Detecting Deception Using Critical Segments

Frank Enos, Elizabeth Shriberg†§*
Martin Graciarena,† Julia Hirschberg, Andreas Stolcke †§*
*Columbia University, USA; †SRI, USA; §ICSI, USA
frank@cs.columbia.edu

Abstract

We present an investigation of segments that map to GLOBAL LIES, that is, the intent to deceive with respect to salient topics of the discourse. We propose that identifying the truth or falsity of these CRITICAL SEGMENTS may be important in determining a speaker's veracity over the larger topic of discourse. Further, answers to key questions, which can be identified *a priori*, may represent emotional and cognitive HOT SPOTS, analogous to those observed by psychologists who study gestural and facial cues to deception. We present results of experiments that use two different definitions of CRITICAL SEGMENTS and employ machine learning techniques that compensate for imbalances in the dataset. Using this approach, we achieve a performance gain of 23.8% relative to chance, in contrast with human performance on a similar task, which averages substantially below chance. We discuss the features used by the models, and consider how these findings can influence future research.
Index Terms: deception, deceptive, speech

1. Introduction

The detection of deception has long been of interest in the domains of law enforcement, national security, business, and research. Interest continues to grow in the general area of 'credibility assessment,' and in the particular task of the detection of deceptive speech. Most work in this area has appeared in the psychology literature (c.f. [1]); work has been recently undertaken by computational linguists and speech scientists who aim to develop systems that classify deceptive and nondeceptive speech using machine learning and speech technologies. How well such systems can be hoped to perform is an open question: Since most human subjects — including trained professionals — perform near chance accuracy at the general deception detection task [2], an automatic system would have to perform considerably *better* than humans to have practical use. This lies in stark contrast to many speech processing tasks such as summarization or even emotion detection, where human performance is usually considered the gold standard.

Studies of automatic deception detection are relatively rare in the literature. Progress has been made recently, particularly in work on the CSC (Columbia-SRI-Colorado) Deception Corpus (see Section 3). Systems have performed from 4% to 6% better than chance (7–10% relative gain) using combined acoustic/prosodic, lexical, and subject-dependent features [3, 4, 5], classifying LOCAL LIES and TRUTHS. While this performance is modest, it substantially exceeds that of human judges on the same data, who performed on average *below* chance[4]. With respect to accuracy at labeling GLOBAL LIES and TRUTHS in the CSC Corpus, the topic of the present work, human judges performed even worse: on average 47.8% versus a chance baseline of 63.6%.

In the current work, we focus on detecting a speaker's more general intention to deceive, i.e., to perpetrate what we term GLOBAL LIES (Section 3). We do so by examining certain systematically identifiable segments — called here CRITICAL SEGMENTS — that may be more emotionally or cognitively charged than segments from the general corpus. The segments examined here bear propositional content that is directly related to the topics of most interest in the mock interrogation paradigm used in the corpus; classification of such segments is thus particularly important. Results reported here substantially exceed human performance at the task of GLOBAL LIE and TRUTH classification[4]. Further, models generated using these segments employ features consistent with hypotheses in the literature [1] and the expectations of practitioners [6] (see Section 5) about spoken cues to deception.

These findings are of interest on a number of fronts. First, they suggest that there may be a speech analog to what psychologists who study behavioral and facial cues to deception call HOT SPOTS, events in which relevant emotion is particularly observable and can thus be more easily detected[7, 8]. Second, such findings can guide the design of future data collection paradigms and real-world approaches, since interviewing techniques might be optimized to induce the subject to produce more CRITICAL SEGMENTS. Finally, continued work at automatic detection can be guided by the general principle that certain kinds of subject responses are more susceptible to detection, and that methods should be developed to identify and examine these sorts of responses.

1.1. Critical segments

Work by psychologists studying behavioral and facial cues to deception [7, 8] suggests that certain events in interviews, termed HOT SPOTS, are particularly useful in determining whether a subject is telling the truth. In directing detection efforts to CRITICAL SEGMENTS, we 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 deceptive statements in the corpus at large. Presumably, such segments will be both emotionally charged — potentially resulting in stronger prosodic and acoustic cues — and cognitively loaded — potentially resulting in more lexical cues to deception.

