Liars and traitors in Diplomacy, and how to catch them

Ludovic Delisle, Clara Grandry and Julian Raub
Team CLJ, EPFL 2020 ADA Course, Milestone P4 : creative extension

I. Introduction

As part of our project on studying linguistic harbingers of betrayal, relying on the work of Niculae et al. [1] on the online game Diplomacy, we found another Diplomacy dataset which explores how players deceive each other by lying or telling the truth. Using this data set we want to explore the relationship between betrayals and lying, to see if there are some common markers that would reveal specific dynamics of betrayal based on the ability of the player to detect a lie.

II. Data

A. Data set

We used two datasets : the "betrayal" dataset from [1], and the new "lies" dataset, from Peskov et al. [2]. The new data set consists of 17000 individual messages in which players were asked to annotate if the message they sent was intended to be a lie or not. The receiver of the message was in turn also asked to annotate if s/he believed the other one to be telling the truth or not. The raw dataset contains some meta-features that we are not interested in, but also useful data, including : the text of the message, the players involved, the annotations (lie or truth) from the players, and the in-game year of the message.

B. Pre-processing

The tricky part of the pre-processing was to extract the features "politeness, sentiment and discourse" the same way they extracted them for the paper: "Linguistic Harbingers of Betrayal: A Case Study on an Online Strategy Game" [1]. It was difficult because it required a lot of natural language processing notions which was quite new to us. We used tools provided by ConvoKit [3].

1) Politeness: For the politeness feature, we trained a classifier model with the Stanford Politeness Corpus. As described in [4], it is a corpus coming from Wikipedia discussions where every message has been labelled "polite/impolite". With this model, we made predictions on the messages from our data set. The predictions are scores between 0 and 1 (log-odds of the logistic regression), 0 meaning "certain that the text is impolite" and 1, "certain that the text is polite".

2) Sentiment: For the sentiment feature, we used the Stanford coreNLP [5], which enables users to extract linguistic annotations from text, including token and sentence boundaries, parts of speech, named entities, numeric and time values, dependency and constituency parses, coreference, sentiment, quote attributions, and relations. Like in the paper, we used it to extract the sentiment feature from each sentences of the messages which returned 1, 0 and -1 when the sentence was respectively positive, neutral and negative. Then, the "sentiment score" was computed as the percentage of sentences with positive sentiment in a message. The specificity of coreNLP is that is runs with Java, and we had to run it through a server to make it work with Python. The whole process could take up to a minute for the longest messages; thus, extracting the sentiment from every messages took almost the whole day.

3) Discourse markers: For the discourse feature, we used the Penn Discourse Treebank 2.0 [6] in order to get a list containing every discourse marker. From that list we deleted the markers that they didn't use in the paper (and, for, but, if, as, or, and so). These are markers that are too common and that would appear in over 20% of the messages. Then we tokenized each messages into single tokens, bigrams, trigrams and fourgrams and searched the discourse markers among those Ngram tokens. Finally, the discourse score was computed as the average number of markers per sentence.

III. Method

A. Prediction of betrayal

First, we trained a model on the original betrayals dataset, in order to predict if a relationship would end in betrayal. We extracted the features that were available in the new dataset : Politeness, Sentiment and Discourse, considering each of these features for both players (the betrayer and the victim), therefore training the model on 6 features, each game season being a classification instance. We used a regularized (L1) logistic regression, and performed 5-fold cross validation to estimate the accuracy of the prediction. The model yielded an average accuracy of 50.1% and a Matthews correlation coefficient of 0.015. This shows that the model cannot predict a betrayal accurately, and the predictions are globally random. In the original paper [1], the authors reported an accuracy of 57% and a Matthews correlation coefficient of 0.14, and used more diverse features as predictors (positive sentiment, planning discourse, sentences, expansion, comparison, contingency, number of words, negative sentiment).

But here, we did not have all these features at hand for the new dataset, and due to the very high computational cost of features extraction and limited resources (computational, and time), it was not possible for us to extract these additional features. We still tried to use the model to predict betrayals on seasons from the new dataset, but it only predicted lasting friendships and did not find any betrayal. However, one interesting observation we made is that the politeness of the victim and the discourse of the betrayer were systematically found to be significant predictors with a 5% threshold.

