Linguistic Cues to Deception and Perceived Deception in Interview Dialogues

Sarah Ita Levitan, Angel Maredia, & Julia Hirschberg
Department of Computer Science
Columbia University
New York, NY, USA
{sarahita@cs, asm2221, julia@cs}.columbia.edu

Abstract

We explore deception detection in interview dialogues. We analyze a set of linguistic features in both truthful and deceptive responses to interview questions. We also study the perception of deception, identifying characteristics of statements that are perceived as truthful or deceptive by interviewers. Our analysis show significant differences between truthful and deceptive question responses, as well as variations in deception patterns across gender and native language. This analysis motivated our selection of features for machine learning experiments aimed at classifying globally deceptive speech. Our best classification performance is 72.74 F1-Score (about 27% better than human performance), which is achieved using a combination of linguistic features and individual traits.

1 Introduction

Deception detection is a critical problem studied by psychologists, criminologists, and computer scientists. In recent years the NLP and speech communities have increased their interest in deception detection. Language cues are inexpensive and easy to collect, and research examining text-based and speech-based cues to deception has been quite promising. Prior work has examined deceptive language in several domains, including fake reviews, mock crime scenes, and opinions about topics such as abortion or the death penalty. In this work we explore the domain of interview dialogues, which are similar to many real-world deception conditions.

Previous work has presented the results of classification experiments using linguistic features, attempting to identify which features contribute most to classification accuracy. However, studies often do not include an empirical analysis of features. We might know that a particular feature set (e.g. LIWC categories) is useful for deception classification, but we lack insight about the nature of the deceptive and truthful language that makes the feature set useful, and whether the differences in language use are statistically significant. In this work we conduct an empirical analysis of feature sets and report on the different characteristics of truthful and deceptive language. In addition, previous work has focused on the characteristics of deceptive language, and not on the characteristics of *perceived* deceptive language. We are also interested in human perception of deception; that is, what are the characteristics of language that listeners perceive as truthful or deceptive? We examine a unique dataset that includes information about both the deceiver and the interviewer, along with interviewer judgments of deception. Along with an analysis of deceptive and truthful speech, we analyze the believed and disbelieved speech, according to reported interviewer judgments. Finally, previous work has focused on general inferences about deception; here we include analysis of gender and native language, to study their effect on deceptive behavior, and also their effect on perception of deception. This work contributes to the critical problem of automatic deception detection, and increases our scientific understanding of deception, deception perception, and speaker differences in deceptive behavior.

The paper is organized as follows: In Section 2 we review related work in language-based cues to deception. Section 3 describes the dataset used for this work, and Section 4 details the different feature sets we employ. In Section 5, we report on the results of our empirical study of indicators of deception and perceived deception, as well as gender and native language differences. Section 6 presents our machine learning classification results using the deception indicator feature sets. We conclude in Section 7 with a discussion and ideas

for future work.

2 Related Work

Language-based cues to deception have been analyzed in many genres. Ott et al. (2011) compared approaches to automatically detecting deceptive opinion spam, using a crowdsourced dataset of fake hotel reviews. Several studies use a fake opinion paradigm for collecting data, instructing subjects to write or record deceptive and truthful opinions about controversial topics such as the death penalty or abortion, or about a person that they like/dislike (Newman et al., 2003; Mihalcea and Strapparava, 2009). Other research has focused on real-world data obtained from court testimonies and depositions (Fornaciari and Poesio, 2013; Bachenko et al., 2008; Pérez-Rosas et al., 2015). Real-world deceptive situations are high-stakes, where there is much to be gained or lost if deception succeeds or fails; it is hypothesized that these conditions are more likely to elicit strong cues to deception. However, working with such data requires extensive research to annotate each utterance for veracity, so such datasets are often quite small and not always reliable.

