Detecting Deception Using Critical Segments

Frank Enos,[*] Elizabeth Shriberg,[†§] Martín Graciarena,[†] Julia Hirschberg,[*] Andreas Stolcke[†§]
[*]Columbia University, USA; [†]SRI, USA; [§]ICSI, USA

Detecting deception is a notoriously difficult task. Most human subjects — including trained professionals — perform near chance at deception detection[2]. A study using our corpus, the Columbia-Colorado-SRI (CSC) Deception Corpus, (see center panel), showed that listeners performed *worse* than chance at detecting deception in recorded speech[4]. The present work examined certain systematically identifiable segments — called CRITICAL SEGMENTS — that bear propositional content directly related to the topics of most interest in the interrogation. We augmented this approach with techniques for adjusting the class imbalance in the data. Our results, as much as 23.8% relative improvement over chance, substantially exceed human performance at the task of **TRUTH** and **LIE** classification[4]. Further, models generated using these segments employ features consistent with hypotheses in the literature and the expectations of practitioners [6] about cues to deception.

Hot Spots

Work by psychologists on behavioral and facial cues to deception [7, 8] suggests that certain events in interviews, termed HOT SPOTS, are particularly useful in determining a subject's veracity. We thus hoped to find that certain segments of speech that deal directly with the most salient topics of the speaker's deception are more easily classified than other deceptive statements. Presumably, such segments would be both emotionally charged and cognitively loaded, potentially resulting in stronger acoustic, prosodic, and lexical cues. In the present work, we developed systematic rules to isolate potential HOT SPOTS, which we term CRITICAL SEGMENTS (see center panel). We used these segments as proxies to determine truthfulness in each section of the interview.

Skewed Class Distributions

Our sets of CRITICAL SEGMENTS contain a majority of **LIE** examples: (67.5% for Critical, 62% for Critical-Plus). Results on the natural distribution were poor but exceeded chance, so we hoped that adjusting the skew might allow the learner to induce more effective rules, since classification algorithms can be hampered by the 'over-prevalence' [16] of majority class examples[14, 16, 17, 18]. We thus used under-sampling [17] to eliminate the bias of models for the majority class.

Table 1 Accuracy Classifying Global Lies and Truths

Dataset	*Relative Improvement*	*Accuracy*	*Baseline*
Critical-Plus	5.8%	65.6	62.0
Critical	1.6%	68.6	67.5
Critical-Plus Under-sampled	22.2%	61.1	50.0
Critical Under-sampled	23.8%	61.9	50.0

About the CSC Corpus

The Columbia-SRI-Colorado (CSC) Deception Corpus [3] is a laboratory collection of 32 recorded interviews of native Standard American English speakers, containing within-subject deceptive and nondeceptive speech. Speakers were motivated via financial incentive and by what social psychologists term the 'self-presentational' perspective [1] to lie about their performance on a pre-test in six areas of general knowledge. Features extracted from the corpus include over 200 lexical, prosodic, and acoustic features.

In this paper, **TRUTH** or **LIE** describes the speaker's overall intention to deceive (or not) with respect to the salient topic of the conversation; that is, the claimed score for each section of the pre-test.

Duration of the interviews ranged from 25 to 50 minutes and comprised 15.2 hours of dialog, providing approximately 7 hours of subject speech.

Hypotheses on Critical Segments

1. CRITICAL SEGMENTS will occur when the propositional content of the segment relates directly to the most salient topics of the interview.
2. CRITICAL SEGMENTS will occur when subjects are directly challenged to explain their claims with regard to salient topics of the interview.

Selection of Critical Segments

Segments were selected by hand as follows:
1. Include segments that are responses to questions that directly ask the subject for his or her score on a particular section.
2. Include segments that respond to immediate follow-up questions requesting a justification of the claimed score, when such a question is posed by the interviewer.
3. Omit everything else.

