Psycholinguistic Features for Deceptive Role Detection in Werewolf *

Codruta Girlea
University of Illinois
Urbana, IL 61801, USA
girlea2@illinois.edu

Roxana Girju
University of Illinois
Urbana, IL 61801, USA
girju@illinois.edu

Eyal Amir
University of Illinois
Urbana, IL 61801, USA
eyal@illinois.edu

Abstract

We tackle the problem of identifying deceptive agents in highly-motivated high-conflict dialogues. We consider the case where we only have textual information. We show the usefulness of psycho-linguistic deception and persuasion features on a small dataset for the game of Werewolf. We analyse the role of syntax and we identify some characteristics of players in deceptive roles.

1 Introduction

Deception detection has gained some attention in the NLP community (Mihalcea and Strapparava, 2009; Ott et al., 2011; Jindal and Liu, 2008). The focus has mostly been on detecting insincere reviews or arguments. However, there has been little work (Hung and Chittaranjan, 2010) in detecting deception and manipulation in dialogues.

When the agents involved in a dialogue have conflicting goals, they are often motivated to use deception and manipulation in order to reach those goals. Examples include trials and negotiations. The high motivation for using deception and the possiblity of tracking the effects on participants throughout the dialogue sets this problem apart from identifying deception in nonce text fragment where motivation is not immediately relevant.

The Werewolf game is an instance of such a dialogue where people are motivated to deceive and manipulate in order to reach their goals. The setting is a village where at least one of the villagers is secretly a werewolf. Each night, a villager falls prey to the werewolves. Each day, the remaining villagers discuss to find the most likely werewolf to execute.

Players are assigned roles that define their goals and available actions. For our purpose, all roles are collapsed together as either werewolf or non-werewolf. There are other important roles in Werewolf, such as seer, vigilante, etc, the goals and available actions of which are not the focus of this paper. (Barnwell, 2012) provides a broader description of the game and roles.

Each player only knows her own role, as assigned by an impartial judge who overlooks the game. The players with a *werewolf* role learn each other's roles in the first round of the game. Every round, they pick another non-werewolf player to be removed from the game. This happens during a *night phase*, and is hidden from the other players. The judge announces the identity and role of the removed player.

All players are then allowed to remove one other player from the game before the next night phase. They discuss and vote during a *day phase*. The *non-werewolves* are motivated to remove the werewolves from the game. The werewolves are motivated to hide their roles, as in every round there is a majority of non-werewolves. Any time werewolves become a majority, they win the game. Any time all werewolves are eliminated, they lose the game.

In this paper we define the task as binary classification of deceptive and non-deceptive roles in Werewolf. Werewolf roles are deceptive, as they rely on deception to win the game, whereas all the

*This research was partially funded owing to a collaboration between Adobe Research and the University of Illinois. We thank Julia Hockenmaier, Dan Roth, Dafna Shahaf, Walter Chang, Trung Bui, and our reviewers for their helpful comments and feedback.

other roles are nondeceptive (Hung and Chittaranjan, 2010). For our purpose, all these roles are collapsed together according to whether they help the werewolves or not. We consider all the utterances of each player in each game as one distinct instance.

This is a first step towards building a model of deception in the Werewolf game, and more generally in scenarios where deception can be used to achieve goals. The model can then be used to predict future actions, e.g. the vote outcomes in Werewolf.

We show that by analyzing this dialogue genre we can gain some insights into the dynamics of manipulation and deception. These insights would then be useful in detecting hidden intentions and predicting decisions in important, real-life scenarios.

2 Previous Work

There has been little work on deception detection in written language and most of it has focused on either discriminating between sincere and insincere arguments (Mihalcea and Strapparava, 2009) or opinion spam (Ott et al., 2011; Jindal and Liu, 2008). One method of data collection has been to ask subjects to argue for both sides of a debate (Mihalcea and Strapparava, 2009). While lies about one's beliefs are also present, the manipulative behaviour and the motivation are missing. Since none of the previous work focuses on dialogue, the change in participants' beliefs, intentions, and plans reflected in the interaction between players is also absent.

