investigate the impact of logical polarity. The feature we use are summarised in Figure 2. We consider several local features of the response. Most of these features are inspired by earlier approaches reviewed in Section 2, such as those by Galley et al. (2004) and Misra and Walker (2013). We use several lexical features that act as cues for acceptance or rejection. For instance, the presence of *"yeah"* is a good cue for acceptance, while the presence of *"but"* is a strong cue for rejection. The bigram *"yeah, but"* is in turn a good indicator for rejection—the *"yeah"* in such cases seems to be an attempt at politeness (Brown and Levinson, 1987; Bousfield, 2008). Since rejections are dispreferred moves, they are frequently initiated with a hedging such as *"well"* or with hesitation or stalling (Byron and Heeman, 1997). These utterance-initial cues are aggregated into one feature. Rejections also tend to be longer than acceptances since the speaker feels the need to justify the unexpected move (Pomerantz, 1984). We take into account the length of the entire utterance containing R with three binary features.[6] We also consider less frequent semantic indicators for acceptance and rejection, respectively, which we group into two aggregate features that record the presence of agreement words such as *"okay"* or *"correct"* and contrast words such as *"however"* or *"although"*. Given our observations regarding polarity and polarity particles in Section 3, in contrast to previous approaches we don't include *"yes"* and *"no"* as local lexical cues. Instead, we add a new local feature encoding the polarity of the response as determined by the R-polarity heuristics in Figure 1.[7]

[6]The use of Boolean features here is motivated by our choice of classifier, as we point out in the next subsection. The length thresholds have been set up manually after qualitative examination of several examples.

[7]We have tested other theoretically motivated local features, such as turn-length and number of disfluencies. The local R features in Fig. 2 combined with the local R polarity feature correspond to the best performing local feature set.

Local features cannot capture the most interesting aspects of logical polarity, which originate from the interaction between the polarities of the proposal and the response in relative response types. To account for this, we introduce four relative P-R polarity features corresponding to the response types described in Table 1. Finally, we introduce a feature that records the presence of some form of parallelism between P and R. As we mentioned at the end of Section 3.2, the markedness expectation predicts that sentential parallelism will occur more frequently in reverse relative responses, i.e., responses to negative proposals. The parallelism feature targets such cases. We restrict ourselves to strict identity between P and R of a pronominal subject and a verb (in negative *vs.* positive form).[8] The feature therefore is only able to capture examples such as (12a) but not (12b), where anaphora resolution would be required.[9]

(12) a. A: But it wouldn't be very attractive.
B: No, it would.
b. A: TVs aren't capable of sending.
B: Yes, they are.

4.3 Results

We conducted the machine learning experiment using `BernoulliNB`, the Bernoulli-distributed Naive Bayesian classifier from *scikit-learn* (Pedregosa et al., 2011), which outperformed several other classifiers, including Random Forests and a Support Vector Machine. We chose this classifier because our main features—the relative polarities—are Boolean and our data is highly imbalanced.[10] Given the high relative frequency of acceptances over rejections in our datasets (see Table 3), measuring accuracy or retrieving acceptances would yield very good results. Hence, as discussed in section 2, we believe that the most discerning task is the retrieval of rejections. Precision, recall and F-scores for this task, with the classifier trained on different combinations of feature sets, are shown in Table 4. We developed the classifier on the whole AMI dataset, as the small number of rejections makes splitting up the corpus into a development and a test set infeasible. The SWB corpus was exclusively used for testing. In the AMI dataset we tested the classifier with 10-fold cross-validation and in the SWB dataset with 5-fold cross-validation, due to the more limited amount of rejections in this corpus. Also, due to the lack of training data, the more specific Relative P-R Parallelism feature could not be applied to the SWB corpus.

[8] We use the NLTK POS tagger to implement this feature (Bird et al., 2009).

[9] Given the high frequency of pronominal forms in spoken dialogue, pronoun identity turns out to be reasonably useful.

[10] The scikit-learn documentation indicates that this classifier is particularly suited for sparse data and Boolean features.

LOCAL R FEATURES

Length of utterance containing R in number of words:

- Three features: $l > 2$, $l > 12$, $l > 24$

Acceptance Indicators:

- R contains *yeah*
- R contains any of *absolutely, okay, accept, agree, correct, either, true, sure,* not preceded by *not*

Rejection Indicators:

- R contains *but*
- R contains the bigram *'yeah, but'*
- R starts with any of *well, oh, uh, mm*
- R contains any of *actually, however, though, although*

LOCAL R POLARITY FEATURE

- *positive* or *negative*, according to R-polarity in Fig. 1

RELATIVE P-R POLARITY FEATURES (*cf.* Fig. 1)

- *positive–positive*
- *positive–negative*
- *negative–negative*
- *negative–positive*

RELATIVE P-R PARALLELISM FEATURE

One of the following patterns appears in P-R, where a pronoun p, an auxiliary verb *aux* and a main verb v are identical in P and R:

- *'p aux not'* – *'p aux'* not followed by {*n't*| *not*}
- *'p (aux) not v'* – *'p v'*
- *'I do*{*n't*| *not*} {*think*|*know*} {*that*|*if*} *p aux'* – *'p aux'* not followed by {*n't*| *not*}

Figure 2: Feature types (all features are Boolean).

