Automatically Identifying Agreement and Disagreement in Speech

Rik Koncel-Kedziorski, Andrea Kahn, Claire Jaja

this slide left intentionally blank


"Welcome to
the Old West"

A little vocabulary

Spurts: periods of speech with no pauses greater than ½ second

Adjacency Pairs:

- fundamental units of conversational organization
- two parts (A and B) produced by different speakers
- Part A makes B immediately relevant
- Need not be directly adjacent

Problem Overview

multiple facets of the same problem:

- identifying adjacency pairs
- identifying contentious spots (“hot spots”) where participants are highly involved
- identifying agreement vs. disagreement (i.e. labeling spurts as agreement or disagreement)

Challenges

- automatic speech recognition errors
- agreement or disagreement not always clear, even to humans

Dataset

International Computer Science Institute (ICSI) Meeting corpus:

- collection of 75 naturally occurring, weekly meetings of research teams
- ~1 hour each
- average 6.5 participants

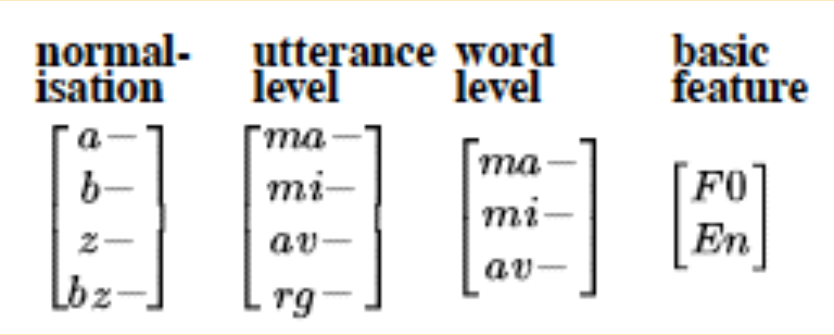
Features

- Acoustic
- Text
- Context

Acoustic Features

- Types:
  - Mean and variance of F0
  - Mean and variance of energy
  - Mean and maximum vowel duration
  - Mean, maximum, and initial pause
  - Duration of overlap of two speakers
- Levels (for F0 and energy features):
  - Utterance-level
  - Word-level
- Normalization schemes:
  - Absolute (no normalization)
  - b-, z-, or bz- normalization

Acoustic Features: An Example Approach



From Wrede & Shriberg (2003b).

Structure of acoustic/prosodic features used for identifying speaker involvement

Acoustic Features: An Example Approach

1 bz-ma-av-F0	13 b-av-F0	25 bz-rg-F0	37 a-ma-F0
2 z-ma-av-F0	14 z-av-En	26 a-av-En	38 a-av-ma-F0
3 z-av-ma-F0	15 z-av-mi-F0	27 a-ma-En	39 a-av-F0
4 bz-av-F0	16 z-av-ma-En	28 a-rg-En	40 a-mi-En
5 bz-av-ma-F0	17 b-av-mi-F0	29 a-ma-av-En	41 b-mi-F0
6 z-ma-F0	18 z-ma-av-En	30 a-av-mi-En	42 z-mi-En
7 z-av-F0	19 b-rg-F0	31 b-mi-av-F0	43 a-rg-F0
8 bz-ma-F0	20 z-rg-En	32 z-mi-av-En	44 a-av-mi-F0
9 b-ma-av-F0	21 z-ma-En	33 a-mi-av-En	45 a-mi-av-F0
10 bz-av-mi-F0	22 z-av-mi-En	34 z-rg-F0	46 z-mi-av-F0
11 b-av-ma-F0	23 a-av-ma-En	35 bz-mi-F0	47 a-mi-F0
12 b-ma-F0	24 bz-mi-av-F0	36 a-ma-av-F0	48 z-mi-F0

From Wrede & Shriberg (2003b).