Linguistic features such as n-grams and language complexity have been analyzed as cues to deception (Pérez-Rosas and Mihalcea, 2015; Yancheva and Rudzicz, 2013). Syntactic features such as part of speech tags have also been found to be useful for structured data (Ott et al., 2011; Feng et al., 2012). Statement Analysis (Adams, 1996) is a text-based deception detection approach that combines lexical and syntactic features. An especially useful resource for text-based deception detection is the Linguistic Inquiry and Word Count (LIWC) (Pennebaker and King, 1999), which groups words into psychologically motivated categories. In addition to lexical features, some studies have examined acoustic-prosodic cues to deception (Rockwell et al., 1997; Enos, 2009; Mendels et al., 2017). (Benus et al., 2006) studied pause behavior in deceptive speech. This work is very promising, but it is more difficult to obtain large, cleanly recorded speech corpora with deception annotations than to obtain text corpora. An excellent meta-study of verbal cues to deception can be found in (DePaulo et al., 2003).

3 Data

3.1 Corpus

For this work, we examined the Columbia X-Cultural Deception (CXD) Corpus (Levitan et al., 2015a) a collection of within-subject deceptive and non-deceptive speech from native speakers of Standard American English (SAE) and Mandarin Chinese (MC), all speaking in English. The corpus contains dialogues between 340 subjects. A variation of a fake resume paradigm was used to collect the data. Previously unacquainted pairs of subjects played a ”lying game” with each other. Each subject filled out a 24-item biographical questionnaire and were instructed to create false answers for a random half of the questions. They also reported demographic information including gender and native language, and completed the NEO-FFI personality inventory (Costa and McCrae, 1989).

The lying game was recorded in a sound booth. For the first half of the game, one subject assumed the role of the interviewer, while the other answered the biographical questions, lying for half and telling the truth for the other; questions chosen in each category were balanced across the corpus. For the second half of the game, the subjects roles were reversed, and the interviewer became the interviewee. During the game, the interviewer was allowed to ask the 24 questions in any order s/he chose; the interviewer was also encouraged to ask follow-up questions to aid them in determining the truth of the interviewees answers. Interviewers recorded their judgments for each of the 24 questions, providing information about human perception of deception. The entire corpus was orthographically transcribed using the Amazon Mechanical Turk (AMT)[1] crowd-sourcing platform, and the speech was segmented into *inter-pausal units* (IPUs), defined as pause-free segments of speech separated by a minimum pause length of 50 ms. The speech was also segmented into turn units, where a turn is defined as a maximal sequence of IPUs from a single speaker without any interlocutor speech that is not a *backchannel*. There are two forms of deception annotations in the corpus: local and global. Interviewees labeled their responses with local annotations by pressing a ”T” or ”F” key for each utterance as they spoke. These keypresses were automatically aligned with speaker IPUs and turns. Global la-

[1] https://www.mturk.com/mturk/

bels were provided by the biographical questionnaire, where each of the 24 questions was labeled as truthful or deceptive.

Consider the following dialogue:
Interviewer: What is your mother's job?
Interviewee: My mother is a doctor (F). She has always worked very late hours and I felt neglected as a child (T).

Is the interviewee response true or false? We differentiate between global and local deception. Globally, the response to the question is deceptive. However, it contains local instances of both truth and deception. In this work we focus on dialogue-based deception, using global deception labels.

3.2 Global Segmentation

Previous work with the CXD corpus has focused on IPU-level and turn-level analysis and classification of local deception, mostly with acoustic-prosodic features (Levitan et al., 2015b; Mendels et al., 2017). Here we are interested in exploring global deception at the dialogue-level for the first time in this corpus. We define response-segments as sets of turns that are related to a single question (of the 24 interview questions). In order to annotate these segments, we first used a question detection and identification system (Maredia et al., 2017) that uses word embeddings to match semantically similar variations of questions to a target question list. This was necessary because interviewers asked the 24 questions using different wording from the original list of questions. On this corpus, (Maredia et al., 2017) obtained an F1-score of .95%.

After tagging interviewer turns with this system, we labeled the set of interviewee turns between two interviewer questions q1 and q2 as corresponding to question q1. The intuition behind this was that those turns were responses to follow up questions related to q1, and while the question detection and identification system discussed above did not identify follow up questions, we found that most of the follow up questions after an interviewer question q1 would be related to q1 in our hand annotation. We evaluated this global segmentation on a hand-annotated test set of 17 interviews (about 10% of the corpus) consisting of 2,671 interviewee turns, 408 interviewer questions, and 977 follow up questions. Our global segmentation approach resulted in 77.8% accuracy on our hand-labeled test set (errors were mostly due to turns that were unrelated to any question).

