Conroy, Rubin, and Chen

Summary:

Humans are only 4% better than chance at determining lies in text. There are two major categories of methods, Linguistic Approaches, which focus on content of texts and analysis to determine patterns, and the Network Approaches, where message metadata, or alternatively a knowledge network such as a database provide aggregate deception measures. They incorporate machine learning techniques to train classifiers. So far there is no known typology of methods in current literature.

The linguistic approach makes use of “predictive deception cues” in the content of the text. The simplest method of representing text is as a bag of words approach, where every word is significant. Individual words and n-gram frequencies can be recorded and analyzed to find cues of deception. Part of speech tagging, or shallow syntax, provide frequency sets to reveal cues. The drawback is the lack of context that comes from isolating n-grams, but this method can also be useful in tandem with others.

One may also analyze deeper syntax, using Probability Context Free Grammars, which are sentences that have been converted to parse trees, with probabilities assigned. These can be used with 85-91% accuracy to detect deception. Once again, alone, these methods may not be sufficient to identify deception, and studies have combined the approach with other techniques.

An extension of this method can be found in semantic analysis, wherein attribute:descriptor pairs are compared to existing analogous data, known as a content profile. For example, if the author claims that an art museum is near a hotel, this can potentially be compared to geographic data, or product/business reviews which may or may not present evidence to the contrary. This method has been shown to be 91% accurate; however, associating attributes and descriptors can be challenging without sufficient data, and finding content profile with confirmed correct information is itself a daunting task.

Rhetorical Structure Theory provides an analytic framework for identifying rhetorical relations between linguistic elements. Systematic differences between true and false messages can be combined with a Vector Space Model that calculates a message vector’s distance from a truth center and a deceptive center. This model makes use of prominent rhetorical relations found in deceptive texts to be indicative of deception.

Support Vector Machines and Naïve Bayesian models are examples of classifiers that utilize sets of words and category frequencies to train them. They predict deception by using numeric clustering and distances. Naïve Bayes algorithms se accumulated evidence of correlations between one variable and the others in the model. Classification results from unintended emotional cues and syntactic evidence resulting from exaggeration of the sentiment required of deception.

Network approaches can complement linguistic approaches, especially with the advent of sites such as twitter. Knowledge networks, such as DBpedia or the Google Relation Extraction Corpus, are used to analyze false “factual statements” by calculating a simple shortest path, to find semantic proximity. The shorter the path, the more likely the claim is factual. One challenge with this method is the reliance on a pre-existing knowledge base.

Another network based method involves the use of metadata and “telltale behavior” of questionable sources, such as the inclusion of hyperlinks or associated metadata, which can be compiled. Centering resonance analysis can be used to link the most important words that link other words in the network.

Combining network with linguistic methods adds the “trust” element by identifying source behavior beyond simply content. Additionally, data is much needed in this field, and is a worthwhile contribution on its own.