Examples of Critical Segments

Subject segment (S) corresponding to Rule 1:

(I) *And what was your score exactly on that section?*

(S) *I got excellent, which was, um, pretty good.*

Follow-up questions were included (Rule 2).:

(I) *Why do you think you did so well on that section?*

(S) *Um my-first of all my grandmother was a really good cook.*

Multiple adjacent segments corresponding to Rules 1 or 2 were included:

(I) *So we'll move on now to what we're calling the civics section. How did you do on that section?*

(S) *Uh I d-you know alright.*

(S) *Not great.*

(S) *Fair.*

Segments not corresponding to Rules 1 or 2 were omitted:

(S) *I went to this in-Indian restaurant my parents call Tamarind's.*

Sets Produced

CRITICAL (Rule 1): 465 segments
CRITICAL-PLUS (Rules 1 & 2) 675 segments

Feature Selection

Feature selection was employed to reduce the feature set to 22 features for the Critical set and 56 for the Critical-Plus set.

Method

We classified critical segments as TRUTH or LIE combining implementations of bagging [12], AdaBoost [13], and c4.5 [14] (called J48) provided by Weka and the Weka Java API [15]. For the **Critical-Plus** dataset, 100 class-balanced training/test sets were created by randomly selecting 50 examples (25 **TRUTH**, 25 **LIE**) for each test set; from the remaining examples, we randomly selected 458 (229 **TRUTH**, 229 **LIE**) examples for each training set. An analogous approach produced 100 sets of 272 training and 30 test examples for the **Critical** dataset.

Results

In Table 1 we report averaged classification results on both the original distribution (10-fold cross-validation) and on the under-sampled datasets (100 random trials). Performance on the original samples is poor but exceeds chance. Results for the under-sampled datasets show substantial improvement, lending support to the hypothesis that resampling can render the learner more capable of producing useful rules[16, 18]. Context for comparison is provided by the performance of humans at the analogous task of labeling subject lies in each section of the interview: 32 humans scored on average 47.8% versus a chance baseline of 63.6%[4]. Systems have performed from 7% to 10% (relative) above chance on the somewhat different task of classifying the veracity of every segment [3,4,5].

Many of the rules induced from the current dataset paint a plausible picture of the correlates of deception, one consistent with previous literature. Lexical cues that reflect positive emotional state [19, 9] appear as cues to truth. Such assertive terms as *yes* or *no* also serve as cues to truth. The presence of qualifiers (such as *absolutely* or *really*) is employed as a cue to deception. Filled pauses appear as a cue to truth in many rules produced. Self-repairs appear in numerous rules as a cue to truth, consistent with the finding of De Paulo et al. [1] . Finally, deviation from a subject's baseline values for various energy features cues deception.

The performance we report here is modest, but it is important, since our CRITICAL SEGMENTS are those that point directly to the topics of most interest in the interview. Future work will include the automatic labeling of such segments, and further exploration of other genres of CRITICAL SEGMENTS, such as cases where the interviewer directly accuses the subject of lying.

Acknowledgments

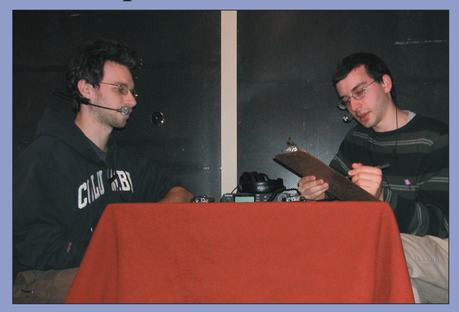
This research was funded in part by NSF IIS-0325399 and the Department of Homeland Security. The authors thank Stefan Benus for many helpful conversations.

CS @CU COMPUTER SCIENCE AT COLUMBIA UNIVERSITY

ICSI INTERNATIONAL COMPUTER SCIENCE INSTITUTE

SRI International®

Setup used to record the CSC Corpus (simulation).



References

[1] B. M. DePaulo, J. J. Lindsay, B. E. Malone, L. Muhlenbruck, K. Charlton, and H. Cooper, "Cues to deception." *Psychological Bulletin*, vol. 129, no. 1, pp. 74–118, 2003.

[2] M. Aamodt and H. Custer, "Who can best catch a liar?" *Forensic Examiner*, vol. 15, no. 1, pp. 6–11, 2006.

