RECOGNITION OF STANCE STRENGTH AND POLARITY IN SPONTANEOUS SPEECH

Gina-Anne Levow, Valerie Freeman, Alena Hrynkevich, Mari Ostendorf,
Richard Wright, Julian Chan, Yi Luan, Trang Tran
University of Washington
Seattle, WA USA

ABSTRACT

From activities as simple as scheduling a meeting to those as complex as balancing a national budget, people take stances in negotiations and decision making. While the related areas of subjectivity and sentiment analysis have received significant attention, work has focused almost exclusively on text, and much stance-taking activity is carried out verbally. This paper investigates automatic recognition of stance-taking in spontaneous speech. It first presents a new annotated corpus of spontaneous, conversational speech designed to elicit high densities of stance-taking at different strengths. Speaker spurts are annotated both for strength of stance-taking behavior and polarity of stance. Based on this annotated corpus, we develop classifiers for automatic recognition of stance-taking behavior in speech. We employ a range of lexical, speaking style, and prosodic features in a boosting framework. The classifiers achieve strong accuracies on both binary detection of stance and four-way recognition of stance strength, well above most common class assignment. Finally, we classify the polarity of stance-taking spurts, obtaining accuracies around 80%. The best classifiers rely primarily on word unigram features, with speaking style and prosodic features yielding lower accuracies but still well above common class assignment.

Index Terms— Stance recognition, spontaneous speech, polarity recognition, corpus annotation

1. INTRODUCTION

Stances, or a speaker's subjective attitudes or opinions about the topic of discussion [1, 2], are an integral part of activities involving collaboration, negotiation, and decision making. In automatic recognition research, stance is similar to sentiment and subjectivity, expressions of an internal mental or emotional "private state" [3]. Recognition research in these areas has grown rapidly following foundational work like [4, 5]. Generally, such work has relied on textual materials and annotated corpora, such as those described in [4, 5, 6]. Predominantly drawing on lexical and syntactic evidence, text-based approaches capitalize on well-formed sentences and complete thoughts; however, our focus is on stance-taking in spoken interactions, which involve ambiguous, fragmentary, or disfluent utterances. A much smaller amount of work has investigated issues of subjectivity, sentiment, or stance in speech, primarily by exploiting existing conversational dyadic ([7] in [8]) or multi-party meeting corpora ([9, 10, 11] in [12, 13, 14], respectively), small portions of which are annotated for elements of subjectivity such as agreement or arguing. Even with speech data, many approaches to automatic subjectivity recognition have leveraged mainly word or n-gram content [15], and efforts to incorporate prosodic information have yielded no significant improvement [16]. This is troubling, as stance-taking in speech harnesses channels of information not available in the textual content, including intonation, speaking rate, emphasis, and precision of articulation [17, 18]. However, [13] found that annotators were better able to identify opinions, especially negative opinions, when they had access to audio recordings than when using transcripts alone.

In the current study, we exploit the ATAROS corpus [19], a novel corpus designed explicitly to elicit high densities of stance-taking at different strengths. This corpus further addresses limitations of prior corpora by providing finer control of audio recording conditions, speaker demographics, and speaker dialect, as well as frequency and strength of stance-taking, while maintaining a relatively naturalistic conversational interaction constrained by task. We describe the manual annotation of the corpus for strength and polarity of stance at the utterance level, in contrast to prior work classifying speeches [20] or full posts [6, 21] in Congressional or online debates, respectively. Based on this annotation, we perform automatic recognition of stance strength and polarity. We contrast the effectiveness of lexical, speaking style, and prosodic features for these tasks. Within a boosting framework, the best results are obtained with word unigram features across all of these tasks.

The remainder of the paper is organized as follows. Section 2 describes the corpus collection, transcription, and stance annotation that form the foundation for stance recognition experiments. Section 3 presents the lexical, speaking style, and acoustic-prosodic features used to identify stance strength and polarity. Section 4 outlines the experimental configuration and reports and discusses the results of the

This work has been supported by NSF IIS: #1351034. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NSF or the U.S. Government.

classification experiments. Finally, Section 5 offers conclusions and outlines of future work.

2. CORPUS DATA COLLECTION

The experiments on stance recognition reported below rely on the ATAROS corpus, a corpus of task-oriented spontaneous, conversational speech designed to elicit a high density of stance-taking behavior at different strengths. Additional details on the corpus and tasks can be found in [19]. Participants working in pairs complete a total of five tasks, each averaging roughly ten minutes in duration. Here we focus analysis and experiments on two tasks, the Inventory task and the Budget task, designed to elicit low and high levels of stance-taking and engagement, respectively. In the Inventory task, participants are asked to imagine that they are managers of a superstore and to place products in appropriate locations in the store. In the Budget task, participants are asked to pretend that they are in charge of balancing an imaginary local budget and are required to cut items across different categories.

