Building a Data Collection for Deception Research

Eileen Fitzpatrick
Montclair State University
Montclair, NJ 07043
fitzpatricke@mail.montclair.edu

Joan Bachenko
Linguistech Consortium, Inc.
Oxford, NJ 07863
jbachenko@linguistech.com

Abstract

Research in high stakes deception has been held back by the sparsity of ground truth verification for data collected from real world sources. We describe a set of guidelines for acquiring and developing corpora that will enable researchers to build and test models of deceptive narrative while avoiding the problem of sanctioned lying that is typically required in a controlled experiment. Our proposals are drawn from our experience in obtaining data from court cases and other testimony, and uncovering the background information that enabled us to annotate claims made in the narratives as true or false.

1 Introduction

The ability to spot deception is an issue in many important venues: in police, security, border crossing, customs, and asylum interviews; in congressional hearings; in financial reporting; in legal depositions; in human resource evaluation; and in predatory communications, including Internet scams, identity theft, and fraud. The need for rapid, reliable deception detection in these high stakes venues calls for the development of computational applications that can distinguish true from false claims.

Our ability to test such applications is, however, hampered by a basic issue: the ground truth problem. To be able to recognize the lie, the researcher must not only identify distinctive behavior when someone is lying but must ascertain whether the statement being made is true or not.

The prevailing method for handling the ground truth problem is the controlled experiment, where truth and lies can be managed. While controlled laboratory experiments have yielded important insights into deceptive behavior, ethical and proprietary issues have put limits on the extent to which controlled experiments can model deception in the "real world". High stakes deception cannot be simulated in the laboratory without serious ethics violations. Hence the motivation to lie is weak since subjects have no personal loss or gain at stake. Motivation is further compromised when the lies are sanctioned by the experimenter who directs and condones the lying behavior (Stiff et al., 1994). With respect to the studies themselves, replication of laboratory deception research is rarely done due to differences in data sets and subjects used by different research groups. The result, as Vrij (2008) points out, is a lack of generalizability across studies.

We believe that many of the issues holding back deception research could be resolved through the construction of standardized corpora that would provide a base for expanding deception studies, comparing different approaches and testing new methods. As a first step towards standardization, we offer a set of practical guidelines for building corpora that are customized for studies of high stakes deception. The guidelines are based on our experiences in creating a corpus of real world language data that we used for testing the deception detection approach described in Bachenko et al. (2008), Fitzpatrick and Bachenko (2010). We hope that our experience will encourage other researchers to build and contribute corpora with the goal of establishing a shared resource that passes the test of ecological validity.

Section 2 of the paper describes the data collection initiative we are engaged in, section 3 describes the methods used to corroborate the claims in the data, section 4 concludes our account and covers lessons learned.

We should point out that the ethical considerations that govern our data collection are subject to the United States Code of Federal

Proceedings of the EACL 2012 Workshop on Computational Approaches to Deception Detection, pages 31–38,
Avignon, France, April 23 - 27 2012.

Regulations (CFRs) for the protection of human subjects and may differ in some respects from those in other countries.

2 Collecting High-Stakes Data

We are building a corpus of spoken and written narrative data used in real world high stakes cases in which many of the claims in the corpus have been corroborated as True or False. We have corroborated claims in almost 35,090 words of narrative. These narratives include statements to police, a legal deposition, and congressional testimony.

In assembling and managing our corpus, two issues have been paramount: the availability of data and constraints on its use. Several types of information must be publicly available, including the primary linguistic data, background information used to determine ground truth, and general information about the case or situation from which the data is taken. In addition, the data must be narrative intensive. There are also several considerations about the data that must be taken into account, including the mode (written or spoken) of the narrative, and considerations involving the needs of the users of the data.

To ensure unconstrained access, data collection must be exempt from human participant restrictions. The restrictions we must adhere to are the regulations of Title 46 of the CFRs.[1] 46 CFR 102 lists the data that is exempt from human participant restrictions. Exempt data includes "[r]esearch involving the collection or study of existing data, documents, records, pathological specimens, or diagnostic specimens, if these sources are publicly available or if the information is recorded by the investigator in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects."

46 CFR 111, section 7 covers protection of privacy: "When appropriate, there are adequate provisions to protect the privacy of subjects and to maintain the confidentiality of data."

It is conceivable that a "real world" high stakes study could involve subjects whose identifiable data would be removed from the collection, but it is highly unlikely that the subjects would consent to having their data – even if sanitized – made available on the Internet. We have therefore used only exempt data, i.e., data that is publicly available with no expectation of privacy on the part of the people involved.

