Automatic Detection of Verbal Deception

Eileen Fitzpatrick, Joan Bachenko, Tommaso Fornaciari
(Montclair State University, Linguistech LLC, Italian National Police)

Morgan & Claypool (Synthesis Lectures on Human Language Technologies, edited by Graeme Hirst, volume 29), 2015, xvii+101 pp; paperback, ISBN 978-1-62705-337-2; ebook, ISBN 978-1-62705-338-9; doi:10.2200/S00656ED1V01Y201507HLT029

Reviewed by
Yoong Keok Lee
IBM T. J. Watson Center

Detecting deception is an ancient problem that continues to find relevance today in many areas, such as border control, financial auditing, testimony assessment, and Internet fraud. Over the last century, research in this area has focused mainly on discovering physiological signals and psychological behaviors that are associated with lying. Using verbal cues (i.e., words and language structure) is not entirely new. But in recent years, data-driven and machine learning frameworks, which are now ubiquitous in the natural language processing (NLP) community, has brought new light to this old field. This highly accessible book puts last decade's research in verbal deception in the context of traditional methods. It is a valuable resource to anyone with a basic understanding of machine learning looking to make inroads or break new ground in the subspecialty of detecting verbal deception.

The book consists of five chapters organized into three parts—background on nonverbal cues, statistical NLP approaches, and future directions. The introductory chapter concisely defines the problem and relates verbal cues to the behavioral ones. It also provides an intuition of why patterns in language would be effective and provides an estimated state-of-the-art performance.

Chapter 2 provides background on a behavioral approach to identifying deception. The first section gets readers acquainted with terms used by nonverbal cues to deception. These include physiological signals (which are the basis of well-known lie detection methods such as polygraphy and thermography), vocal cues (such as speech disfluencies), and body and facial expressions (e.g., pupil dilation). Although seemingly detached from the focus of the book, this preliminary material is an interesting introduction that also serves as a terminology reference later. The remaining chapter covers two topics: psychology of deception and applied criminal justice. The part on psychology presents a literature review of two definitive meta-analysis of the literature in the twentieth century. It first gives a theoretical account of deceptive behavior, such as motivation to lie and emotional states of liars. Next, it reports the experimental effectiveness of measurable cues, whether objectively or subjectively, such as complexity and amount of information. The second meta-analysis examines conditions that tend to make lying behavior more obvious, for example, interrogation. Although seemingly unrelated to NLP, I expect these reviews to be a source of inspiration for novel feature and model engineering. Because the material is very comprehensive and possibly foreign to the NLP community, I would like to see this part organized by the type of behavior cues (in the same vein as the preliminary material on physiological

doi:10.1162/COLI_r_00282

signals), instead of being presented as a summary of two individual papers. A tabular summary of the effectiveness of each cue would also be very useful. The next major topic in this chapter describes how theory is put to practice in criminal justice. This part describes three real-world methodologies used by law enforcement agencies to vet court witnesses. The relevance to NLP is more apparent here because verbal cues, whether spoken or written, are used to score legal statements, although rather subjectively. The chapter ends with an evaluation of a commercial product for assessing the veracity of legal statements. The concluding topic of evaluation naturally brings us to the question of obtaining ground truth data, which is the focus of the next chapter.

Chapter 3 is dedicated to data collection in spoken and written media. It begins by discussing, on one hand, imperfections of the laboratory setting for producing realistic data, and, on the other hand, challenges of establishing ground truth in realistic environments. The rest of the chapter describes how linguistic corpora are acquired in three domains—legal, financial, and mass media. There are numerous examples of the sources of raw data and how ground truth is determined. I am particularly interested in the use of online resources to construct data sets, although the discussion of the mass media domain is unexpectedly short in comparison to the other two domains. Exploiting crowdsourcing efforts also comes to mind immediately, but it is omitted from this chapter. I later realized that the next chapter actually reports a data set that is constructed with the use of Amazon Mechanical Turk, but I think a brief mention here would be very appropriate.

Chapter 4 is the core of this book—NLP systems that distinguish false narratives from true ones. It is organized into three sections and a conclusion. The first section presents the text classification set-up that is now standard in many NLP tasks—supervised learning models, training/test data splits, classification performance metrics, baseline measures, and standard NLP preprocessing procedures. It also discusses the difficulty of establishing baselines and upper bound performance measures. Nevertheless, readers can find solutions that are adopted in practice. The next section describes variations of the task which I find interesting, because they are rather unique to this area. For example, instead of viewing deception identification as a standard document classification problem, we can also aim to identify fraudulent authors, incorrect claims, or even recover omissions of critical information. The remainder of this section repeats issues related to feature variations and obtaining ground truth that are detailed in other sections, so I think they can be omitted. The third section presents systems that have been built in the last decade. It begins with an investigation of models based on the standard n-gram representation. The first system detects scam tweets using the suffix-tree algorithm. I think the objective that it aims to optimize deserves a brief mention. The other remaining n-gram-based systems use data artificially generated from Amazon Mechanical Turk. There is also an informative comparison of different variants of the basic set-up, for example, cross-domain evaluations and the interaction of learning algorithms and various sizes of n-grams. There is also an interesting system that performs one-class learning (i.e., learning only from positive [deceptive] and unlabeled data). The rest of the section details a wide variety of systems that go beyond n-grams. They utilize elaborate lexical features, word categories, and syntactic features obtained from parse trees. A particularly useful resource is the Linguistic Inquiry and Word Count lexicon, which is constructed by psychologists to characterize cognitive change. Overall, the section contains a comprehensive and fairly well organized coverage of models. What I hope to see is a summary of data sets and a systematic comparison of models to consolidate research results in this area.

Chapter 5 discusses the limitations of current research and offers a number of directions for future work. For instance, the NLP community has mostly been working with online reviews. Here you will find opportunities related to expanding the scope of data sets. Also, inspired by recent work that distinguishes imaginative from informative writing, this chapter also explores the possibility of using linguistic cues to capture emotional states that are well-documented in the psychology literature. More broadly, it discusses gaps in our understanding of how verbal and non-verbal cues interact. In summary, this chapter presents a series of thought-provoking questions that are much needed to advance the field.

All in all, this book is a timely reference for an emerging subspecialty at the intersection of deception research and NLP. It synthesizes a survey of recent NLP systems for identifying deceptive verbal data and background material from mainstream behavioral research. I see it as a convenient resource for researchers looking to advance the field in the following ways: improving the state-of-the-art performance on existing benchmark data sets, constructing new evaluation resources and metrics, and investigating the interaction between verbal and non-verbal cues.

Yoong Keok Lee is a research staff member at the IBM T. J. Watson Research Center, Yorktown Heights, New York. His research interests are in machine learning approaches to natural language processing, particularly tagging, morphology, information extraction, and word sense disambiguation. Lee's e-mail address is yklee@us.ibm.com.