For comparison we report the results of a simple unigram baseline: Each content word that occurs at least 5 times in the dataset is used as a Boolean feature (occurrence *vs.* non-occurrence). This achieves F-scores of 31.66 in AMI and 16.63 in SWB. As a more substantial baseline we consider a system that uses only local features of the response, including local polarity. This feature-set is expected to capture relatively well the accepting/rejecting function of absolute responses and default relative responses, since their function aligns with their local polarity. This yields an F-score of 52.24 in AMI and of 33 in SWB. The Relative Polarity features were conceived to reduce classification confusion grounded in reverse polarity: If only local features are considered, a reverse polarity acceptance would appear to be a rejection,

Feature sets	AMI Precision	AMI Recall	AMI F1	SWB Precision	SWB Recall	SWB F1
Unigrams	35.61%	28.97%	31.66	24.20%	12.93%	16.63
Local + Local Polarity	44.13%	64.12%	52.24	20.80%	82.46%	33.00
Local + Relative Polarity	58.08%	61.63%	59.75	49.12%	72.93%	58.49
Local + Relative Pol. + Parallelism	58.23%	64.04%	60.96	n/a	n/a	n/a

Table 4: Precision, Recall, and F-scores for rejection identification.

while a reverse polarity rejection would seem to be an acceptance. Moving from local to relative polarity features should therefore reduce this confusion. Indeed, in both corpora the precision is increased substantially (from 44.13% to 58.08% in AMI and from 20.8% to 49.12% in SWB), causing a great increase in F-scores: 59.75 in AMI and 58.49 in SWB (paired t-tests show all these increases are significant, with $p < 0.001$).

However, in both datasets we observe a reduction in recall when moving from local to relative polarity. We believe that this is in part due to the relative polarity features ignoring some absolute uses of polarity particles, which may have been captured by Local Polarity.[11] The Relative Parallelism feature should be able to help in such cases. For instance, in example (12a) B's utterance would be assigned negative polarity and therefore the relative polarity features would contribute to classify it as an acceptance (since in the large majority of cases negative-negative P-R pairs do correspond to acceptances). In this case, however, *"no"* is used absolutely, i.e., as a rejection. Due to the markedness expectation, this is likely to show up in the form of contrastive parallelism, which we can—at least in part—capture with our simple feature. Indeed, adding this feature to the AMI dataset raises recall back to baseline level: 64.04% *vs.* 61.63% ($p < 0.005$). This, in turn, increase the AMI F-score from 59.75 to 60.96 ($p < 0.05$).

5 Conclusions

The overall aim of this paper has been to investigate the influence of logical polarity in interpreting utterances as acceptance or rejection moves in dialogue. We have built on recent work on the semantics of negation and polarity particles by Cooper and Ginzburg (2011; 2012) and Farkas and Roelofsen (2013) to develop an approach to polarity that is theoretically motivated and that can be computationally tested on corpus data. Although there is a substantial amount of previous work on automatically detecting agreement and disagreement in dialogue corpora, to our knowledge the role of logical polarity had not been explicitly investigated before.

Our focus has been on relative responses, i.e., responses where simply taking into account clues from the utterance to be classified is insufficient—or can even be misleading—to infer acceptance or rejection. We have argued that relative responses require taking into account how the polarities of the response and of the current proposal under discussion interact, and have put forward a model that captures such interaction. Our experiments show that the use of information on relative polarity substantially helps to distinguish acceptances from rejections. This indicates, on the one hand, that our model does a reasonably good job at capturing this phenomenon, and on the other hand, that relative polarity responses are not merely a theoretically interesting phenomenon but are in fact widespread in actual dialogue.

There is certainly room for improving the implementation of our heuristics, for instance by using finer-grained semantic and syntactic information: e.g., we cannot currently capture acceptance/rejection of a subclause, implicature rejections, rhetorical questions, nor sarcasm—all of which affect the recall of our system. Interestingly, the classification experiments yield very similar results in the two corpora with the Local + Relative Polarity feature set—F-scores of 59.75 in AMI and 58.49 in SWB. This indicates that our theoretical observations are applicable independently of setting, domain and number of speakers. There seem to be some differences across the two corpora, however, since the impact of relative polarity information is much higher in SWB than in AMI (the F-score goes up around 7 in AMI when moving from local to relative polarity, while in SWB it increases by 25). A deeper investigation into the shortcomings of our implemented model and of where these shortcomings affect AMI differently than SWB are issues we leave for future work.

[11] We note that the featureset Local + Local Polarity + Relative Polarity does not outperform Local + Local Polarity in the classification experiments. We believe this indicates that polarity is indeed mostly a contextual phenomenon.

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.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *Proceedings of LREC*, pages 2200–2204.

Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

Derek Bousfield. 2008. *Impoliteness in Interaction*. John Benjamins.