2.1 Public availability of data

There is a large body of narrative data in the public domain, data that is also likely to have a rich source of ground truth evidence and general background information. Typical public sources for this data would be crime investigation websites, published police interviews, legal websites, including findlaw.com and justice.gov, quarterly earnings conference calls, and the U.S. Congressional Record. Such data includes publicly available

- Face-to-face interviews
- Depositions
- Court and other public testimony
- Phone conversations[2]
- Recorded statements to police
- Written statements to police
- Debates of political figures and candidates for public office
- Online product endorsements
- Blogs
- Webpages

High profile cases are particularly well represented on websites. In the U.S., police reports, which are a matter of public record, may also be obtained for a small fee from local police departments. Other data aggregators, like FactSet.com, provide data for higher fees.

2.2 Types of Data

2.2.1 Primary linguistic data

The narrative data is the data to be analyzed for cues to deception. Written data is, of course, available as text, but spoken data may also only be available as transcripts. Our current dataset includes recorded data only from the Enron testimony, but ideally speech data would include high quality recorded speech to enable analysis of the prosodic qualities of the speech.

To support robust analysis, it is important that the data be narrative intense. The 'yes'/'no'

[1] These regulations are enforced either by the Institutional Review Board (IRB) of the institution where the research takes place or by an independent IRB contracted by the researchers if there is no housing institution.

[2] For example, the quarterly earnings conference calls analyzed in Larcker and Zakolyukina (2010).

responses of a polygraph interview are not usable for language analysis.

Additionally, we have so far limited our collection to spontaneously produced data. Prepared, rehearsed narrative provides the opportunity to carefully craft the narrative putting the narrator in control not only of the story but of the language used to convey the story. This enables the speaker/writer to avoid the cues that we are looking for. We would be open to adding prepared data to the collection, but have not considered the guidelines for it.

2.2.2 Background data

Background information on the primary data is the basis for the ground truth annotation of the claims made in the primary data. Ground truth investigation can use various types of information, including that coming from interviews, police reports, public records posted on local and national government web sites, fact checking sites like FactCheck.org [3] and PolitiFact.com[4] that analyze political claims and provide sources for the information they use in their own judgments, and websites such as truTV.com that offer the facts of a case, the final court judgment, and interviews with the people involved in the case.

Many of these sources are available on the web – an advantage of using data where there is no expectation of privacy.[5] Some data requires filing for a police report or a court document. The sources for our current data set are given in Appendix A.

Another source of verification can be the narrative itself in situations where the narrator contradicts a prior claim. For example, one narrator, after denying a theft for most of the interview, says "All right, man, I did it," enabling us to mark his previous denials as False.

2.2.3 General information about the case/situation

Ideally, the corpus will include background information on the situation covered by the narrative. If the situation is a legal case, the background information should include the verdict of the judge or jury, the judgment of conviction given by the judge, and the sentence. If the case is on appeal, then that should be noted.

Information on the amount of control the narrator has over the story is also valuable. Is the narrative elicited or freely given? The former gives the narrator less control over the narrative, possibly increasing the odds for the appearance of cues to deception. Is the narrator offering a monologue or a written statement, both of which give the author more control of the narrative than an interview.

2.2.4 Speaker information

General information on the speaker can be valuable in gauging the performance of a deception model, including information on gender, age, and education. We found information on first language background and culture to be useful in analyzing the speech of non-native speakers of English, whose second language speech characteristics sometimes align with deception cues. Other sociolinguistic traits may also be important, although we have found that, while sociolinguistic background may determine word choice, the deceptive behavior is invariant. We have not encountered issues of competency to stand trial in the criminal cases we have included, but such evaluations should be noted if the issue arises in a legal case.

2.2.5 Spoken and written data

Two of the narratives in our current collection are written; the others are spoken. Both written statements were produced as parts of a police interview. The purpose of requesting the statement is to obtain an account in the interviewee's own words and to do this before time and questioning affect the interviewee's thinking. Hence the written statement is analogous to a lengthy interview answer, and the language used is much closer to speech than writing, as the opening of the Routier statement illustrates:

Darin and my sister Dana came home from working at the shop. The boys were playing with the neighborhood kids outside. I was finishing up dinner.

2.3 Other considerations

In providing data for general use by researchers, the collector must be aware of varying needs of researchers using the data. The general needs we

[3] FactCheck is a project of the Annenberg Public Policy Center of the University of Pennsylvania.