We performed our analysis and classification on two segmentations of the data using this tagging method: (1) **first turn:** we analyzed only the single interviewee turn directly following the original question, and (2) **multiple turns** we analyzed the entire segment of interviewee turns that were responding to the original interviewer question and subsequent follow-up questions. In our classification experiments, we explore whether a deceptive answer is be better classified by the interviewee's initial response or by all of the follow-up conversation between interviewer and interviewee.

4 Features

LIWC Previous work has found that deceivers tend to use different word usage patterns when they are lying (Newman et al., 2003). We used LIWC (Pennebaker et al., 2001) to extract semantic features from each utterance. LIWC is a text analysis program that computes features consisting of normalized word counts for 93 semantic classes. LIWC dimensions have been used in many studies to predict outcomes including personality (Pennebaker and King, 1999), deception (Newman et al., 2003), and health (Pennebaker et al., 1997). We extracted a total of 93 features using LIWC 2015 [2], including standard linguistic dimensions (e.g. percentage of words that are pronouns, articles), markers of psychological processes (e.g. affect, social, cognitive), punctuation categories (e.g periods, commas), and formality measures (e.g. fillers, swear words).

Linguistic We extracted 23 linguistic features [3] which we adopted from previous deception studies such as (Enos, 2009; Bachenko et al., 2008). Included in this list are binary and numeric features capturing hedge words, filled pauses, laughter, complexity, contractions, and denials. We include Dictionary of Affect Language (DAL) (Whissell et al., 1986) scores that measure the emotional meaning of texts, and a specificity score which measures level of detail (Li and Nenkova, 2015). The full list of features is: 'hasAbsolutelyReally', 'hasContraction', 'hasI', 'hasWe', 'hasYes', 'hasNAposT' (turns

[2] A full description of the features is found here: https://s3-us-west-2.amazonaws.com/downloads.liwc.net/LIWC2015_OperatorManual.pdf

[3] A detailed explanation of these linguistic features and how they were computed is found here: http://www.cs.columbia.edu/speech/cxd/features.html

that contain words with the contraction "n't"), 'hasNo', 'hasNot', 'isJustYes', 'isJustNo', 'noYesOrNo', 'specificDenial', 'thirdPersonPronouns', 'hasFalseStart', 'hasFilledPause', 'numFilledPauses', 'hasCuePhrase', 'numCuePhrases', 'hasHedgePhrase', 'numHedgePhrases', 'hasLaugh', 'complexity', 'numLaugh', 'DAL-wc', 'DAL-pleasant', 'DAL-activate', 'DAL-imagery', 'specScores' (specificity score).

Response Length Previous work has found that response length, in seconds, is shorter in deceptive speech, and that the difference in number of words in a segment of speech is insignificant between deceptive and truthful speech (DePaulo et al., 2003). For our question-level analysis, we used four different measures for response length: the total number of seconds of an interviewee response-segment, the total number of words in an interviewee response-segment, the average response time of a turn in an interviewee response-segment, and the average number of words per turn in an interviewee response-segment.

Individual Traits We analyzed gender and native language of the speakers to determine if these traits were related to ability to deceive and to detect deception. We also analyzed linguistic cues to deception across gender and native language, and used gender and native language information in our classification experiments. All speakers were either male or female, and their native language was either Standard American English or Mandarin Chinese. In addition, we used the NEO-FFI (5 factor) personality inventory scores as features in classification experiments, but not for the statistical analysis in this paper.

Follow-up Questions Follow-up questions are questions that an interviewer asks after they ask a question from the original prescribed set of questions. We hypothesized that if an interviewer asked more follow-up questions, they were more likely to identify deceptive responses, because asking follow-up questions indicated interviewer doubt of the interviewee's truthfulness. For each interviewee response-segment, we counted the number of follow-up questions interviewees were asked by the interviewer.