The sample used for stance recognition experiments includes 23 dyads, comprising 26 female speakers and 20 male speakers. All are native English-speakers age 18-75 who grew up in the Pacific Northwest (Washington, Oregon, Idaho). There are nine female-female dyads, five male-male dyads, and nine mixed gender dyads. The speakers are distributed across three age groups: roughly half are under 30, a quarter 30-60, and a quarter 60 and above. The 42 tasks include 22 Inventory and 20 Budget tasks.

All interactions are recorded in a sound-treated booth using close-talking head-mounted microphones on separate channels at a 44.1 kHz sampling rate. All speech is manually transcribed in Praat [22] at the level of the “spurt”, a span of speech by one speaker bounded by at least 500 ms of silence. Spurt boundaries are manually aligned to the audio, and speech is transcribed according to a simplified version of the ICSI Meeting Recorder transcription guidelines [9], which uses conventional spelling, capitalization, and punctuation. Filled pauses are transcribed as “um” or “uh” (nasal/non-nasal). Disfluencies are marked with a short dash, directly after a truncated word (e.g., categ-) or after a space following uncompleted thoughts (e.g., I thought -), which may end an utterance or precede a repetition or restart (e.g., I don’t - I’m not - I’m not sure.). A few common vocalizations are transcribed with tags (e.g., {VOC laugh}, {VOC cough}), and notable voice qualities are marked with a following descriptive tag (e.g., {QUAL laughing}). Words which are noticeably emphasized are marked with an asterisk. From this transcription, we employ the Penn Phonetics Lab Forced Aligner (P2FA; [23]) to identify word- and phone-level boundaries time-aligned to the audio signal.

2.1. Coarse-grained stance annotation

After manual orthographic transcription, the tasks are manually annotated at a coarse (inter-pausal) level. Each spurt is marked with one of the stance presence/strength labels listed below. Spurts with a discernible stance strength (label 1, 2, or 3) are also labeled for polarity, as described below. As a result, each spurt is marked with one of 14 possible strength-polarity label combinations.

Stance presence/strength

0 No stance (list reading, backchannels, fact-exchange). Ex: ”Next we have cookies.” ”Mm-hm.”

1 Weak stance (cursory agreement, suggesting solutions, soliciting other’s opinion, bland opinion/reasoning). Ex: ”What do you think?” ”Let’s put it here.” ”Okay, sure.”

2 Moderate stance (more emphatic/energetic/firm versions of items in #1, disagreement, offering alternatives, questioning other’s opinion). Ex: ”Uh, how about here instead?” ”Are you sure?” ”Yes! Perfect.”

3 Strong stance (very emphatic/strong/excited versions of items in #1-2). Ex: ”Oh my god! I can’t have that happen on my watch!” ”Screw that!”

x Unclear (cannot be determined, excited pronunciations of no-stance content). Ex: ”Ooh, buckets!” ”I don’t know what that means.”

Polarity (applied to strength labels 1, 2, 3)

+ Positive (agreement, approval/affinity, willing acceptance, encouragement, positive evaluation, etc.). Ex: ”Sure. Good idea.” ”Yes! Perfect.”

- Negative (disagreement, disapproval/dislike, rejection/grudging acceptance, hedging, negative evaluation, etc.). Ex: ”No, I don’t think so.” ”Well, I guess. If you really want to.”

(NA) Neutral (none of the above, non-evaluative offering or solicitation of opinions/solutions). Ex: ”What should we cut next?” ”Let’s do this one.”

x Unclear (cannot be determined).

Both textual content and prosody are taken into account when determining labels, as prosody can be used to enhance or even reverse the meaning of text alone. Because strength is relative, the scheme is applied on a per-speaker basis. Before labeling, annotators listen to a portion of the task or a prior task to get a general sense of each speaker’s styles and strategies. For example, for speakers with small pitch and intensity ranges, small deviations are more meaningful than for the most energetic speakers, whose modulations must be more extreme to indicate differences in stance. Annotators listen to the audio in Praat while labeling one speaker's transcription and then listen again while labeling the other's. After a task is labeled by one annotator, a second reviews and verifies or corrects each label while listening to the audio. Asterisks are used to indicate uncertainty, with the second annotator providing a second opinion as needed. If the second annotator remains uncertain about a label, a third annotator serves as a tie-breaker. This method yields very high inter-rater agreement. Weighted Cohen's kappas with equidistant penalties are 0.87 for stance strength labels and 0.93 for for polarity labels (p = 0), with the unweighted kappa for combined labels at 0.88 (p = 0).