B. Exploration of lies and truths

Each message of the new dataset is annotated by the two players involved : for the sender, whether s/he was lying or not, and for the received, whether s/he thought s/he was being lied to. Initial exploration of these features show that the dataset contains a total of 882 intended lies and 16354 intended truths. Among these, 14275 truths (87,3% of truths) were correctly perceived as truths, and 661 (4% of truths) perceived as lies (Cassandra's curse) ; only 109 lies (12.4% of lies) were caught, and the remaining 688 (78% of lies) managed to deceive the receiver. The missing percentages are messages where the perception of the receiver is not known. We explored the distribution of our features of interest (politeness, sentiment, discourse) between lies and truth, and also compared these distribution to how the message were perceived. Finally, we still tried to get some insight about the dynamics between the lies and the betrayals during the game. We assigned the roles of betrayer and victim to the players for each pair of players (which was also needed for the classification task). This information being absent of the new dataset, we based the role attribution on the politeness of the players, because we know from Figure 1 [1] that betrayers are significantly more polite than victim in situations of betrayal, and if it is a lasting friendship, there is no significant imbalance, so we can arbitrarily choose the more polite player to be the betrayer.

IV. Results

The plots were generated using Seaborn. In all the figures, the error bars mark 68% confidence intervals.

V. Discussion

In figure 1 we attempted to replicate the results of the paper for the two remaining panels (since we already replicated panel (c) for the project milestone 2). We did not get the same results quantitatively, but they are qualitatively similar. Figure 2 shows interesting results : for intended lies we observe that politeness and planning discourse have significantly higher scores than for the truth counterparts. However, positive sentiment does not have a significant difference. These results seem to indicate that when a player lies, s/he is probably more careful about not revealing his/her intentions, and therefore express more politeness and more temporal discourse markers. Lies seem to be similar to betrayals in terms of politeness, but not in discourse. In Figure 3 we show the same scores, based on speaker intention and receiver perception pairs (ie. correctly detected truths/lies and wrongly detected truths/lies). This figure shows no significant (at 68%) difference in scores between correctly and wrongly guessed scores. This means this it is effectively very difficult for the players to detect lies, at least based on these parameters, which would explain the very low proportion of lies caught by the players. Finally, with the roles of betrayer and victim attributed to the players in the new dataset, we see that both players express more positivity, discourse markers and politeness when they are lying. Comparing these results to Figure 2, the difference in sentiment is now clearly visible. This effect would come from the criterion of politeness we used to assign the roles, which means that the variability in sentiment was only high when the players where not separated by politeness. This would indicate that the positive sentiment and the politeness of a player are be correlated, which seems intuitively true.

Since we do not actually know who is the betrayer and the victim and if there is a betrayal or not, we cannot draw real conclusions about the dynamics between lies and betrayals. However, we can still observe that a lie is more polite than a truth, and link this to the results found by Niculae et al. [1]. In their temporal analysis of the same features over the different seasons, they found that the betrayer is globally more polite than the victim, but this phenomenon reverses just before the betrayal happens. This leads us to two hypotheses regarding the role of lies in a betrayal : either the betrayer is lying earlier in the relationship, e.g. in seasons 2 and 3 before betrayal, where s/he is the most polite, setting the trap for the victim and deceived him/her until the last moment; or it is the victim who provokes its own betrayal by lying, and the other player detecting the lie, immediately breaks the relationship (this is less probable, given the poor ability of players to detect a lie). To conclude, there are a lot of interesting answers that could be find using a better predictor for betrayals on this new dataset (e.g. if a players betrays their friend when they perceive a lie, even if their perception is wrong), but we still managed to get useful insights about lies in the Diplomacy game. Another interesting extension could be to build a lie detector, as Peskov et al. [2] did with the same dataset.

Acknowledgements

We would like to thank the whole ADA team for the opportunity to work and this hands-on project and learn by ourselves. Special thanks to TA Lars Klein for giving us valuable insight into what was, and was not, relevant in our project ideas.