In the present work, we attempted to develop systematic rules to isolate potential HOT SPOTS, which in the speech domain we term CRITICAL SEGMENTS. These rules are based on two simple hypotheses about the nature of 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.

In what follows, we describe existing work on detecting deception (Section 2), describe the corpus (Section 3), explain our approach to operationalizing our hypotheses (Section 4), and report results obtained by experiments performed on the data thus extracted from the CSC Corpus (Section 5).

2. Related Work

Most results on deception detection have appeared in the psychology literature; work on the automatic detection of deception on the part of speech scientists and computational linguists has begun only recently. A number of studies report machine performance (7–10% improvement relative to chance on LOCAL LIES) [3, 4, 5] and human performance (below chance) [4] on the CSC Corpus. A study by Newman et al. [9] uses automatically extracted linguistic features, but it is difficult to infer performance gains with respect to unseen data given the details provided.

A human baseline for the general deception-detection task can be found in a recent meta-analysis [2] of the results of 108 studies of human deception detection. The majority of studies employed college students, who scored on average 54.22% compared to a baseline of 50%. Police and federal officers also performed near chance. A 2003 meta-analysis of 116 studies performed primarily by psychologists [1] reports 23 cues to deception that were significant across multiple studies.

3. The CSC Deception Corpus

The CSC Deception Corpus [3] is a laboratory collection of 32 recorded interviews containing within-subject deceptive and nondeceptive speech. Speakers were motivated via financial incentive to deceive successfully. In addition, speakers were led to believe that the ability to deceive correlates with other desirable personal and social qualities; this linked success at deception to what social psychologists term the 'self-presentational' perspective[1].

Subjects were native speakers of Standard American English, recruited from the Columbia University student population and from the larger community in exchange for payment. Subjects were solicited for a 'communication study' that sought individuals matching a profile based on the 'top twenty-five entrepreneurs' of America (this was false). Prior to the interview, subjects completed a test in six areas. The difficulty of the tasks and questions was manipulated so that each speaker scored too high in two areas, too low in two areas, and correctly in two. Four target profiles were constructed to balance the distribution of lies among the six areas. After completing the test, subjects were told that the study actually sought individuals who did not match the profile, but could lead an interviewer to believe that they did. Those who successfully deceived the interviewer would participate in a drawing for $100.

During the interview, speakers indicated whether each of their statements was true or contained some element of deception by pressing one of two pedals hidden from the interviewer. Ground truth was known *a priori* with respect to the claimed score (the most salient topic of conversation) since it was based on speakers' scores on the six-topic 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. Speech was segmented on several levels: segmentations using sentence-like units (EARS slash units or SUs) [10] are used in the present experiments. Full details regarding data collection can be found in [3]. The standard CSC corpus feature set [3] consists of 251 features: acoustic and automatically extracted prosodic features (as in [11]): automatically extracted lexical features, and features extracted automatically based on individual subject behaviors (such as ratio of laughs in lies vs. truths).

The CSC paradigm results in the production of two kinds of lies. GLOBAL LIES describe the speaker's overall intention to deceive (or not) with respect to a salient topic of the conversation; here, the claimed score for each section. LOCAL LIES refer to the propositional content of statements made to support the overall argument; this content will be either true or false. This distinction is important to the findings of the present study: while earlier work has focused on the detection of LOCAL LIES, the current work presents an approach to classifying GLOBAL LIES and TRUTHS.

4. Methods and Materials

We performed machine learning classification experiments — classifying **TRUTH** or **LIE** — on CRITICAL SEGMENTS identified in the CSC corpus. These were performed using implementations of bagging [12], AdaBoost [13], and c4.5 [14] (called J48) provided by Weka and the Weka Java API [15]. Feature selection was performed on features from the CSC Corpus feature set during the current experiments; features used are described in further detail in Section 5.

4.1. Selection of critical segments

CRITICAL SEGMENTS were selected by hand from the full set of segments (EARS slash units or SUs [10]) using the following rules:

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.