3. CLASSIFICATION FEATURES

The experiments reported below exploit three main classes of features: text-based, speaking style, and acoustic-prosodic features. The text-based features are drawn directly from the manual transcriptions of the spurts in the ATAROS corpus. In particular, we employ word unigram features of the tokenized text of the transcript, with case-folding. We conduct comparative experiments with higher order word n-grams, character n-grams (similar to [15, 16]), and stemmed text. Higher order word n-grams produce no improvement over unigrams, character n-grams perform no better than word unigrams, and stemming yields a small reduction in classification accuracy.

For speaking style features, we compute several measures associated with higher levels of stance-taking and engagement in the different experimental tasks of the ATAROS data. As reported in [19], tasks intended to elicit higher levels of stance-taking and engagement exhibit increased spurt durations and greater rates of disfluent speech, as indicated by increased rates of repetition, filled pauses, and truncated words. Spurt duration is captured in terms of number of syllables, number of words, and total duration in seconds. Finally, we include the number of emphasized words, number of unintelligible spans, and number of filled pauses and truncated words, as marked in the transcript.

For prosodic features, we extract pitch and intensity measures calculated over each full spurt and over the last 500 milliseconds of the spurt. Pitch is computed using KALDI-pitch [24, 25], considered to be a current state-of-the-art pitch tracker. Intensity is calculated using Praat's [22] "To Intensity..." function. All values are log-scaled and z-score normalized on a per-speaker, per-task basis. For each span, we compute maximum, minimum, mean, and off-slope for pitch and intensity. Finally, for each spurt, we compute speaking rate in syllables per second.

Strength		Polarity	
Label	Proportion	Label	Proportion
0	27.5%	Neutral	56.2%
1	48.7%	Positive	37.1%
2	23.2%	Negative	6.7%
3	0.6%		

Table 1. Distribution of stance strength and polarity labels in classification experiments

4. EXPERIMENTS

The following section describes our experimental configuration and contrastive settings.

4.1. Experimental Configuration

All experiments employ ICSIboost [26], a freely available public domain reimplementation of BoosTexter [27]. ICSIboost provides a boosting classifier based on decision stump weak learners and supports categorical and real-valued features as well as n-gram features over text spans. Given the relatively small amount of data available, we chose to apply a cross-validation framework to our experiments with five folds. No speakers appear in both training and testing for a single fold. Based on initial exploratory experiments, we chose to run 500 rounds of training for each fold when word unigram features are included and 25 when they are not.

We perform experiments on automatic recognition of both stance strength and stance polarity. The unit for annotation of stance strength and polarity and thus experimentation is the "spurt", a span of speech delimited by at least 500 ms of silence, as determined by manual transcription and alignment to the audio signal. For stance recognition, we compare binary recognition, distinguishing spurts that involve stance-taking (stance strength levels 1-3) from those that do not (strength level 0), with full recognition of stance strength, labeling each spurt with a value between 0 and 3. For stance polarity, we perform three-way classification, distinguishing neutral, negative, and positive. Spurts with stance label "x" (unclear), are excluded from classification experiments. The distribution of labels in the stance strength and polarity classification experiments appears in Table 1. For each of these classification tasks, we compare the effectiveness of different feature sets, specifically: word unigrams, speaking style features, and acoustic prosodic features, as described in Section 3.

4.2. Stance Detection

The stance detection experiments consider the binary task of detecting whether a spurt exhibits any stance-taking activity (strength levels 1, 2, or 3) or not (strength level 0). The most common class is the "stance" class, accounting for 72.6%

of the samples. We find that word unigrams on their own achieve an accuracy of 80.5%. This represents a reduction in error of almost 30% over most common class assignment. A combination of prosodic and speaking style features only achieves an accuracy of 75%, still surpassing most common class assignment. However, no combination of features improves over the word unigram features.

4.3. Stance Strength Classification

Stance strength classification experiments consider the four-way task of recognizing the assigned stance strength level: 0, 1, 2, 3. The most frequent class is class 1 (weak stance), which is assigned to mild agreement; this class accounts for 48.7% of instances in the data.

For word unigrams alone, we obtain an accuracy of 71%, a 22% absolute improvement over that of most common class assignment. Prosodic features alone reach 54.1% and speaking style features 53.7%, well below that of the text features, but still solidly above most common class assignment. In combination, speaking style and prosodic features reach 55.2% accuracy. However, no combination with other features improves over word unigram feature effectiveness; simple combination of all features achieves only 64% accuracy.