More related is the work of (Hung and Chittaranjan, 2010), who recorded a total of 81.17 hours of people playing the Werewolf game, and used phonetic features to detect werewolves. While their results are promising, our focus is on written text only.

We have also been inspired by psycho-linguistic studies of deception detection (Porter and Yuille, 1996) as well as by psycho-linguistic research on persuasive or powerless language (Greenwald et al., 1968; Hosman, 2002; Sparks and Areni, 2008). We build upon findings from both of these lines of research, as Werewolf players use both deception, to hide their roles and intentions, and persuasion, to manipulate other players' beliefs and intentions.

3 Experiments

3.1 Data

The raw data consists of 86 game transcripts collected by Barnwell (Barnwell, 2012). The transcripts have an average length of 205 messages per transcript, including judge comments.

In the transcripts, the judge is a bot, which means there is a small fixed set of phrases it uses to make announcements. As part of the system, the judge knows all the roles as they are assigned. It reveals those roles in an annoucement as follows: every time a player is removed from the game, their role is made known; and the roles of the remaining players are revealed after the game is concluded.

We automatically extracted role assignments by looking for phrases such as *was a, is a, turns out to have been, carried wolfsbane*. We manually checked the assignments and found the werewolf roles were correctly assigned for 72 out of the 86 transcripts. The remaining 14 games end before they should because the judge bot breaks down. However, the players do reveal their own roles after the game ends. By looking at their comments, we manually annotated these remaining games.

All the utterances from each player in each transcript translate to one data instance. The label is 1 or 0, for whether the player is or isn't a werewolf. We do not consider the judge as part of the data. The resulting data set consists of 701 instances, of which 116 are instances of a werewolf role.

Given the small size and the skewed distribution of the dataset, we balanced the data with resampling so that we have enough instances to learn from.

3.2 Features

3.2.1 Psycholinguistic Features

(Tausczik and Pennebaker, 2010) suggest word count and use of negative emotions, motion, and sense words are indicative of deception.

We counted the negative emotion words using the MPQA subjectivity lexicon of (Wilson et al., 2005). We also experimented with the NRC word-emotion association lexicon of (Mohammad and Yang, 2011), but found the MPQA lexicon to perform better. Since we didn't have access to LIWC (Tausczik and Pennebaker, 2010), we used manually created lists of motion (*arrive, run, walk*) and sense

(*see, sense, appearance*) words. The lists are up to 50 words long. We also considered the number of verbs, based on our intuition that heavy use of verbs can be associated to motion. However, we don't expect the number of motion words to be as important in our domain. This is because deception in the Werewolf game does not refer to a fabricated story that other players have to be convinced to believe, but rather to hiding one's identity and intentions.

(Tausczik and Pennebaker, 2010) also talk about honesty features : number of exclusion words (*but, without, exclude, except, only, just, either*) and number of self references (we used a list of first person singular pronoun forms). They claim that cognitive complexity is also correlated with honesty. This is because maintaining the coherence of a fabricated story is cognitively taxing. Cognitive complexity manifests in the use of: long words (longer than 6 letters), exclusion words (differentiating between competing solutions), conjunctions, disjunctions, connectives (integrating different aspects of a task), and cognitive words (*think, plan, reason*).

(Porter and Yuille, 1996) observe that for highly motivated deception, people use longer utterances, more self references, and more negative statements. We used those as features, as the average number of words per utterance and the number of dependencies of type *negation* from the Stanford parser.

Another set of features cited by (Porter and Yuille, 1996) comes for ex-polygrapher Sapir's training program for police investigators. He notes that liars use too many unneeded connectors, and display deviations in pronoun usage – most of the times by avoiding first-person singular. This seems to contradict the discussion on highly motivated deception (Porter and Yuille, 1996), and is aligned with (Tausczik and Pennebaker, 2010)'s findings. It is possible that the natural tendency of a liar is to avoid self references (e.g. due to cognitive dissonance), but that a strong motivation can cause one to purposefully act against this tendency, ignoring any mental discomfort it may cause. In our experiments, we didn't observe any tendency of werewolf to either avoid or increase use of self references.

