Syntactic Stylometry for Deception Detection

Song Feng **Ritwik Banerjee** **Yejin Choi**
Department of Computer Science
Stony Brook University
Stony Brook, NY 11794-4400
songfeng, rbanerjee, ychoi@cs.stonybrook.edu

Abstract

Most previous studies in computerized deception detection have relied only on shallow lexico-syntactic patterns. This paper investigates syntactic stylometry for deception detection, adding a somewhat unconventional angle to prior literature. Over four different datasets spanning from the product review to the essay domain, we demonstrate that features driven from Context Free Grammar (CFG) parse trees consistently improve the detection performance over several baselines that are based only on shallow lexico-syntactic features. Our results improve the best published result on the hotel review data (Ott et al., 2011) reaching 91.2% accuracy with 14% error reduction.

1 Introduction

Previous studies in computerized deception detection have relied only on shallow lexico-syntactic cues. Most are based on dictionary-based word counting using LIWC (Pennebaker et al., 2007) (e.g., Hancock et al. (2007), Vrij et al. (2007)), while some recent ones explored the use of machine learning techniques using simple lexico-syntactic patterns, such as n-grams and part-of-speech (POS) tags (Mihalcea and Strapparava (2009), Ott et al. (2011)). These previous studies unveil interesting correlations between certain lexical items or categories with deception that may not be readily apparent to human judges. For instance, the work of Ott et al. (2011) in the hotel review domain results in very insightful observations that deceptive reviewers tend to use verbs and personal pronouns (e.g., "I", "my") more often, while truthful reviewers tend to use more of nouns, adjectives, prepositions. In parallel to these shallow lexical patterns, might there be deep syntactic structures that are lurking in deceptive writing?

This paper investigates syntactic stylometry for deception detection, adding a somewhat unconventional angle to prior literature. Over four different datasets spanning from the product review domain to the essay domain, we find that features driven from Context Free Grammar (CFG) parse trees consistently improve the detection performance over several baselines that are based only on shallow lexico-syntactic features. Our results improve the best published result on the hotel review data of Ott et al. (2011) reaching 91.2% accuracy with 14% error reduction. We also achieve substantial improvement over the essay data of Mihalcea and Strapparava (2009), obtaining upto 85.0% accuracy.

2 Four Datasets

To explore different types of deceptive writing, we consider the following four datasets spanning from the product review to the essay domain:

I. TripAdvisor—Gold: Introduced in Ott et al. (2011), this dataset contains 400 truthful reviews obtained from `www.tripadviser.com` and 400 deceptive reviews gathered using Amazon Mechanical Turk, evenly distributed across 20 Chicago hotels.

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 171–175,
Jeju, Republic of Korea, 8-14 July 2012.

TripAdvisor–Gold		TripAdvisor–Heuristic	
Deceptive	Truthful	Deceptive	Truthful
NP^PP → DT NNP NNP NNP	S^ROOT → VP .	NP^S → PRP	VP^S → VBZ NP
SBAR^NP → S	NP^NP → $ CD	SBAR^S → WHADVP S	NP^NP → NNS
NP^VP → NP SBAR	PRN^NP → LRB NP RRB	VP^S → VBD PP	WHNP^SBAR → WDT
NP^NP → PRP$ NN	NP^NP → NNS	S^SBAR → NP VP	NP^NP → NP PP PP
NP^S → DT NNP NNP NNP	NP^S → NN	S^ROOT → PP NP VP .	NP^S → EX
VP^S → VBG PP	NP^PP → DT NNP	VP^S → VBD S	NX^NX → JJ NN
NP^PP → PRP$ NN	NP^PP → CD NNS	NP^S → NP CC NP	NP^NP → NP PP
VP^S → MD ADVP VP	NP^NP → NP PRN	NP^S → PRP$ NN	VP^S → VBZ RB NP
VP^S → TO VP	PRN^NP → LRB PP RRB	NP^PP → DT NNP	PP^NP → IN NP
ADJP^NP → RBS JJ	NP^NP → CD NNS	NP^PP → PRP$ NN	PP^ADJP → TO NP