Features sorted according to the difference between the means of involved vs. uninvolved speakers

Text Features

structural

relate to structure of utterances, mostly used for AP identification

- # of speakers between A and B
- # of spurts between A and B
- # of spurts of speaker B between A and B
- do A and B overlap?
- is previous/next spurt of same speaker?
- is previous/next spurt involving same B speaker?

Text Features

lexical

counts

- # of words
- # of content words
- # of positive/negative polarity words
- # of instances of each cue word
- # of instances of each cue phrase and agreement/disagreement token

Text Features

lexical

pairs

- ratio of words in A also in B (and vice versa)
- ratio of content words in A also in B (and vice versa)
- # of n-grams in both A and B
- does A contain first/last name of B?

content

- first and last word
- class of first word based on keywords
- perplexity w/ respect to different language models (one for each class)

Context Features: Pragmatic Function

Whether B (dis)agrees with A is influenced by

- the previous statement in the discourse
- Whether B (dis)agreed with A recently
- Whether A (dis)agreed with B recently
- Whether B (dis)agreed recently with some speaker X who (dis)agrees with A

Context Features: Empirical Result

	$p(c_i \mid c_{i-1})$	$p(c_i^{B\to A} \mid \text{pred}_{B\to A}(c_i^{B\to A}))$	$p(c_i^{B\to A} \mid \text{pred}_{A\to B}(c_i^{B\to A}))$
$p(\text{AGREE} \mid \text{AGREE})$	**.213**	**.250**	.175
$p(\text{OTHER} \mid \text{AGREE})$	.713	.643	.737
$p(\text{DISAGREE} \mid \text{AGREE})$	.073	.107	.088
$p(\text{AGREE} \mid \text{OTHER})$	.187	.115	.177
$p(\text{OTHER} \mid \text{OTHER})$	.714	.784	.710
$p(\text{DISAGREE} \mid \text{OTHER})$	.098	.100	.113
$p(\text{AGREE} \mid \text{DISAGREE})$	.139	.087	**.234**
$p(\text{OTHER} \mid \text{DISAGREE})$	.651	.652	.638
$p(\text{DISAGREE} \mid \text{DISAGREE})$	**.209**	**.261**	.128

From *Identifying agreement and disagreement in conversational speech: Use of bayesian networks to model pragmatic dependencies.*

Context Features: Empirical Result

	$p(c_k^{B\to A}\|c_i, c_j)$, where $c_j = \text{pred}_{B\to X}(c_k^{B\to A})$ and $c_i = \text{pred}_{X\to A}(c_j)$			
	c_i = AGREE c_j = AGREE	c_i = AGREE c_j = DISAGREE	c_i = DISAGREE c_j = AGREE	c_i = DISAGREE c_j = DISAGREE
$p(\text{AGREE}\|c_i, c_j)$	**.225**	.147	.131	.152
$p(\text{OTHER}\|c_i, c_j)$	.658	.677	.683	.668
$p(\text{DISAGREE}\|c_i, c_j)$	.117	**.177**	**.186**	**.180**

From *Identifying agreement and disagreement in conversational speech: Use of bayesian networks to model pragmatic dependencies.*

Spotting “Hot Spots”

Wrede, B. and Shriberg, E. (2003b). Spotting "hotspots" in meetings: Human judgments and prosodic cues. In Proceedings of Eurospeech, pages 2805-2808, Geneva.

problem: identifying features correlated with speaker involvement

features used: acoustic/prosodic features (mean and variance in F0 and energy)