5 Analysis

In order to analyze the differences between deceptive and truthful speech, we extracted the above features from each question response-segment, and calculated a series of paired t-tests between the features of truthful speech and deceptive speech. All tests for significance correct for family-wise Type I error by controlling the false discovery rate (FDR) at $\alpha = 0.05$. The k^{th} smallest p value is considered significant if it is less than $\frac{k*\alpha}{n}$.

5.1 Interviewee Responses

Table 1 shows the features that were statistically significant indicators of truth and deception in interviewee response-segments consisting of multiple turns. Below, we highlight some interesting findings.

In contrast to (DePaulo et al., 2003), we found that the total duration of an interviewee response-segment was longer for deceptive speech than for truthful speech. Additionally, while (DePaulo et al., 2003) showed that the number of words in a segment of speech was not significantly different between deceptive and truthful speech, we found that deceptive response-segments had more words than truthful response-segments. Furthermore, we found that longer average response time per turn and more words per sentence were significant indicators of deception. These results show that when interviewees are trying to deceive, not only is their aggregate response longer in duration and number of words, but their individual responses to each follow-up question are also longer. Consistent with (DePaulo et al., 2003), we found that more filled pauses in an interviewee response-segment was a significant indicator of deception. Deceivers are hypothesized to experience an increase in cognitive load (Vrij et al., 1996), and this can result in difficulties in speech planning, which can be signaled by filled pauses. Although (Benus et al., 2006) found that, in general, the use of pauses correlates more with truthful than with deceptive speech, we found that filled pauses such as "um" were correlated with deceptive speech. The LIWC *cogproc* (cognitive processes) dimension, which includes words such as "cause", "know", "ought" was significantly more frequent in truthful speech, also supporting the theory that cognitive load is increased while practicing deception.

We found that increased $DALimagery$ scores, which compute words often used in speech to create vivid descriptions, were indicators of deception. We also found that the LIWC language summary variables of $authenticity$ and $adjectives$

Feature	Deception	Truth	Neutral
Lexical	DAL.activate, DAL.imagery, DAL.pleasant noYesOrNo, numCuePhrase, numFilledPauses, numHedgePhrases specScores, thirdPersonPronouns	isJustNo	complexity, DAL.wc isJustYes numLaugh specificDenial
LIWC	achieve, adj, adverb, affiliation analytic, article, authentic cause, clout, compare, conj dash, discrep, drives, family feel, focusfuture, focuspast, friend health, interrog, ipron, male motion, percept, ppron, prep pronoun, power, relativ, reward shehe, social, space, swear verb, WC, we, WPS, you	certain, dic function, negate, netspeak	affect, apostro, assent auxverb, body, cogproc colon, comma, death differ, female, filler i, ingest, insight, leisure posemo, quant, quote relig, sad, see sixltr, they, tone, work
Response length	num words, response length avg response length, avg num words		
Followup	num turns		

Table 1: Statistically significant indicators of truth and deception in interviewee response-segments consisting of multiple turns related to a single question.

Feature	Deception	Truth	Neutral
Lexical	DAL.imagery, DAL.pleasant numCuePhrases, numFilledPauses numHedgePhrases, specificDenial specScores, thirdPersonPronoun	DAL.actvate isJustNo	complexity, DAL.wc isJustYes, noYesOrNo numLaugh
LIWC	adverb, article,authentic, body conj, focuspast, interrog, ipron prep, pronoun, WC, WPS	negate	apostro, bio, cause certain, clout, cogproc, compare discrep, focusfuture, function insight, money, motion negemo, nonflu, number posemo, ppron, relative
Response length	num words response length avg num words avg response length		
Followup	num turns		

Table 2: Statistically significant indicators of perceived truth and deception in interviewer judgments of interviewee responses.

were indicators of deception: in an effort to sound more truthful and authentic, interviewees may have provided a level of detail that is uncharacteristic of truthful speech. Similarly, the $specificity$ metric was indicative of deception: deceptive responses contained more detailed language. Words in the LIWC $clout$ category - a category describing words that indicate power of influence - were more prevalent in deceptive responses, suggesting that subjects sounded more confident while lying. $Interrogatives$ were an indicator of deception. In the context of the interviewer-interviewee paradigm, these are interviewee questions to the interviewer. Perhaps this was a technique used to stall so that they had more time to develop an answer (e.g. "Can you repeat the question?"), or to deflect the interviewer's attention from their deception and put the interviewer on the spot. We observed that $hedge$ words and phrases, which speakers use to distance themselves from a proposition, were more frequent in deceptive speech. This is consistent with Statement Analysis (Adams, 1996), which posits that hedge words are used in deceptive statements to intentionally create vagueness that obscures facts. Consistent with this finding, $certainty$ in language (words such as "always" or "never") was a strong indicator of truthfulness.