[4] PolitiFact is sponsored by the Tampa Bay Times.

[5] Information may be withdrawn from the web, however, if there are changes in a case, such as the filing of an appeal or simply fading interest in the case.

consider are the ground truth yield and the question of the scope of the True/False label.

2.3.1 Ground truth yield

The amount of background data that can be gathered to yield ground truth judgments can vary widely depending on the type of narrative data collected. We have worked with private criminal data where the ratio of verified propositions to words in the primary data is as high as .049 and with private job interview data where the ratio is as low as .00043. The low yield may be problematic for some types of experiment, as well as frustrating for the data collector. It is important to have some assurance that there are a reasonable number of resources that can provide ground truth data before collecting the narrative data, particularly if the narrative data is difficult to collect.

2.3.2 The Scope of the T/F label

With the exception of Fornaciari and Poesio (2011), Hirschberg et al. (2005), Bachenko et al. (2008) and Fitzpatrick and Bachenko (2010), the ML/NLP deception literature distinguishes True from False at the level of the narrative, not the proposition. In other words, most of the studies identify the liar, not the lie. For real world data, the choice to label the full narrative as True or False usually depends on the length of the narrative; a narrator giving trial testimony or a job interview will have many claims, while someone endorsing a product may have just one: this product is good.

There are high stakes narratives that are short, such as TSA airport interviews. However, the computational models of such data will be different from those of longer narratives where true and false statements are interspersed throughout. We currently have no data of this type.

3 Providing Ground Truth

In longer real-world narratives people lie selectively and the interviewer usually needs to figure out which statements, or propositions, are lies. To enable the capture of this situation in a model, we engage in a two-step process: the scope of selected verifiable propositions in the data is marked, and then the claim in each proposition is verified or refuted in the background investigation.

3.1 Marking the scope of each proposition

We currently mark the scope of verifiable propositions in the narrative that are likely to have supporting background ground truth information before we establish the ground truth. For example, statements made about a domestic disturbance that involved the police are likely to have a police report to supply background information, while "my mother walked me to school every day," while technically verifiable, will not.

A verifiable proposition, or claim, is any linguistic form that can be assigned a truth value. Propositions can be short; the transcribed answers below are all fragmented ground truth units:

{my neck%T}
{Correct%T}
{Yep%T}

Examples such as these are common in spoken dialogue. Although they do not correspond syntactically to a full proposition, they have propositional content.

Propositions can also be quite long. For example, in the 34 words of the sentence

Any LJM transaction that involved a cash disbursement that would have been within my signing authority either had to be signed by me or someone else higher in the hierarchical chain of the company.

there is only a single claim: I or someone above me had to sign LJM transactions that involved cash disbursements.

Some material is excluded from proposition tagging. Utterances that attest only to the frame of mind of the narrator, e.g. expressions such as *I think, it's my belief,* cannot be refuted or confirmed empirically. Similarly, a sentence like *Ms. Watkins said that rumor had it* contains an assertion (*rumor had it)* not made by the narrator and therefore has no value in testing a verbal deception hypothesis. For the same reason, direct quotes are excluded from verification.

3.2 Marking the Ground Truth

Once the scope of the propositions in a narrative is marked, the annotated narrative is checked against the background ground truth information, and each proposition that can be verified is marked as T or F. We represent this judgment as follows:

But as far as the relationship between {Jeff McMahon moving from the finance group into the industrial products group%T}, {there was no connection whatsoever%F} (Enron)

{*At that time Philip Morris owned the Clark Gum Company%T} and {we were trying to get into the candy business%T}* (Johnston)

3.2.1 The fact checker

It is critical that the person who marks the ground truth has no contact with the persons who are checking the narrative for markers of deception – to the extent that the latter task is done by hand.

We have employed a law student to fact check the claims in the one legal deposition (Johnston) we have in our current data set. We plan to employ an accounting student with a background in forensic accounting to fact check Lehmann Bros. quarterly earnings conference calls (see Larcker and Zakolyukina (2010) for similar data). For the other data, we have employed graduate assistants in linguistics who do not work on the deception markers.

3.2.2 Sources of background information

At a minimum, the background information used to mark the ground truth should include the source of the data used to establish the truth. That said, no data source is perfect. A confession may be coerced, an eyewitness may forget, a judgment may be faulty. However, at some point, we have to make a decision as to what a credible source is. We have assumed that the sources given in section 2.2.2 above, as well as claims made by the narrator that refute prior claims, all function as reliable sources of background information upon which to make decisions about the truth of a claim.