There are differences in language when used to recount a true memory versus a false one (reality monitoring) (Porter and Yuille, 1996). In this context, a true memory means a memory of reality, i.e. of a story that actually happened, whereas a false memory is a mental representation of a fabricated story. People talking about a true memory tend to focus on the attributes of the stimulus that generated the memory (e.g. shape, location, color), whereas people talking about a false memory tend to use more cognitive words (e.g. *believe*, *think*, *recall*) and hedges (e.g. *kind of*, *maybe*, *a little*). An explanation is that the process of fabricating a story engages reasoning more than it does memory, and people tend to resist committing to a lie. We used the noun and adjective count as a rough approximation of the number of stimulus attributes, as adjectives and nouns in prepositional phrases can be used to enrich a description, e.g. of a memory. However, it is important to note that Werewolf players do not actively lie, in the sense that the discussion does not involve events not directly accessible to all players. Therefore it's impossible for players to lie about the course of events, so there is no false memory to recount.

Another characteristic of the game is that the werewolves actively try to *persuade* other players that their intentions are not harmful. (Hosman, 2002) notes that language complexity is indicative of persuasive power. A measure of language complexity is the type-token ratio (TTR). On the other hand, hesitations (*um, er, uh*), hedges (*sort of, kind of, almost*), and polite forms are markers of powerless language (Sparks and Areni, 2008). We did not find any polite forms in our data, the context being a game where players adopt a familiar tone.

The complete list of features is as follows (words are stemmed): *TTR* (type-token ratio), *number of hesitations, number of negative emotions, number of words, number of words longer than 6 letters, number of self references, number of negations, number of hedges* (50 hedge words), *number of cognitive words* (50 words), *number of motion words* (20 words), *number of sense words* (17 words), *number of exclusion words, number of connectors* (prepositions and conjunctions), *number of pronouns, number of adjectives, number of nouns, number of verbs, number of conjunctions, number of prepositions*.

3.2.2 POS and Syntactic Features

Following the intuition that cognitive complexity can also be reflected in sentence structure, we de-

cided to look beyond lexical level for markers of deception and persuasion and experimented with POS and syntactic features. We used the Stanford POS parser (Lee et al., 2011) to extract part of speech labels as well as dependencies and production rules.

The syntactic features are based on both constituency and dependency parses, i.e. both production rules and dependency types.

3.3 Results

We used Weka and 10-fold cross-validation. We experimented with: logistic regression (LR, 10^8 ridge), SVM, Naive Bayes, perceptron, decision trees (DT), voted perceptron (VP), and random forest (RF).

The results are summarized in Table 1. DT and LR performed best among basic classifiers. DT outperforms LR, and the ensemble methods (VP and RF) far outperform both. An explanation is that there are deeper nonlinear dependencies between features. We believe such dependencies are worth further investigation beyond the scope of this paper. We plan to address this in future work.

In Table 1, we underlined the results for the two best *basic* classifiers, since we further analyze the features for these. Given space constraints, ensemble methods (VP and RF) are left to future work as analyzing the features and interactions based on their internal structure needs special attention.

Table 2 summarizes the feature selection results. We used Weka's feature selection. The selected features (with a positive/negative association with a deceptive role) were: number of words (negative); number of pronouns, adjectives, nouns (positive).

In order to observe each feature's individual contribution, we also performed manual feature selection, removing one feature at a time. Removing the following features improved or did not affect the performance, increasing the F1 score from 64.9 to 66.1 : number of self references, number of adjectives, number of long words, number of conjunctions or of connectors (but not both), number of cognitive words, number of pronouns.