It is also interesting to note the features that were not significant indicators of truth or decep-

tion. For example, there was no significant difference in laughter frequency or apostrophes (used for contractions in this corpus) between truthful and deceptive responses.

When we compared indicators of truth vs. deception across multiple turns to indicators of truth vs. deception in just the first turns of interviewee response-segments, we found that, generally, indicators in first turns are a subset of indicators across multiple turns. In some cases there were interesting differences. For example, although $tone$ (emotional tone - higher numbers indicate more positive, and lower indicate negative) was not a significant indicator of deception for the entire interviewee response-segment, negative tone was a moderate indicator of deception in first turns. This suggests that the tone of interviewees, when they have just started their lie, is different from when they are given the opportunity to expand on that lie. The findings from our analysis of first turns suggest that there might be enough information in the first response alone to distinguish between deceptive and truthful speech; we test this in our classification experiments in Section 6.

5.2 Interviewer Judgments of Deception

In addition to analyzing the linguistic differences between truthful and deceptive speech, we were interested in studying the characteristics of speech that is believed or disbelieved. Since the CXD corpus includes interviewer judgments of deception for each question asked, we have the unique opportunity to study human perception of deception on a large scale. Table 2 shows the features that were statistically significant indicators of truth and deception in interviewee responses - consisting of multiple turns - that were perceived as true or false by interviewers. Here we highlight some interesting findings. There were many features that were prevalent in speech that interviewers perceived as deceptive, which were in fact cues to deception. For example, speech containing more words in a response-segment and more words per sentence was generally perceived as deceptive by interviewers, and indeed, this perception was correct. Disbelieved answers had a greater frequency of filled pauses and hedge words, and greater specificity, all of which were increased in deceptive speech.

There were also several features that were indicators of deception, but were not found in higher rates in statements that were perceived as false. For example, the LIWC dimensions $clout$ and $certain$ were not significantly different in believed vs. disbelieved interviewee responses, but $clout$ was increased in deceptive speech and $certain$ language was increased in truthful speech. There were also features that were significantly different between believed and disbelieved statements, but were not indicators of deception. For example, statements that were perceived as false by interviewers had a greater proportion of $specificDenials$ (e.g. "I did not") than those that were perceived as true; this was not a valid cue to deception. Number of turns was increased in dialogue segments where the interviewer did not ultimately believe the interviewee response. That is, more follow up questions were asked when an interviewer did not believe their interlocutor's response, which is an intuitive behavior. When we compared indicators of speech that was perceived as deceptive across multiple turns to indicators of speech that was perceived as deceptive in just the first turns, we found that, generally, indicators in first turns are a subset of indicators across multiple turns.

On average, human accuracy at judging truth and deception in the CXD corpus was 56.75%, and accuracy at judging deceptive statements only was 47.93%. The average F1-score for humans was 46. Thus, although some cues were correctly perceived by interviewers, humans were generally poor at deception perception. Nonetheless, characterizing the nature of speech that is believed or not believed is useful for applications where we would ultimately like to synthesize speech that is trustworthy.