Table 1: Most discriminative rewrite rules ($\hat{r}$): hotel review datasets

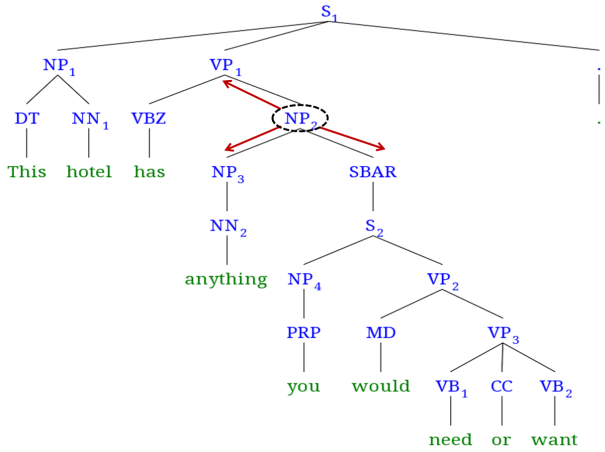

Figure 1: Parsed trees

II. TripAdvisor—Heuristic: This dataset contains 400 truthful and 400 deceptive reviews harvested from `www.tripadviser.com`, based on fake review detection heuristics introduced in Feng et al. (2012).[1]

III. Yelp: This dataset is our own creation using `www.yelp.com`. We collect 400 *filtered* reviews and 400 *displayed* reviews for 35 Italian restaurants with average ratings in the range of [3.5, 4.0]. Class labels are based on the meta data, which tells us whether each review is *filtered* by Yelp's automated review filtering system or not. We expect that *filtered* reviews roughly correspond to deceptive reviews, and *displayed* reviews to truthful ones, but not without considerable noise. We only collect 5-star reviews to avoid unwanted noise from varying degree of sentiment.

IV. Essays: Introduced in Mihalcea and Strapparava (2009), this corpus contains truthful and deceptive essays collected using Amazon Mechanic Turk for the following three topics: "Abortion" (100 essays per class), "Best Friend" (98 essays per class), and "Death Penalty" (98 essays per class).

3 Feature Encoding

Words Previous work has shown that bag-of-words are effective in detecting domain-specific deception (Ott et al., 2011; Mihalcea and Strapparava, 2009). We consider unigram, bigram, and the union of the two as features.

Shallow Syntax As has been used in many previous studies in stylometry (e.g., Argamon-Engelson et al. (1998), Zhao and Zobel (2007)), we utilize part-of-speech (POS) tags to encode shallow syntactic information. Note that Ott et al. (2011) found that even though POS tags are effective in detecting fake product reviews, they are not as effective as words. Therefore, we strengthen POS features with unigram features.

Deep syntax We experiment with four different encodings of production rules based on the Probabilistic Context Free Grammar (PCFG) parse trees as follows:

- r: unlexicalized production rules (i.e., all production rules except for those with terminal nodes), e.g., $NP_2 \rightarrow NP_3$ SBAR.
- $r*$: lexicalized production rules (i.e., *all* production rules), e.g., PRP → "you".
- $\hat{r}$: unlexicalized production rules combined with the grandparent node, e.g., NP_2^VP

[1] Specifically, using the notation of Feng et al. (2012), we use data created by Strategy-$dist\Phi$ heuristic, with $H_S, \mathcal{S}$ as deceptive and $H'_S, \mathcal{T}$ as truthful.

		TripAdvisor		Yelp	Essay		
		Gold	Heur		Abort	BstFr	Death
words	unigram	*88.4*	**74.4**	59.9	*70.0*	*77.0*	**67.4**
	bigram	85.8	71.5	**60.7**	71.5	79.5	55.5
	uni + bigram	**89.6**	73.8	60.1	**72.0**	**81.5**	65.5
shallow syntax +words	pos(n=1) + unigram	87.4	74.0	62.0	70.0	80.0	66.5
	pos(n=2) + unigram	88.6	74.6	59.0	67.0	82.0	66.5
	pos(n=3) + unigram	88.6	74.6	59.3	67.0	82.0	66.5
deep syntax	r	78.5	65.3	56.9	62	67.5	55.5
	$\hat{r}$	74.8	65.3	56.5	58.5	65.5	56.0
	$r*$	89.4	74.0	64.0	70.1	77.5	66.0
	$\hat{r}*$	90.4	75	63.5	71.0	78	67.5
deep syntax +words	r + unigram	89.0	74.3	62.3	76.5	82.0	69.0
	$\hat{r}$ + unigram	88.5	74.3	62.5	**77.0**	81.5	70.5
	$r*$ + unigram	90.3	75.4	**64.3**	74.0	**85.0**	**71.5**
	$\hat{r}*$ + unigram	**91.2**	**76.6**	62.1	76.0	84.5	71.0

Table 2: Deception Detection Accuracy (%).