References

[1] V. Niculae, S. Kumar, J. Boyd-Graber, and C. Danescu-Niculescu-Mizil, "Linguistic harbingers of betrayal: A case

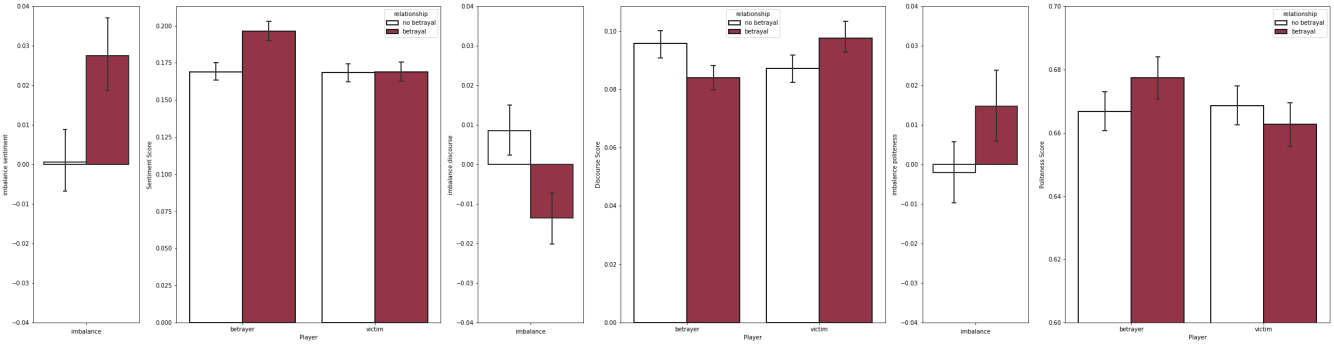

Figure 1. Positive sentiment (% of sentences), discourse markers (avg. per sentence) and politeness (avg. message score) for betrayers and victims. *Friendships that will end in betrayal are imbalanced. The eventual betrayer is more positive, more polite, but plans less than the victim.[1]*

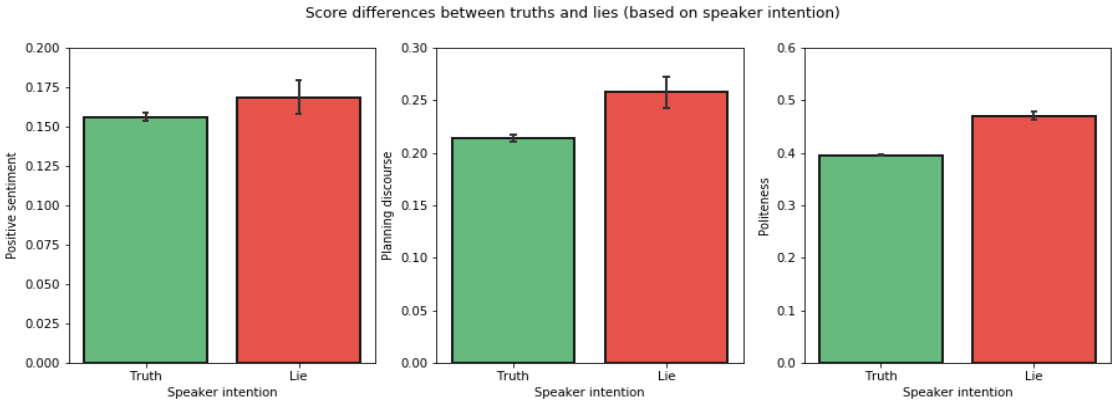

Figure 2. Sentiment, discourse and politeness scores for intended truths and lies. *Lies are systematically perceived as having a higher politeness score as well as higher planning discourse. The difference in positive sentiment is not significant at 95%*