5.3 Gender and Native Language Differences in Deception Behavior

Having discovered many differences between deceptive and truthful language across all speakers, we were interested in analyzing differences in deceptive language across groups of speakers. Using gender and native language (English or Mandarin Chinese) as group traits, we conducted two types of analysis. First, we directly compared deception performance measures (ability to deceive as interviewee, and ability to detect deception as interviewer) between speakers with different traits, to assess the effect of individual characteristics on deception abilities. In addition, we compared the features of deceptive and truthful language in sub-

Group	Deception	Truth
Male Female	analytic, friend, interrog achieve, adverb, article authentic, cause compare discrep,family, feel focusfuture, percept, power relativ, we	posemo
English	acheve, adverb, affiliation compare, interrog, power relativ, space, swear	
Chinese	analytic, bio cause discrep, feel, health percep, (filler)	certain (informal) (netspeak)

Table 3: Gender-specific and language-specific indicators of deception and truth. We consider a result to approach significance if its uncorrected p value is less than 0.05 and indicate this with () in the table.

sets of the corpus, considering only people with a particular trait, in order to determine group-specific patterns of deceptive language. As before, tests for significance correct for family-wise Type I error by controlling the false discovery rate (FDR) at $\alpha = 0.05$. The k^{th} smallest p value is considered significant if it is less than $\frac{k*\alpha}{n}$.

5.3.1 Gender

There were no significant differences in deception ability between male and female speakers. However, there were many differences in language between male and female speakers. Further, some features were only discriminative between deception and truth for a specific gender. Table 3 shows linguistic features that were significantly different between truthful and deceptive speech, but only for one gender. In some cases the feature was found in different proportions in male and females, and in other cases there was no significant difference. For example, $family$ words were indicative of deception only in female speakers, and these words were also used more frequently by female speakers than male speakers.

The LIWC category of $compare$ was also indicative of deception for females only, and this feature was generally found more frequently in female speech. Article usage was only significantly different between truthful and deceptive speech in females (more articles were found in deceptive speech), but articles were used more frequently in male speech. On the other hand, the LIWC category of $posemo$ (positive emotion) was increased in truthful speech for male speakers only, and there was no significant difference of $posemo$ frequency across gender.

5.3.2 Native Language

Interviewees were more successful at deceiving native Chinese speakers than at deceiving native English speakers ($t(170) = -2.13, p = 0.033$). This was true regardless of interviewee gender and native language, and slightly stronger for female interviewers ($t(170) = -2.22, p = 0.027$). When considering only female interviewers, interviewees were more successful at deceiving non-native speakers than native speakers, but this difference was not significant when considering only male interviewers. As with gender, there were several features that were discriminative between deception and truth for only native speakers of English, or only native speakers of Mandarin. Table 3 shows LIWC categories and their relation to deception, broken down by native language. For example, power words were found more frequently in deception statements, when considering native English speakers only. In general, power words were used more by native Mandarin speakers than by native English speakers. LIWC categories of $compare$, $relative$, and $swear$ were more prevalent in deceptive speech, only for English speakers. On the other hand, $feel$ and $perception$ dimensions were only indicators of deception for native Mandarin speakers, although there was no significant difference in the use of these word categories across native language. $Informal$ and $netspeak$ word dimensions tended to be more frequent in truthful speech for native Chinese speakers only (approaching significance), and these word categories were generally more frequent in native Mandarin speech. Finally, $filler$ words tended to be more frequent in deceptive speech (approaching significance) only for native Mandarin speakers, and these were used more frequently by native Mandarin speakers than native English speakers.

Overall, our findings suggest that deceptive behavior in general, and deceptive language in particular, are affected by a person's individual characteristics, including gender and native language. When building a deception classification system, it is important to account for this variation across speaker groups.

Features	Segmentation	Accuracy	Precision	Recall	F1-score
Human baseline	Multiple turns	56.75	56.50	40.00	46.50
LIWC	Single turn	65.75	65.79	65.74	65.72
	Multiple turns	72.78	72.84	72.74	72.74
Lexical	Single turn	66.95	66.97	66.95	66.94
	Multiple turns	70.33	70.46	70.28	70.25
LIWC+lexical	Single turn	68.35	68.36	68.35	68.35
	Multiple turns	71.66	71.77	71.60	71.58
LIWC+individual	Single turn	67.50	67.50	67.50	67.49
	Multiple turns	71.85	71.93	71.80	71.79
Lexical+individual	Single turn	69.32	69.33	69.32	69.31
	Multiple turns	69.95	70.06	69.89	69.86
LIWC+lexical+individual	Single turn	70.87	70.87	70.87	70.87
	Multiple turns	72.40	72.50	72.34	72.33

Table 4: Random Forest classification of single turn and multiple turn segmentations, using text-based features and individual traits (gender, native language, NEO-FFI personality scores).