$_1 \rightarrow \text{NP}_3$ SBAR.

- $\hat{r}*$: lexicalized production rules (i.e., *all* production rules) combined with the grandparent node, e.g., PRP$\hat{}$NP $_4 \rightarrow$ "you".

4 Experimental Results

For all classification tasks, we use SVM classifier, 80% of data for training and 20% for testing, with 5-fold cross validation.[2] All features are encoded as tf-idf values. We use Berkeley PCFG parser (Petrov and Klein, 2007) to parse sentences. Table 2 presents the classification performance using various features across four different datasets introduced earlier.[3]

4.1 TripAdvisor–Gold

We first discuss the results for the TripAdvisor–Gold dataset shown in Table 2. As reported in Ott et al. (2011), bag-of-words features achieve surprisingly high performance, reaching upto 89.6% accuracy. Deep syntactic features, encoded as $\hat{r}*$ slightly improves this performance, achieving 90.4% accuracy. When these syntactic features are combined with unigram features, we attain the best performance of 91.2% accuracy, yielding 14% error reduction over the word-only features.

Given the power of word-based features, one might wonder, whether the PCFG driven features are being useful only due to their lexical production rules. To address such doubts, we include experiments with unlexicalized rules, r and $\hat{r}$. These features achieve 78.5% and 74.8% accuracy respectively, which are significantly higher than that of a random baseline (∼50.0%), confirming statistical differences in deep syntactic structures. See Section 4.4 for concrete exemplary rules.

Another question one might have is whether the performance gain of PCFG features are mostly from local sequences of POS tags, indirectly encoded in the production rules. Comparing the performance of [shallow syntax+words] and [deep syntax+words] in Table 2, we find statistical evidence that deep syntax based features offer information that are not available in simple POS sequences.

4.2 TripAdvisor–Heuristic & Yelp

The performance is generally lower than that of the previous dataset, due to the noisy nature of these datasets. Nevertheless, we find similar trends as those seen in the TripAdvisor–Gold dataset, with respect to the relative performance differences across different approaches. The sig-

[2]We use LIBLINEAR (Fan et al., 2008) with L2-regulization, parameter optimized over the 80% training data (3 folds for training, 1 fold for testing).

[3]Numbers in *italic* are classification results reported in Ott et al. (2011) and Mihalcea and Strapparava (2009).

TripAdvisor–Gold		TripAdvisor–Heur	
Decep	Truth	Decep	Truth
VP	PRN	VP	PRN
SBAR	QP	WHADVP	NX
WHADVP	S	SBAR	WHNP
ADVP	PRT	WHADJP	ADJP
CONJP	UCP	INTJ	WHPP

Table 3: Most discriminative phrasal tags in PCFG parse trees: TripAdvisor data.

nificance of these results comes from the fact that these two datasets consists of real (fake) reviews in the wild, rather than manufactured ones that might invite unwanted signals that can unexpectedly help with classification accuracy. In sum, these results indicate the existence of the statistical signals hidden in deep syntax even in real product reviews with noisy gold standards.

4.3 Essay

Finally in Table 2, the last dataset Essay confirms the similar trends again, that the deep syntactic features consistently improve the performance over several baselines based only on shallow lexico-syntactic features. The final results, reaching accuracy as high as 85%, substantially outperform what has been previously reported in Mihalcea and Strapparava (2009). How robust are the syntactic cues in the cross topic setting? Table 4 compares the results of Mihalcea and Strapparava (2009) and ours, demonstrating that syntactic features achieve substantially and surprisingly more robust results.