3.4 POS and Syntactic Features

We repeated the experiments with POS tags as features (POS model), and then with syntactic features, i.e. dependency types and production rules (POS+dep, POS+con, and POS+syn models). For each model, the *baseline* feature set is the set of psycholinguistic features used in the previous section. The subsequent models use both the baseline features and the syntactic features, e.g. POS+con uses lexical-level psycholinguistic features, POS tags, and production rules. We also used tf-idf weighting.

Model	Acc.	F1	Prec.	Rec.	AUC
SVM	57.2	56.2	58.9	57.2	57.9
Perc	62.77	62.6	63.5	62.8	67
LR	**64.91**	**64.9**	**65.1**	**64.9**	66.8
NB	55.92	53.7	58.9	55.9	68.6
DT	**84.45**	**84.4**	**84.9**	**84.5**	**87.4**
VP	65.34	62.2	70.7	65.3	65.6
RF	90.87	90.8	91.2	90.9	98.1

Table 1: Werewolf classification: *Perc* - Perceptron, *LR* - Logistic Regression, *NB* - Naive Bayes, *DT* - Decision Tree, *VP* - Voted Perceptron, *RF* - Random Forest

Model	Acc.	F1	Prec.	Rec.	AUC
bfs	64.91	64.9	65.1	64.9	66.8
afs	62.625	62.4	62.6	63.4	63.4
mfs	**66.76**	**66.8**	**66.9**	**66.8**	**68.9**

Table 2: Experimental results using logistic regression: *bfs* - baseline feature set; *afs* - model on a subset of features generated with CFS-BFS feature selection; *mfs* - model on a manually selected subset of features

Table 3 suggests that production rules highly improve performance. An explanation is that complex syntax reflects cognitive complexity. For example, the utterance: *Player A said that I was innocent, which I know to be true* has many subordinates (SBAR nodes , SBAR $\rightarrow$ IN S, SBAR $\rightarrow$ WHNP S), whereas *Anyone feeling particularly lupine?* has elliptical structure (missing S $\rightarrow$ NP VP). There is also overlap with lexical features (IN nodes).

4 Discussion and Conclusions

Inspecting the decision tree, we found that most non-werewolf players used few words, no connectors, and no negations. Most werewolves use more words, adjectives, few negative emotion words, and not many words greater than 6 letters. Some werewolves use sense words and few negative emotion words, whereas others use no sense words and few or no hedges, self references, or cognitive words.

Feature set	Acc.	F1	Prec.	Rec.	AUC
baseline	64.91	64.9	65.1	64.9	66.8
POS	67.33	67.3	67.3	67.3	72.1
POS+dep	76.87	76.8	76.9	76.9	79.9
POS+con	**90.59**	**90.6**	**90.8**	**90.6**	**92.1**
POS+syn	**91.58**	**91.6**	**91.8**	**91.6**	**91.2**
POS+syn (tf-idf)	**92.287**	**92.3**	**92.3**	**92.3**	**91.8**

Table 3: POS and Syntactic Features (Logistic Regression): *baseline* - lexical psycholinguistic features, also used in subsequent models together with new features; *POS* - POS features; *dep* - dependency features; *con* - constituency features (production rules); *syn* - syntactic features; *POS+dep/con/syn* - POS and dependency/constituency/syntactic features

The conclusion is that werewolves are more verbose and moderately emotional, whereas non-werewolves are usually quiet, non-confrontational players. Werewolves also use moderately complex language, which can be explained by the fact that they are both actively trying to persuade other players, and under the cognitive load of constantly adjusting their plans to players' comments, and maintaining a false image of themselves and others.

This aligns with previous findings on low cognitive complexity for maintaining a lie (Tausczik and Pennebaker, 2010) and verbosity for highly motivated deception (Porter and Yuille, 1996).

Inspecting the odds ratios (OR) of the features in the logistic regression classifier, we found the following features to be most relevant: *TTR* (3.49), *number of hesitations* (0.91), *number of negative emotions* (1.16), *number of motion words* (1.25), *number of sense words* (0.76), *number of exclusions* (1.31), *number of connectors* (0.92), *number of conjunctions* (0.92), *number of prepositions* (1.32).