4.4. Stance Polarity Classification

For stance polarity classification, we consider the three-way classification task of assigning known stance-bearing instances to the classes of “neutral”, “positive”, or “negative.” Here the most common class is “neutral”, accounting for roughly 56% of the samples in the data.

For word unigrams alone, we achieve an accuracy of 80%, yielding over 50% relative error reduction over most common class assignment. For prosodic features alone, we achieve an accuracy of 71%, a relative error reduction of 34% over most common class assignment, but a drop of 9% absolute from the text-based features. For speaking style features, the accuracy is similar to the prosodic features at about 71%. Here also, adding prosodic and speaking style features to the word unigram features yields no further improvement, but only a small decrease in accuracy to 79.3%.

4.5. Discussion

The experiments on stance recognition and polarity recognition above highlight the strong evidence provided by lexical information, in the form of word unigrams. These features yield the best effectiveness across all three tasks by a wide margin, with accuracies ranging from 71% to 81%. To gain insight into types of words indicating stance, we select unigrams that appear among the first 25 word unigrams selected by the classifier across all folds in the stance strength classification task; they appear in Table 2. The unigrams include indicators of agreement, fillers and floor holders, punctuation marks, evaluative content words (“important”,”good”), and closed class words, including modals and disjunctions. The corresponding words for polarity classification are shown in Table 3. This list is dominated by affirmative words but includes two terms denoting negative polarity. Punctuation items are correspondingly reduced. Such explicit terms will be highly informative for polarity classification.

yeah	okay	um	hm
need	maybe	important	good
the	this	could	but
?	,	!	*

Table 2. Unigrams selected early in all folds for stance strength recognition

yeah	yes	yep	true
sure	okay	mm-hm	kay
no	but		
the	here	.	

Table 3. Unigrams selected early in all folds for stance polarity recognition

Outside the unigram features, the effectiveness of the prosodic and speaking style features is much more limited. Given the heavy reliance of the stance strength classifier on punctuation features noted above, we speculate that this punctuation is conveying information that would otherwise be provided through prosodic means. For example, spurts with interrogative and exclamatory punctuation are likely to involve stance-taking and also to be prosodically distinguished. The prosodic measures would provide a (possibly noisy) information source that would be more reliably signalled by text features when manually assigned gold-standard punctuation is available. We tested this hypothesis by excluding punctuation features for stance strength classification and comparing effectiveness with and without prosodic and speaking style features. We find that four-way stance strength prediction accuracy using only word unigram features *excluding* punctuation drops to 61.5%. When we now add prosodic and speaking style features (again excluding those reliant on punctuation), accuracy improves to 63%. This increase suggests that the manually annotated punctuation features were masking the effect of prosodic cues.

It is also interesting to note the greater effectiveness of prosodic and speaking style features in stance polarity classification (15% absolute and 34% relative error reduction improvement over most common class) than in stance detection and stance strength recognition (3-6% absolute and roughly 10% relative error reduction), in both absolute and relative terms.

We further speculate that the current spurt segmentation, which can both oversegment a dialog act based on pauses and undersegment by merging multiple dialog acts, limits the effectiveness of the prosodic features. Key cues may be segmented into discrete spurts, and combinations of acts, such as backchannels or floor holders, may muddy stance-taking cues when prosodic measures such as pitch and intensity are aggregated across the spurt.

5. CONCLUSION & FUTURE WORK

This work has investigated stance strength and stance polarity in spontaneous speech. A novel stance-annotated corpus has been presented, describing the annotation methodology and the strong level of interannotator agreement achieved on this task. We have further explored automatic recognition of stance strength and stance polarity in a boosting framework, comparing textual, speaking style, and prosodic features. Good effectiveness has been achieved on these tasks, with accuracies from 71-81% and relative reduction in error from common class assignment of 30-50%, depending on the task. Word unigram features yielded the best results across all tasks, with prosodic and speaking style features exhibiting much more limited utility.

We plan to explore two main strategies to improve both stance strength and stance polarity recognition. First, we plan to exploit a more fine-grained dialog act segmentation as our unit of analysis. The current spurt-based units, while simple to extract, can often subsume multiple dialog acts with potentially different stance-taking behaviors, for example, when a back-channel (0 stance) is followed by a stance-taking move. In some cases during annotation, annotators indicated a desire to further segment the spurt or to assign multiple, possibly opposing labels. Second, we plan to investigate alternative, novel prosodic measures to better capture the dynamics of speech associated with stance-taking, including measures of changes in the vowel space and measures related to the modulation spectrum.