Spotting “Hot Spots”: Approach

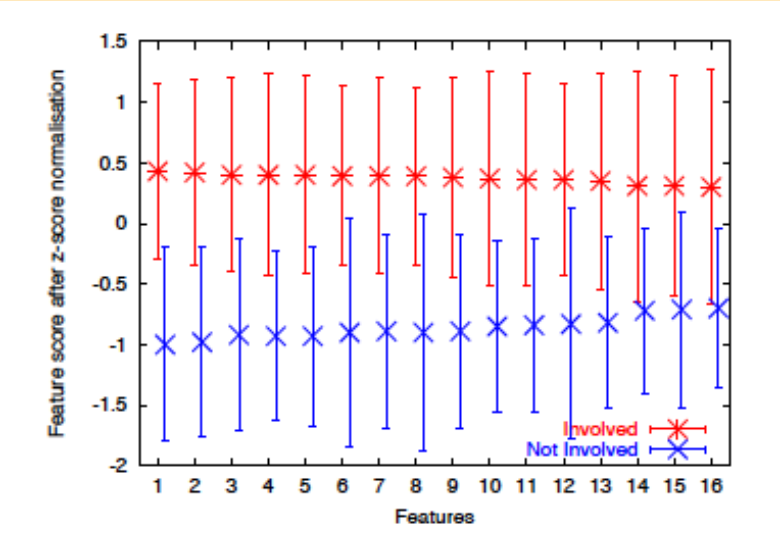
- Considered 88 utterances for which at least 3 ratings were available
- Gold label (involved vs. uninvolved) assigned was a weighted average of the ratings
- Sorted features according to their usefulness in determining speaker involvement
  - i.e., differences between the means of involved vs. uninvolved speakers

Spotting “Hot Spots”: Inter-annotator Agreement

- Utterances initially labeled as “involved: amused”, “involved: disagreeing”, “involved: other”, or “not particularly involved”
- Utterances were presented in isolation (no context)
- Used 9 raters who were familiar with the speakers

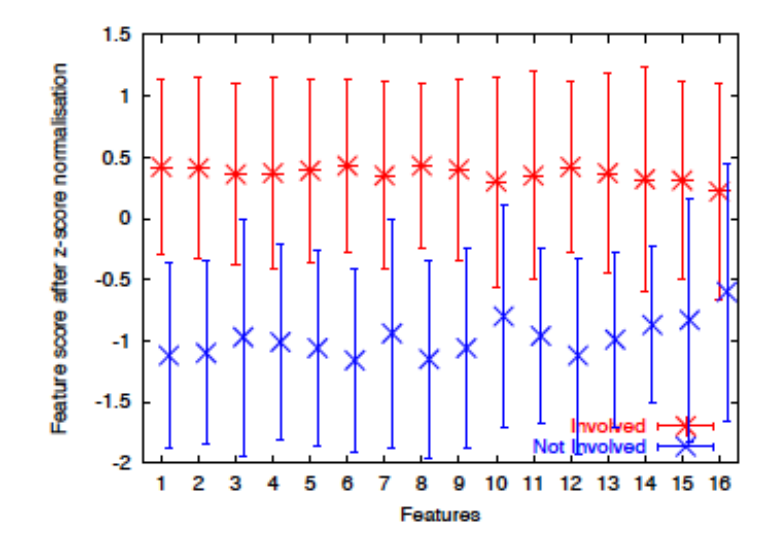
- Found that high and low pairwise kappa seemed to correlate with particular raters
  - i.e., some raters simply better at the task than others
- Found that native speakers had a higher pairwise kappa agreement

Spotting “Hot Spots”: Results



Mean and standard deviations of top 16 normalized features of all speakers rated as involved or not involved.

Spotting “Hot Spots”: Results



Mean and standard deviations of top 16 normalized features of one speaker* rated as involved or not involved.

*They don’t say how they selected this speaker. (Maybe results for other speakers don’t look as good.)

Spotting “Hot Spots”: Issues

- Really, a feature selection study: Ideally, they’d subsequently test these features on a different dataset and see what kinds of results they got

- Paper allegedly about “identifying hotspots”, but in actuality they’re just attempting to detect whether a particular utterance by a particular speaker is involved vs. uninvolved

- Despite the fact that they reported high agreement between annotators, they also identified sources of annotation discrepancy, highlighting the subjective nature of the task of labeling involvement

Detection of Agreement vs. Disagreement

Hillard, D., Ostendorf, M., and Shriberg, E. (2003). Detection of agreement vs. disagreement in meetings: Training with unlabeled data. In Proceedings of HLT-NAACL Conference, Edmonton, Canada.