On the connection between werewolf roles and persuasion, TTR is indicative of persuasive power as well as of a werewolf, and the number of hesitations is a marker of powerless language, and is negatively associated with a werewolf role.

The fact that the number of prepositions is indicative of a werewolf role aligns with Sapir's findings, whereas the positive influence of negative emotion and motion words and the negative influence of connectors and conjunctions is as predicted by (Tausczik and Pennebaker, 2010). However, (Tausczik and Pennebaker, 2010) cite the number of sense words as highly associated with deception and the number of exclusions, with honesty. We found that in our case these associations are reversed.

One possible explanation regarding the number of sense words can be the fact that *seeing*, a family of sense words, is overloaded in this data set, since *seer* is a legitimate game role, with actions (*seeing*) that carry a specific meaning. Another explanation is that, since the transcripts are from online game, there is no actual sensing involved.

As for the number of exclusions, (Tausczik and Pennebaker, 2010) list it as a marker of cognitive complexity, which should be affected by any attempt to maintain a false story. But here most players do not actively lie, so there is no false story to maintain, and therefore no toll on cognitive complexity.

Another observation is that features suggested for highly motivated deception (longer utterances, more self references, and more negations) are not important for this data set. It is possible that we do not have highly motivated deception, since any motivation is mitigated by the context, which is a game. This suggests that deception in dialogue contexts as well as game contexts is different than in the story-telling contexts analyzed in previous work in psycholinguistics (Hosman, 2002; Porter and Yuille, 1996). On the other hand, *identity concealment* is different than other kinds of highly motivated deception – in this particular case it might be more helpful to appear logical, rather than emotional.

In this paper we presented a simple model to serve as a baseline for further models of deception detection in dialogues. We did not consider word sequence, player interaction, individual characteristics of players, or non-literal meaning. However, the data set is too small for any more complex models. We believe our results shed light on some mechanisms of deception in the Werewolf game in particular, and of deception and manipulation in dialogues in general. We plan to collect more data on which we can employ richer models that also take into account utterance sequence and dialogue features.

References

Brendan Barnwell. 2012. brenbarn.net. `http://www.brenbarn.net/werewolf/`. [Online; accessed 1-April-2016].

Anthony G. Greenwald, Rosita Daskal Albert, Dallas Cullen, Robert Love, and Joseph Sakumura Who Have. 1968. Cognitive learning, cognitive response to persuasion, and attitude change. pages 147–170. Academic Press.

M. A. Hall. 1998. *Correlation-based Feature Subset Selection for Machine Learning*. Ph.D. thesis, University of Waikato, Hamilton, New Zealand.

Lawrence A. Hosman. 2002. Language and persuasion. In James Price Dillard and Michael Pfau, editors, *The Persuasion Handbook: Developments in Theory and Practice*. Sage Publications.

Hayley Hung and Gokul Chittaranjan. 2010. The idiap wolf corpus: Exploring group behaviour in a competitive role-playing game. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, pages 879–882, New York, NY, USA. ACM.

Nitin Jindal and Bing Liu. 2008. Opinion spam and analysis. In *Proceedings of the Conference on Web Search and Web Data Mining (WSDM)*, pages 219–230.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. In *Proceedings of the CoNLL-2011 Shared Task*.

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*, ACLShort '09, pages 309–312, Stroudsburg, PA, USA. Association for Computational Linguistics.

Saif M Mohammad and Tony Wenda Yang. 2011. Tracking sentiment in mail: how genders differ on emotional axes. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (ACL-HLT 2011*, pages 70–79.

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.

Stephen Porter and John C. Yuille. 1996. The language of deceit: An investigation of the verbal clues to deception in the interrogation context. *Law and Human Behavior*, 20(4):443–458.

John R. Sparks and Charles S. Areni. 2008. Style versus substance: Multiple roles of language power in persuasion. *Journal of Applied Social Psychology*, 38(1):37–60.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 347–354, Stroudsburg, PA, USA. Association for Computational Linguistics.