6. REFERENCES

[1] Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan, *Longman grammar of spoken and written English*, Longman, 1999.

[2] Pentti Haddington, "Stance taking in news interviews," *SKY Journal of Linguistics*, vol. 17, pp. 101–142, 2004.

[3] Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik, *A Comprehensive Grammar of the English Language*, Longman, New York, 1985.

[4] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2002, pp. 79–86.

[5] Janyce Wiebe, Theresa Wilson, and Claire Cardie, "Annotating expressions of opinions and emotions in language," *Language Resources and Evaluation*, vol. 39, no. 2–3, pp. 165–210, 2005.

[6] Swapna Somasundaran and Janyce Wiebe, "Recognizing stances in online debates," in *Proceedings of ACL 2009: Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, 2009.

[7] John Godfrey, Edward Holliman, and Jane McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *Proceedings of ICASSP-92*, 1992, pp. 517–520.

[8] Gabriel Murray and Giuseppe Carenini, "Detecting subjectivity in multiparty speech," in *Proceedings of Interspeech 2009*, 2009, pp. 2007–2010.

[9] Nelson Morgan, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Adam Janin, Thilo Pfau, Elizabeth Shriberg, and Andreas Stolcke, "The meeting project at ICSI," in *Proceedings of Human Language Technologies Conference*, 2001.

[10] Susanne Burger, Victoria MacLaren, and Hua Yu, "The ISL meeting corpus: The impact of meeting type on speech type," in *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, 2002.

[11] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI meeting corpus," in *Proceedings of the Measuring Behavior Symposium on "Annotating and Measuring Meeting Behavior"*, 2005.

[12] Dustin Hillard, Mari Ostendorf, and Elizabeth Shriberg, "Detection of agreement vs. disagreement in meetings: Training with unlabeled data," in *Proceedings of HLT-NAACL Conference*, Edmonton, Canada, 2003.

[13] Swapna Somasundaran, Janyce Wiebe, Paul Hoffmann, and Diane Litman, "Manual annotation of opinion categories in meetings," in *ACL Workshop: Frontiers in Linguistically Annotated Corpora(Coling/ACL 2006)*, 2006.

[14] Theresa Wilson, "Annotating subjective content in meetings," in *Proceedings of the Language Resources and Evaluation Conference*, 2008.

[15] Theresa Wilson and Stephan Raaijmakers, "Comparing word, character, and phoneme n-grams for subjective utterance recognition," in *Proceedings of Interspeech 2008*, 2008.

[16] Stephan Raaijmakers, Khiet Truong, and Theresa Wilson, "Multimodal subjectivity analysis of multiparty conversation," in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, Honolulu, Hawaii, October 2008, pp. 466–474, Association for Computational Linguistics.

[17] Valerie Freeman, "Using acoustic measures of hyperarticulation to quantify novelty and evaluation in a corpus of political talk shows," M.S. thesis, University of Washington, 2010.

[18] Valerie Freeman, "Hyperarticulation as a signal of stance," *Journal of Phonetics*, vol. 45, pp. 1–11, 2014.

[19] Valerie Freeman, Julian Chan, Gina-Anne Levow, Richard Wright, Mari Ostendorf, and Victoria Zayats, "Manipulating stance and involvement using collaborative tasks: An exploratory comparison," in *Proceedeings of Interspeech 2014*, 2014.

[20] M. Thomas, B. Pang, and L. Lee, "Get out the vote: Determining support or opposition from congressional floor-debate transcripts," in *Proceedings of EMNLP 2006*, 2006.

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

[22] Paul Boersma and David Weenink, "Praat: doing phonetics by computer [computer program], version 5.3.55," 2013, http://www.praat.org.

[23] Jiahong Yuan and Mark Liberman, "Speaker identification on the SCOTUS corpus," in *Proceedings of Acoustics '08*, 2008.

[24] D. Povey, A. Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsk, Georg Stemmer, and Karel Vesel, "The Kaldi speech recognition toolkit," in *Proceedings of ASRU 2011*, 2011.

[25] Pegah Ghahremani, Bagher BabaAli, Daniel Povey, Korbinian Riedhammer, Jan Trmal, and Sanjeev Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *Proceedings of ICASSP 2014*, 2014.

[26] Benoit Favre, Dilek Hakkani-Tür, and Sebastien Cuendet, "Icsiboost," `http://code.google.come/p/icsiboost`, 2007.

[27] Robert E. Schapire and Yoram Singer, "Boostexter: A boosting-based system for text categorization," *Machine Learning*, vol. 39, no. 2–3, pp. 135–168, 2000.