4.4 Discriminative Production Rules

To give more concrete insights, we provide 10 most discriminative unlexicalized production rules (augmented with the grand parent node) for each class in Table 1. We order the rules based on the feature weights assigned by LIBLINEAR classifier. Notice that the two production rules in bolds — [SBAR^NP → S] and [NP ^VP → NP SBAR] — are parts of the parse tree shown in Figure 1, whose sentence is taken from an actual fake review. Table 3 shows the most discriminative phrasal tags in the PCFG parse trees for each class. Interestingly, we find more frequent use of VP, SBAR (clause introduced by subordinating conjunction), and WHADVP in deceptive reviews than truthful reviews.

training:	**A** & **B**	**A** & **D**	**B** & **D**
testing:	**D**eathPen	**B**estFrn	**A**bortion
M&S 2009	58.7	58.7	62.0
$r*$	66.8	70.9	69.0

Table 4: Cross topic deception detection accuracy: Essay data

5 Related Work

Much of the previous work for detecting deceptive product reviews focused on related, but slightly different problems, e.g., detecting duplicate reviews or review spams (e.g., Jindal and Liu (2008), Lim et al. (2010), Mukherjee et al. (2011), Jindal et al. (2010)) due to notable difficulty in obtaining gold standard labels.[4] The Yelp data we explored in this work shares a similar spirit in that gold standard labels are harvested from existing meta data, which are not guaranteed to align well with true hidden labels as to deceptive v.s. truthful reviews. Two previous work obtained more precise gold standard labels by hiring Amazon turkers to write deceptive articles (e.g., Mihalcea and Strapparava (2009), Ott et al. (2011)), both of which have been examined in this study with respect to their syntactic characteristics. Although we are not aware of any prior work that dealt with syntactic cues in deceptive writing directly, prior work on hedge detection (e.g., Greene and Resnik (2009), Li et al. (2010)) relates to our findings.

6 Conclusion

We investigated syntactic stylometry for deception detection, adding a somewhat unconventional angle to previous studies. Experimental results consistently find statistical evidence of deep syntactic patterns that are helpful in discriminating deceptive writing.

[4]It is not possible for a human judge to tell with full confidence whether a given review is a fake or not.

References

S. Argamon-Engelson, M. Koppel, and G. Avneri. 1998. Style-based text categorization: What newspaper am i reading. In *Proc. of the AAAI Workshop on Text Categorization*, pages 1–4.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

S. Feng, L. Xing, Gogar A., and Y. Choi. 2012. Distributional footprints of deceptive product reviews. In *Proceedings of the 2012 International AAAI Conference on WebBlogs and Social Media*, June.

S. Greene and P. Resnik. 2009. More than words: Syntactic packaging and implicit sentiment. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 503–511. Association for Computational Linguistics.

J.T. Hancock, L.E. Curry, S. Goorha, and M. Woodworth. 2007. On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes*, 45(1):1–23.

Nitin Jindal and Bing Liu. 2008. Opinion spam and analysis. In *Proceedings of the international conference on Web search and web data mining*, WSDM '08, pages 219–230, New York, NY, USA. ACM.

Nitin Jindal, Bing Liu, and Ee-Peng Lim. 2010. Finding unusual review patterns using unexpected rules. In *Proceedings of the 19th ACM Conference on Information and Knowledge Management*, pages 1549–1552.

X. Li, J. Shen, X. Gao, and X. Wang. 2010. Exploiting rich features for detecting hedges and their scope. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning—Shared Task*, pages 78–83. Association for Computational Linguistics.

Ee-Peng Lim, Viet-An Nguyen, Nitin Jindal, Bing Liu, and Hady Wirawan Lauw. 2010. Detecting product review spammers using rating behaviors. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 939–948, New York, NY, USA. ACM.

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

Arjun Mukherjee, Bing Liu, Junhui Wang, Natalie S. Glance, and Nitin Jindal. 2011. Detecting group review spam. In *Proceedings of the 20th International Conference on World Wide Web (Companion Volume)*, pages 93–94.

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*, pages 309–319, Portland, Oregon, USA, June. Association for Computational Linguistics.

J.W. Pennebaker, C.K. Chung, M. Ireland, A. Gonzales, and R.J. Booth. 2007. The development and psychometric properties of liwc2007. *Austin, TX, LIWC. Net.*

S. Petrov and D. Klein. 2007. Improved inference for unlexicalized parsing. In *Proceedings of NAACL HLT 2007*, pages 404–411.

A. Vrij, S. Mann, S. Kristen, and R.P. Fisher. 2007. Cues to deception and ability to detect lies as a function of police interview styles. *Law and human behavior*, 31(5):499–518.

Ying Zhao and Justin Zobel. 2007. Searching with style: authorship attribution in classic literature. In *Proceedings of the thirtieth Australasian conference on Computer science - Volume 62*, ACSC '07, pages 59–68, Darlinghurst, Australia, Australia. Australian Computer Society, Inc.