Konstantinos Bousmalis, Marc Mehu, and Maja Pantic. 2013. Towards the Automatic Detection of Spontaneous Agreement and Disagreement Based on Nonverbal Behaviour: A Survey of Related Cues, Databases, and Tools. *Image Vision Computing*, 31(2):203–221.

Penelope Brown and Stephen Levinson. 1987. *Politeness: Some universals in language usage*. Cambridge University Press.

Donna K. Byron and Peter A. Heeman. 1997. Discourse marker use in task-oriented spoken dialog. In *Proceedings of Eurospeech*, pages 2223–2226.

Jean Carletta. 2007. Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus. *Language Resources and Evaluation*, 41(2):181–190.

Herbert H. Clark and Edward F. Schaefer. 1989. Contributing to discourse. *Cognitive Science*, 13(2):259–294.

Herbert H. Clark. 1996. *Using language*. Cambridge University Press.

Robin Cooper and Jonathan Ginzburg. 2011. Negation in dialogue. In *Proceedings of the 15th SemDial Workshop (Los Angelogue)*, pages 130–139.

Robin Cooper and Jonathan Ginzburg. 2012. Negative inquisitiveness and alternatives-based negation. In *Logic, Language and Meaning: Proceedings of the 18th Amsterdam Colloquium*, Lecture Notes in Computer Science, pages 32–41. Springer.

Donka Farkas and Floris Roelofsen. 2013. Polar initiatives and polar particle responses in an inquisitive discourse model. Available from `http://www.illc.uva.nl/inquisitivesemantics/`.

Raquel Fernández, Matthew Frampton, Patrick Ehlen, Matthew Purver, and Stanley Peters. 2008. Modelling and Detecting Decisions in Multi-Party Dialogue. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pages 156–163.

Michel Galley, Kathleen McKeown, 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 42nd Meeting of the Association for Computational Linguistics (ACL'04)*, pages 669–676.

Sebastian Germesin and Theresa Wilson. 2009. Agreement detection in multiparty conversation. In *Proceedings of the 2009 international conference on Multimodal interfaces*, pages 7–14. ACM.

John J. Godfrey, E. C. Holliman, and J. McDaniel. 1992. SWITCHBOARD: Telephone Speech Corpus for Research and Development. *IEEE Conference on Acoustics, Speech, and Signal Processing*, 1:517–520.

Sangyun Hahn, Richard Ladner, and Mari Ostendorf. 2006. Agreement/disagreement classification: Exploiting unlabeled data using contrast classifiers. In *Proceedings of the HLT-NAACL*, pages 53–56.

Vasileios Hatzivassiloglou and Kathleen R McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 174–181.

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 2003*, pages 34–36.

Julia L. B. Hirschberg. 1985. *A theory of scalar implicature*. Ph.D. thesis, University of Pennsylvania.

Laurence R. Horn. 1989. *A Natural History of Negation*. University of Chicago Press.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elisabeth Shriberg, Andreas Stolcke, and Chuck Wooters. 2003. The ICSI Meeting Corpus. In *Proceedings of ICASSP'03*, pages 364–367.

Dan Jurafsky, Elizabeth Shriberg, and Debra Biasca. 1997. Switchboard SWBD-DAMSL shallow-discourse-function-annotation coder's manual, draft 13. Technical Report TR 97-02, Institute for Cognitive Science, University of Colorado at Boulder.

Alex Lascarides and Nicholas Asher. 2009. Agreement, disputes and commitments in dialogue. *Journal of Semantics*, 26(2):109–158.

Amita Misra and Marilyn Walker. 2013. Topic independent identification of agreement and disagreement in social media dialogue. In *Proceedings of the SIGDIAL 2013 Conference*, pages 41–50, Metz, France. Association for Computational Linguistics.

Barbara Partee. 1989. Possible worlds in model-theoretic semantics: A linguistic perspective. In S. Allen, editor, *Possible Worlds in Humanities, Arts and Sciences*, pages 93–123. Walter de Gruyter.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Anita Pomerantz. 1984. Agreeing and disagreeing with assessments: Some features of preferred/dispreferred turn shapes. In *Structures of Social Action*. Cambridge University Press.

Matthew Purver, John Dowding, John Niekrasz, Patrick Ehlen, Sharareh Noorbaloochi, and Stanley Peters. 2007. Detecting and summarizing action items in multi-party dialogue. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 18–25.

Robert Stalnaker. 1978. Assertion. In P. Cole, editor, *Pragmatics*, volume 9 of *Syntax and Semantics*, pages 315–332. New York Academic Press.

Marilyn A Walker, Jean E Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *LREC*, pages 812–817.

Marilyn A. Walker. 1996. Inferring acceptance and rejection in dialog by default rules of inference. *Language and Speech*, 39(2-3):265–304.

Wen Wang, Sibel Yaman, Kristin Precoda, Colleen Richey, and Geoffrey Raymond. 2011. Detection of agreement and disagreement in broadcast conversations. In *Proceedings of ACL*, pages 374–378.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of HLT-EMNLP*, pages 347–354.

Jie Yin, Paul Thomas, Nalin Narang, and Cecile Paris. 2012. Unifying Local and Global Agreement and Disagreement Classification in Online Debates. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, pages 61–69.