Here is a representative example of a subject segment (labeled **(S)**) that corresponds to Rule 1:

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

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

The interviewer frequently posed a follow-up question requesting immediate justification of the score claimed by the subject, as described in Rule 2. Responses to such questions were included:

(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.*

Often, a subject used multiple adjacent SUs in a response that corresponded to Rules 1 or 2. In such a case, all segments representing the response 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.*

Finally, many subject segments did not correspond to either Rules 1 or 2 because they were not produced in response to questions of the two genres described above. Such segments were omitted:

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

From the corpus of 9068 SUs, we thus produced two sets of CRITICAL SEGMENTS: one set of 465 based only on Rule 1 (termed **Critical**) and one set of 675 based on Rules 1 and 2 (termed **Critical-Plus**). Feature selection was employed to reduce the feature set to 22 features for the **Critical** set and 56 for the **Critical-Plus** set.

4.2. Coping with skewed class distributions

It is well known that classification algorithms — particularly those using decision trees, such as c4.5 [14] — can be negatively affected by datasets in which the class distribution is skewed (c.f [16, 17, 18]). In simple terms, this results in a bias on the part of the induced decision tree toward the majority class due to the 'over-prevalence' [16] of majority class examples.

For CRITICAL SEGMENTS, the CSC Corpus is such a dataset. The present sets of CRITICAL SEGMENTS contain a majority of **LIE** examples: (67.5% for **Critical**, 62% for **Critical-Plus**). Because initial classification results on the natural class distribution were poor but exceeded chance, we hypothesized that adjusting the class imbalance might allow the learner to induce more effective rules. We follow a commonly used approach to adjust the imbalance.

In this approach, termed under-sampling[1] [17], examples from the majority class are eliminated in order to create a balanced distribution. For the **Critical-Plus** dataset, combined training/test sets of 508 examples[2] were used. Under-sampled training/test sets were created as follows: for each of 10 training/test sets, randomly select 50 examples (25 **TRUTH**, 25 **LIE**) for the test set; from the remaining examples, randomly select 458 (229 **TRUTH**, 229 **LIE**) for the test set. An analogous approach was used with the **Critical** dataset, producing sets of 272 training and 30 test examples.

For each dataset, the above procedure was repeated 10 times with different random seeds to account for the exclusion of some data; results reported here thus reflect average performance on 100 individual training/test sets for each dataset.

5. Results and Discussion

In Table 1 we report classification results for the two datasets, both for the original samples (using 10-fold cross-validation) and for the under-sampled datasets, using 100 random trials as described in Section 4.2. Both raw accuracy and improvement relative to chance are reported. Given the difference in baselines, the relative scores represent the best basis for comparison since these scores are normalized with respect to the baseline chance accuracy, which varies among the configurations of the data. Performance on the original samples is poor but exceeds chance: 5.8% relative to chance for the **Critical-Plus** dataset, 1.6% for the **Critical** dataset. Results for the under-sampled datasets show 22.2% relative improvement for the **Critical-Plus** set and 23.8% relative improvement for the **Critical** set. This lends support to our hypothesis with respect to the skew of the distribution: in cases where the over-prevalence of one class interferes with c4.5's modeling, resampling can render the learner more capable of producing useful rules[16, 18].

There are no previous results for classification of GLOBAL LIES and TRUTHS on the corpus to provide a standard for comparison. Some context is provided, however, by the performance of humans at the analogous task of labeling GLOBAL LIES with respect to each section of the interview: 32 human listeners scored on average 47.8% versus a chance baseline of 63.6%[4].

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

An interesting aspect of these results is that performance is slightly better for the **Critical** dataset than for the **Critical-Plus** dataset, despite the smaller size of the **Critical** set (272 training examples in each trial, versus 414). We suspect that this difference is due to the increased cognitive and emotional stakes of the questions involved: The **Critical** dataset contains only subject segments that respond directly to the interviewer's most salient questions (e.g., 'What was your score on section X?'); the **Critical-Plus** dataset includes additional segments that contextualize that question but do not respond directly to it. It is possible that the latter differ enough with respect to emotional and cognitive load to produce a less effective learner when included with the smaller **Critical** set.