3.2.3 Verifying a claim

To verify a claim, we use both direct and circumstantial evidence. However, the latter is used only to direct us to a potentially false claim and must be supported by additional, direct facts.

Direct evidence requires no additional inferencing. In a narrative we have studied but not marked for ground truth, the police return to the apartment from which the suspect's wife has gone missing to find her body in the closet, at which point the suspect admits to suffocating his wife and describes the events leading up to the murder. His narrative prior to the confession described contrasting events that occurred in the same timeframe; this will enable us to mark these as False based on the direct evidence of the body and the confession.

Circumstantial evidence requires that a fact be inferred. For example, in his testimony before the U.S. Congress, Jeffrey Skilling claims that when he left Enron four months before the company collapsed, he thought "the company was in good shape." Circumstantial evidence of Skilling's reputation as an astute businessman and the well-known knowledge of his deep involvement with the company make this unlikely, as the interviewing congressman points out. However, we relied as well on direct testimony from other members of the Enron Board of Directors to affirm that Skilling knew the disastrous state of Enron when he left.

Verifying claims is a difficult, time consuming and sometimes tedious process. For the 35,090 words of narrative data currently in our collection, we have been able to verify 184 propositions, 110 as True and 74 as False. Appendix B gives the T/F counts for each of our narratives.

3.3 Enron: Examples of verification

Jeffrey Skilling was the Chief Operating Officer of the Enron Corporation as it was failing in 2001; he left the company in August 2001. In his testimony before the U.S. Congress the following year, which we used as our primary narrative data, Skilling made several important claims that were contradicted either by multiple parties involved in the case or by facts on record. This section illustrates how we apply the evidence to several of Skilling's claims.

1. The financial condition of Enron at the time of Skilling's departure.

MR. SKILLING: Congressman, I can just say it again – {on the date I left I absolutely, unequivocally thought the company was in good shape.F%}

Congressman Edward Markey provides circumstantial evidence that this claim is false, stating that Skilling's reputation, competence and hands-on knowledge makes this claim hard to believe. Direct evidence comes from Jeffrey McMahon, a former Enron treasurer, and Jordan Mintz, a senior attorney, who testified that they had told Skilling their concerns that limited

partnerships that the company was involved in created a conflict of interest for certain Enron board members, and were damaging Enron itself.

2. The presence of Mr. Skilling at a critical meeting to discuss these limited partnerships, which enabled Enron to hide its losses.

MR. SKILLING: Well, {there's an issue as to whether I was actually at a%F} -- the particular meeting that you're talking about was in Florida, Palm Beach, Florida. . . .

But when Greenwood brandished a copy of the meeting's minutes, which confirmed Skilling's presence, the former COO hedged his answer, saying,

MR. SKILLING: "I could have been there for a portion of the meeting. Was I there for the entire meeting? I don't know."

3. The issue of whether Skilling, as Enron's Chief Operating Officer, was required to approve Enron-LJM limited partnership transactions.

Mr. SKILLING: {I was not required to approve those transactions.%F}

Minutes of the Finance Committee of Enron's Board of Directors, October 6, 2000 (referenced in the congressional testimony) show that "Misters Buy, Causey, and Skilling approve all transactions between the company and LJM funds."

4 Conclusion and lessons learned

Research in high stakes deception has been held back by the difficulty of ground truth verification. Finding suitable data "in the wild" and conducting the fact checks to obtain ground truth is costly, time-consuming and labor intensive. This is not an unknown problem in computational linguistics. Other research efforts that rely on fact checking, such as Sauri and Pustejovsky (2009), face similar ground truth challenges.

We have described our work in building a corpus customized for high stakes deception studies in hopes of encouraging other researchers to build and share similar corpora. We envision the eventual goal as a multi-language resource with standardized methods and corpora available to the community at little or no cost.

We have made several mistakes that we hope we and others can avoid in collecting high stakes data. Some errors cost us time and others aggravating work trying to correct them.

Our first lesson was to establish a strict separation between the people who annotate the data for ground truth and those who mark it for deception – if any portion of the latter is being done manually. It is important that the fact checkers are not influenced by anything in the language of the narrator that might skew them toward marking a claim one way or the other.

With respect to the narrative data, it is important in selecting new data for annotating and ground truth checking to establish that the data is of the types approved by the research institution's compliance board; in the United States, this is the Institutional Review Board of the housing institution.

It is also important to have assurance that there is a robust body of background data with which to establish ground truth. While it is impressive to be able to find 13 of the 15 verifiably false statements in 240,000 words of narrative—a situation we experienced with a private data set—it does not give us the statistical robustness we would hope for.