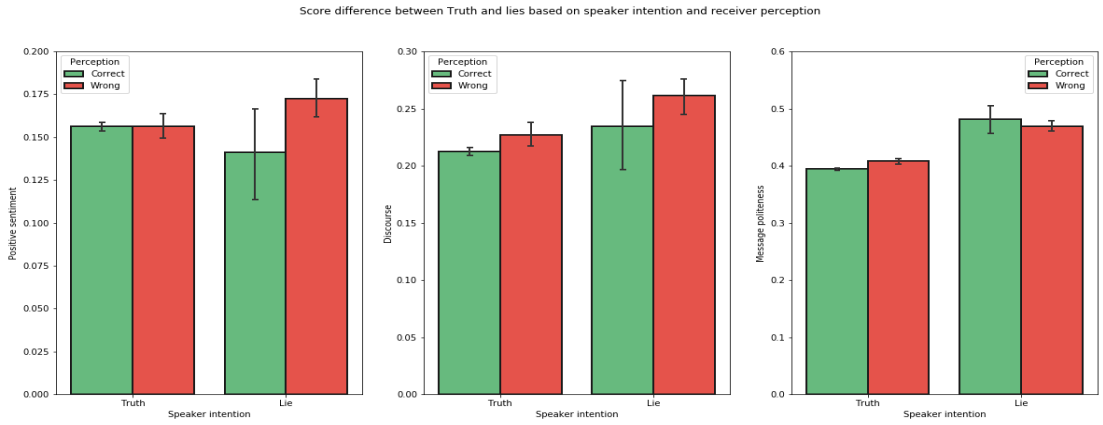

Figure 3. Sentiment, discourse and politeness scores for intended and perceived lies and truths. *Correctly and wrongly guessed truths and lies have no significant difference in either one of the three scores.*

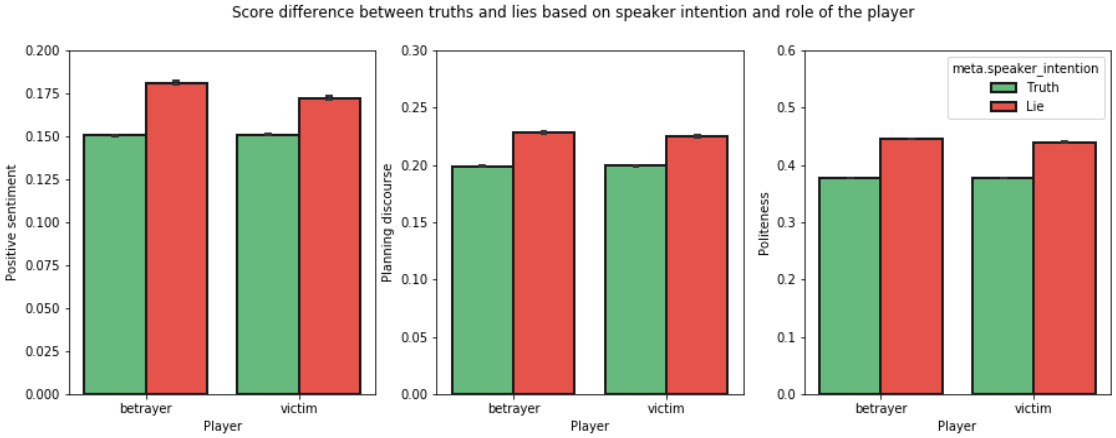

Figure 4. Sentiment, discourse and politeness scores for intended lies and truths by category of player (victim or betrayer). *Both players systematically express more positivity, discourse markers and politeness when they are lying.*

study on an online strategy game," pp. 1650–1659, jul 2015. [Online]. Available: https://www.aclweb.org/anthology/P15-1159

[2] D. Peskov, B. Cheng, A. Elgohary, J. Barrow, C. Danescu-Niculescu-Mizil, and J. Boyd-Graber, "It takes two to lie: One to lie, and one to listen," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 3811–3854. [Online]. Available: https://www.aclweb.org/anthology/2020.acl-main.353

[3] J. P. Chang, C. Chiam, L. Fu, A. Wang, J. Zhang, and C. Danescu-Niculescu-Mizil, "ConvoKit: A toolkit for the analysis of conversations," in *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 1st virtual meeting: Association for Computational Linguistics, Jul. 2020, pp. 57–60. [Online]. Available: https://www.aclweb.org/anthology/2020.sigdial-1.8

[4] C. Danescu-Niculescu-Mizil, M. Sudhof, D. Jurafsky, J. Leskovec, and C. Potts, "A computational approach to politeness with application to social factors," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 2013, pp. 250–259. [Online]. Available: https://www.aclweb.org/anthology/P13-1025

[5]

[6] R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber, "The Penn Discourse TreeBank 2.0." in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA), May 2008. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2008/pdf/754_paper.pdf