problem: identifying agreement/disagreement

features: text (lexical), acoustic

Detection of Agreement vs. Disagreement

methodology: decision tree classifier

- 450 spurts x 4 meetings (1800 spurts total) hand-labeled as negative (disagreement), positive (agreement), backchannel, or other
- upsampled data for same number of training points per class
- iterative feature selection algorithm
- unsupervised clustering strategy for incorporating unlabeled data (8094 additional spurts)
    - first, heuristics, then, LM perplexity (iterated until no movement between groups), used as “truth” for training

Detection of Agreement vs. Disagreement

Features	Hand Transcriptions			ASR Transcriptions		
	Overall Accuracy	A/D Confusion	A/D Recovery	Overall Accuracy	A/D Confusion	A/D Recovery
Keywords	82%	2%	87%	61%	7%	53%
Hand Trained LM	71%	13%	74%	64%	10%	67%
Unsupervised LM	78%	10%	81%	67%	14%	70%
All word based	79%	8%	83%	71%	3%	78%

Table 1: Results for detection with different classifiers using word based features.

Transcripts Train/Test	Overall Accuracy	A/D Confusion	A/D Recovery
Hand/Hand	62%	17%	62%
Unsup./Hand	66%	13%	72%
Hand/ASR	62%	16%	61%
Unsup./ASR	64%	14%	75%

Table 2: Results for classifiers using prosodic features.

Detection of Agreement vs. Disagreement

Issues

- choice of labeling - label backchannel and agreement separately, but then merge for presenting 3-way classification accuracy
- unbalanced dataset (6% neg, 9% pos, 23% backchannel, 62% other) - upsampling may be extreme
- inter-annotator agreement not high (kappa coefficient of .6), not really discussed in paper
- report results on word-based and prosodic features separately - briefly mention no performance gain by

Identifying Agreement and Disagreement

Galley, M., McKeown, K., Hirschberg, J., and Shriberg, E. (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), Main Volume, pages 669-676, Barcelona, Spain.

Identifying Agreement and Disagreement

Problem: Determine whether the speaker of a spurt is agreeing, disagreeing, backchanelling, or none of these.

Features: Structural, Durational, Lexical, Pragmatic

Structural features:

· is the previous/next spurt of the same speaker?
· is the previous/next spurt involving the same B speaker?

Durational features:

· duration of the spurt
· seconds of overlap with any other speaker
· seconds of silence during the spurt
· speech rate in the spurt

Lexical features:

· number of words in the spurt
· number of content words in the spurt
· perplexity of the spurt with respect to four language models, one for each class
· first and last word of the spurt
· number of instances of adjectives with positive polarity (Hatzivassiloglou and McKeown, 1997)
· idem, with adjectives of negative polarity
· number of instances in the spurt of each cue phrase and agreement/disagreement token listed in (Hirschberg and Litman, 1994; Cohen, 2002)

Identifying Agreement and Disagreement

Feature sets	Accuracy
(Hillard et al., 2003)	82%
Lexical	84.95%
Structural and durational	71.23%
All (no label dependencies)	85.62%
All (with label dependencies)	86.92%

Table 6. 3-way classification accuracy

Feature sets	Label dep.	No label dep.
Lexical	83.54%	82.62%
Structural, durational	62.10%	58.86%
All	84.07%	83.11%

Table 7. 4-way classification accuracy

Identifying Agreement and Disagreement

Response and Critique

- Very interesting computational pragmatics study
- Does pragmatic information really improve classification accuracy? 1% is an improvement I guess…

Issues/Critical Response

- assumes spurts are valid segmentation
- agreement and disagreement are not categorical variables (agreement spectrum) -- and involvement/lack of involvement certainly aren’t either
- all on same dataset, and presumably some of the features are domain-specific (or speaker-specific)
- does not incorporate visual data such as expression, posture, gesture, and et cetera
- no analysis of effect on downstream applications

Thanks for listening!

Any questions?