[3] J. Hirschberg, S. Benus, J. M. Brenier, S. F. F. Enos, S. Gilman, C. Girand, M. Graciarena, L. M. A. Kathol, B. Pellom, E. Shriberg, and A. Stolcke, "Distinguishing deceptive from non-deceptive speech," in *Proc. Eurospeech*. Lisbon: ISCA, 2005.

[4] F. Enos, S. Benus, R. Cautin, M. Graciarena, J. Hirschberg, and E. Shriberg, "Personality factors in human deception detection: Comparing human to machine performance," in *Proc. Inter- speech*. Pittsburgh: ISCA, 2006.

[5] M. Graciarena, E. Shriberg, A. Stolcke, J. H. F. Enos, and S. Kajarekar, "Combining prosodic, lexical and cepstral systems for deceptive speech detection," in *Proc. IEEE ICASSP*. Toulouse, France: IEEE, 2006.

[6] J. Reid and Associates, *The Reid Technique of Interviewing and Interrogation*. Chicago: John E. Reid and Associates, Inc., 2000.

[7] R. Adelson, "Detecting deception," *APA Monitor on Psychology*, vol. 35, no. 7, 2004.

[8] M. Frank, Personal Communication, 2005.

[9] M. L. Newman, J. W. Pennebaker, D. S. Berry, and J. M. Richards, "Lying words: Predicting deception from linguistic style." *Personality and Social Psych. Bull.*, vol. 29, pp. 665–675, 2003.

[10] NIST, "Fall 2004 rich transcription (rt-04f) evaluation plan," http://www.nist.gov/speech/tests/rt/rt2004/fall/ docs/rt04f-evalplan-v14.pdf, 2004.

[11] E. Shriberg and A. Stolcke, "Direct modeling of prosody: An overview of applications in automatic speech processing," in *Proc. International Conference on Speech Prosody*, Nara, Japan, 2004.

[12] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996. [Online]. Available: cite- seer.ist.psu.edu/breiman96bagging.html

[13] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," in *European Conference on Computational Learning Theory*, 1995, pp. 23–37. [Online]. Available: cite- seer.ist.psu.edu/article/freund95decision theoretic.html

[14] J. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, pp. 81–106, 1986.

[15] S. Garner, "Weka: The waikato environment for knowledge analysis," in *Proc. New Zealand Computer Science Research Students Conference*, pages 57–64, 1995. [Online]. Available: citeseer.ist.psu.edu/garner95weka.html

[16] N. Chawla, "C4.5 and imbalanced data sets: Investigating the effect of sampling method, probabilistic estimate, and decision tree structure," in *Proc. Workshop on Learning from Imbalanced Data Sets II*. Washington, DC: ICML, August 2003.

[17] C. Drummond and R. Holte, "C4.5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling," in *Proc. Workshop on Learning from Imbalanced Data Sets II*. Washington, DC: ICML, August 2003.

[18] V. Hoste, "Optimization issues in machine learning of coreference resolution," Ph.D. dissertation, University of Antwerp, http://www.cnts.ua.ac.be/ hoste/proefschrift.html, 2005.

[19] C. Whissel, "The dictionary of affect in language," in *Emotion: Theory, Research and Experience*, R. Plutchik and H. Kellerman, Eds. Academic Press, 1989, pp. 113–131.

[20] S. Benus, F. Enos, J. Hirschberg, and E. Shriberg, "Pauses in deceptive speech," in *Proc. ISCA 3rd International Conference on Speech Prosody*. Dresden, Germany: ISCA, 2006.

[21] M. O'Sullivan and P. Ekman, "The wizards of deception detection," in *The Detection of Deception in Forensic Contexts*, P. Granhag and L. Strömwall, Eds. Cambridge: Cambridge University Press, 2004.

[22] L. A. Streeter, R. M. Krauss, V. Geller, C. Olson, and W. Apple, "Pitch changes during attempted deception." *Journal of Personality and Social Psychology*, vol. 35, no. 5, pp. 345–350, 1977.