6 Deception Classification

Motivated by our analysis showing many significant differences in the language of truthful and deceptive responses to interview questions, we trained machine learning classifiers to automatically distinguish between truthful and deceptive text, using the feature sets described in section 4. We compared classification performance for the two segmentation methods described in section 3.2: first turn and multiple turns. This allowed us to explore the role of context in automatic deception detection. When classifying interviewee response-segments, should the immediate response only be used for classification, or is inclusion of surrounding turns helpful? This has implications not only for deception classification, but for practitioners as well. Should human interviewers make use of responses to follow up questions when determining response veracity, or should the initial response receive the most consideration?

We compared the performance of 3 classification algorithms: Random Forest, Logistic Regression, and SVM (sklearn implementation). In total, there were 7,792 question segments for both single turn and multiple turns segmentations. We divided this into 66% train and 33% test, and used the same fixed test set in experiments for both segmentations in order to directly compare results. The random baseline performance is 50, since the dataset is balanced for truthful and deceptive statements. Another baseline is human performance, which is 46.0 F1 in this corpus. The Random Forest classifier was consistently the best performing, and we only report those results due to space constraints. Table 4 displays the classification performance for each feature set individually, as well as feature combinations, for both single turn and multiple turn segmentations. It also shows the human baseline performance, obtained from the interviewers' judgments of deception in the corpus, which were made after asking each question along with related follow-up questions (i.e. multiple turn segmentation).

The best performance (72.74 F1-score) was obtained using LIWC features extracted from multiple turns. This is a 22.74% absolute increase over the random baseline of 50, and a 26.74% absolute increase over the human baseline of 46. The performance of classifiers trained on multiple turns was consistently better than those trained on single turns, for all feature sets. For multiple turns, LIWC features were better than the lexical feature set, and combining lexical with LIWC features did not improve over the performance of LIWC features alone. Adding individual traits information was also not beneficial. However, when considering the first turn only, the best results (70.87 F1-score) were obtained using a combination of LIWC+lexical+individual features. Using the first turns segmentation, lexical features were slightly better than LIWC features, and interestingly, adding individual traits helped both feature sets. A combination of LIWC and lexical features was better than each on its own.

These results suggest that contextual informa-

tion, in the form of follow up questions, is beneficial for deception classification. It seems that individual traits, including gender, native language, and personality scores, are helpful in deception classification under the condition where contextual information is not available. When the contextual information is available, the the additional lexical content is more useful than individual traits.

7 Conclusions and Future Work

In this paper we presented a study of deceptive language in interview dialogues. Our analysis of linguistic characteristics of deceptive and truthful speech provides insight into the nature of deceptive language. We also analyzed the linguistic characteristics of speech that is *perceived* as deceptive and truthful, which is important for understanding the nature of trustworthy speech. We explored variation across gender and native language in linguistic cues to deception, highlighting cues that are specific to particular groups of speakers. We built classifiers that use combinations of linguistic features and individual traits to automatically identify deceptive speech. We compared the performance of using cues from the single first turn of an interviewee response-segment with using cues from the full context of multiple interviewee turns, achieving performance as high as 72.74% F1-score (about 27% better than human detection performance).

This work contributes to the critical problem of automatic deception detection, and increases our scientific understanding of deception, deception perception, and individual differences in deceptive behavior. In future work, we plan to conduct similar analysis in additional deception corpora in other domains, in order to identify consistent domain-independent deception indicators. In addition, we plan to conduct cross-corpus machine learning experiments, to evaluate the robustness of these and other feature sets in deception detection. We also would like to explore additional feature combinations, such as adding acoustic-prosodic features. Finally, we plan to conduct an empirical analysis of deception behavior across personality types.

Acknowledgments

This work was partially funded by AFOSR FA9550-11-1-0120 and by NSF DGE-11-44155. Thank you to Bingyan Hu for her assistance with feature extraction. We thank the anonymous reviewers for their helpful comments.

References

Susan H Adams. 1996. Statement analysis: What do suspects' words really reveal. *FBI L. Enforcement Bull.* 65:12.