5.1. Importance of critical segments

The findings we report here are particularly relevant to the general deception detection task since our CRITICAL SEGMENTS are those that point directly to the topic of most interest with regard to the interview: the test scores claimed by the subjects. Earlier studies have attempted the separate task of classifying all segments in the corpus with respect to LOCAL LIES with relative accuracy gains of 7–10% above chance. However, the primary task embodied by the paradigm (and attempted by human listeners with little success in an earlier perception study [4]) is to determine the veracity of the subjects' claims with regard to their scores. Thus, while performance achieved here is modest, it is significant since this performance is obtained specifically on the segments whose veracity is of greatest interest, those that reflect the GLOBAL LIE category.

We have also shown that a more powerful classifier can be trained using resampling techniques that compensate for the corpus's skewed class distributions. The substantially improved performance indicates that the learner is better able to infer more useful rules when the present data are distributed evenly — and more importantly that such rules exist.

5.2. Relevant features

Because the bagging/boosting approach used here in 100 trials per dataset produced a large number of c4.5 decision trees, it is impractical to give an exhaustive description of the features employed in the models. We can, however, make some general observations about features that applied to a large number of cases in the induced trees.

Many of the rules induced from the current dataset paint a very plausible picture of the correlates of deception and one that is consistent with previous literature. First, lexical cues that

[1] Under-sampling is generally preferable to over-sampling; see [17] for details.

[2] The total number of examples available after subtracting the 167 'excess' **LIE** examples is 508.

speak to emotional state, such as the presence of negative or positive emotion words [19, 9], appear prominently. In particular, the presence of positive emotion words correlates positively with truth in many of the models produced. Likewise, many decision trees include rules based on features that could be interpreted to relate to the quality of being 'compelling' [1]. The use of such assertive terms as *yes* or *no*, for example, serves as a cue to deception in the models produced. Likewise, the presence of a specific, direct denial that the subject is lying is used in many rules as a cue to truth. This feature in particular has been cited by law enforcement practitioners as a cue to deceit [6], but we are unaware of previous evidence in the scientific deception literature that supports this claim. The presence of qualifiers (such as *absolutely* or *really*) is employed as a cue to deception in the models; this again is a feature gleaned from conversations with practitioners. Filled pauses appear as a cue to truth in many rules produced; this is consistent with an analysis of filled pauses in the CSC Corpus reported by Benus et al. [20]. Self-repairs appear in numerous rules as a cue to truth; this is consistent with the finding of De Paulo et al. [1] that liars exhibit fewer ordinary imperfections in their speech. Finally, various energy features (captured using a number of normalization schemes [11]) are employed in complicated rules that suggest that extreme values for energy — either high or low — correlate with deception. This is consistent with suggestions in the literature [21] that a subject's deviation from his or her baseline behavior is a useful cue to deception. It is interesting that although some studies have shown a correlation between increased F_0 and deception (c.f. [22]), F_0 features do not appear prominently in most of the rules induced here. One notable exception is that a number of F_0 slope features do appear in rules induced on the **Critical-Plus** dataset; we hesitate to make inferences about the nature of the correlation, however, since these features are generally embedded in complicated subtrees. A difference between our two datasets is that the presence of past tense verbs appears to correlate with deception in the **Critical-Plus** dataset, while it is not employed in the **Critical** set.

6. Conclusions and Future Work

The work reported here uses systematically identifiable CRITICAL SEGMENTS to detect deception on the GLOBAL LIE level in the CSC Corpus. Results substantially exceed human performance at a similar task. This finding can guide future research on a number of fronts. First, future paradigms can be designed to optimize subjects' production of CRITICAL SEGMENTS. For example, interviewers can be instructed to focus primarily on questions that require direct assertions about the most salient facts of the paradigm. Further, methods should be investigated that will allow for the automatic labeling of such segments, possibly using a combination of lexical features from both the interviewer and the subject. Finally, further investigation of the CSC Corpus is warranted, since other genres of CRITICAL SEGMENTS may exist, such as cases where the interviewer directly accuses the subject of lying.

7. 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.

8. 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. Interspeech*. 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-eval-plan-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: citeseer.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: citeseer.ist.psu.edu/article/freund95decisiontheoretic.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.