We also found it important to save the data sources locally. Websites disappear and the possibility of further fact checking goes with them.

Finally, it is important to provide formal training for proposition tagging and ground truth tagging to ensure consistency and quality. Tutorials, user manuals and careful supervision should be available at all times.

Acknowledgments

We are thankful to the anonymous EACL reviewers for their incisive and helpful comments. Any errors or oversights are strictly the responsibility of the authors.

References

Joan Bachenko, Eileen Fitzpatrick and Michael Schonwetter. 2008. Verification and Implementation of Language-based Deception Indicators in Civil and Criminal Narratives. *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008).* University of Manchester, Manchester, UK.

Eileen Fitzpatrick and Joan Bachenko. 2010. Building a Forensic Corpus to Test Language-based Indicators of Deception. *Corpus Linguistics in*

North America 2008: Selections from the Seventh North American Symposium of the American Association for Corpus Linguistics. Gries, S., S. Wulff and M. Davies (eds.). *Series in Language and Computers.* Rodopi.

Tommaso Fornaciari and Massimo Poesio. 2011. Lexical vs. Surface Features in Deceptive Language Analysis, Workshop: Legal Applications of Human Language Technology. *13th International Conference on Artificial Intelligence and Law*. June 6-10. University of Pittsburgh.

Julia Hirschberg, Stefan Benus, Jason M. Brenier, Frank Enos, Sarah Friedman, Sarah Gilman, Cynthia Girand, Martin Graciarena, Andreas Kathol, Laura Michaelis, Bryan L. Pellom, Elizabeth Shriberg, Andreas Stolcke. 2005. "Distinguishing Deceptive from Non-Deceptive Speech," *INTERSPEECH 2005*, Lisbon, September.

David F. Larcker and Anastasia A. Zakolyukina. 2010. Detecting deceptive discussions in conference calls. Rock Center for Corporate Governance. Working Paper Series No. 83.

Roser Sauri and James Pustejovsky. 2009. FactBank 1.0. *Linguistic Data Consortium*, Philadelphia.

James B. Stiff, Steve Corman, Robert Krizek, and Eric Snider. 1994. Individual differences and changes in nonverbal behavior; Unmasking the changing faces of deception. *Communication Research, 21,* 555-581.

Aldert Vrij. 2008. Detecting Lies and Deceit: Pitfalls and Opportunities, 2nd. Edition. Wiley-Interscience.

Code of Federal Regulations. Retrieved Jan. 26, 2012 http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.html#46.102

Appendix A. Sources of Background Data that has been verified[6]

Case	Source	
Johnston	Documents available from the *State of Minnesota and Blue Cross and Blue Shield of Minnesota v Philip Morris Inc et al* during the discovery process of the trial.	
Routier	Police report from first responder, Sgt. Matthew Walling. No longer available online	
Enron[7]	Kenneth L. Lay and Jeffrey K. Skilling Jury Trial – Govt. Exhibits[8] Enron Special Investigations Report (The Powers Report) Employee letters and emails	
Kennedy	Police report from Edgartown MA, and transcript of the inquest	
Peterson	Modesto Police Dept. website Gomez Peterson interview Sawyer Peterson interview Findlaw.com International call code database	Mobile number lookup Mapquest U.S. Time Zones Livermore Chevron Station

Appendix B. Distribution of T and F Propositions in Collection

Case	Words	Trues	Falses
Johnston	12,762	34	48
Routier	1,026	8	2
Enron	7,476	23	21
Kennedy	245	8	2
Peterson	13,581	37	1
TOTAL	35,090	110	74

[6] We included data from two cases of theft in the original set, which was collected prior to the creation of an IRB at our university. Incomplete documentation requires us to exclude these cases. Another case, which we called 'Guilty Nurse,' was not sufficiently sourced to be included.
[7] http://news.findlaw.com/legalnews/lit/enron/#documents
[8] http://www.justice.gov/enron/

Appendix C. Attributes of the Data Set

S=spoken; W=written

Case	Case Type	Mode	Narrator
Johnston	Civil; sale of tobacco to teens	S	Male 60+; retired tobacco CEO
Routier	Criminal; murder	W	Female 26; homemaker
Enron (Skilling)	Criminal; fraud	S	Male 53; former Enron COO
Kennedy	Criminal; leaving the scene of an accident	W	Male 37; former US Senator, deceased
Peterson	Criminal; murder	S	Male 30; agriculture chemical salesman