Joan Bachenko, Eileen Fitzpatrick, and Michael Schonwetter. 2008. Verification and implementation of language-based deception indicators in civil and criminal narratives. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*. Association for Computational Linguistics, pages 41–48.

Stefan Benus, Frank Enos, Julia Hirschberg, and Elizabeth Shriberg. 2006. Pauses in deceptive speech. In *Speech Prosody*. volume 18, pages 2–5.

PT Costa and RR McCrae. 1989. Neo five-factor inventory (neo-ffi). *Odessa, FL: Psychological Assessment Resources* .

Bella M DePaulo, James J Lindsay, Brian E Malone, Laura Muhlenbruck, Kelly Charlton, and Harris Cooper. 2003. Cues to deception. *American Psychological Association, Inc.* pages 74–118.

Frank Enos. 2009. *Detecting deception in speech*. Ph.D. thesis, Citeseer.

Song Feng, Ritwik Banerjee, and Yejin Choi. 2012. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*. Association for Computational Linguistics, pages 171–175.

Tommaso Fornaciari and Massimo Poesio. 2013. Automatic deception detection in italian court cases. *Artificial intelligence and law* 21(3):303–340.

Sarah I Levitan, Guzhen An, Mandi Wang, Gideon Mendels, Julia Hirschberg, Michelle Levine, and Andrew Rosenberg. 2015a. Cross-cultural production and detection of deception from speech. In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*. ACM, pages 1–8.

Sarah I Levitan, Guzhen An, Mandi Wang, Gideon Mendels, Julia Hirschberg, Michelle Levine, and Andrew Rosenberg. 2015b. Cross-cultural production and detection of deception from speech. In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*. ACM, pages 1–8.

Junyi Jessy Li and Ani Nenkova. 2015. Fast and accurate prediction of sentence specificity. In *Proceedings of the Twenty-Ninth Conference on Artificial Intelligence (AAAI)*. pages 2281–2287.

Angel S Maredia, Kara Schechtman, Sarah I Levitan, and Julia Hirschberg. 2017. Comparing approaches for automatic question identification. SEM.

Gideon Mendels, Sarah Ita Levitan, Kai-Zhan Lee, and Julia Hirschberg. 2017. Hybrid acoustic-lexical deep learning approach for deception detection. *Proc. Interspeech 2017* pages 1472–1476.

Rada Mihalcea and Carlo Strapparava. 2009. The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*. Association for Computational Linguistics, pages 309–312.

Matthew L Newman, James W Pennebaker, Diane S Berry, and Jane M Richards. 2003. Lying words: Predicting deception from linguistic styles. *Personality and social psychology bulletin* 29(5):665–675.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, pages 309–319.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates* 71:2001.

James W Pennebaker and Laura A King. 1999. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology* 77(6):1296.

James W Pennebaker, Tracy J Mayne, and Martha E Francis. 1997. Linguistic predictors of adaptive bereavement. *Journal of personality and social psychology* 72(4):863.

Verónica Pérez-Rosas, Mohamed Abouelenien, Rada Mihalcea, and Mihai Burzo. 2015. Deception detection using real-life trial data. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, pages 59–66.

Verónica Pérez-Rosas and Rada Mihalcea. 2015. Experiments in open domain deception detection. In *Proceedings of EMNLP 2015*. ACL, pages 1120–1125.

Patricia Rockwell, David B Buller, and Judee K Burgoon. 1997. The voice of deceit: Refining and expanding vocal cues to deception. *Communication Research Reports* 14(4):451–459.

Aldert Vrij, Gun R Semin, and Ray Bull. 1996. Insight into behavior displayed during deception. *Human Communication Research* 22(4):544–562.

Cynthia Whissell, Michael Fournier, René Pelland, Deborah Weir, and Katherine Makarec. 1986. A dictionary of affect in language: Iv. reliability, validity, and applications. *Perceptual and Motor Skills* 62(3):875–888.

Maria Yancheva and Frank Rudzicz. 2013. Automatic detection of deception in child-produced speech using syntactic complexity features. In *ACL (1)*. pages 944–953.