Navigating Diplomatic Discourse: A Binary Classification Approach to *Diplomacy* Generation

Aadesh Bajaj, Angela Nguyen, Srinidhi Yerraguntala
Department of Computer Science
University of Southern California
{aadeshba,angelang,syerragu}@usc.edu

Abstract

Natural Language is a powerful tool in detecting the truthfulness of online statements. This is especially true in the field of international relations where nations are often trying to maneuver conflicting interests and deceitful parties. There still exists a sizable gap in real-time human interactions in this field due to variability in human communication.

We explored the role of real-time language in a similar setting through a detailed corpus from the online seven player game *Diplomacy*, where players assume the role of a Great Power of Europe in the early 20th century and compete for dominance. The game is inherently collaborative and reflects common war strategies in international relations including forming alliances and spreading misinformation. This was originally explored in a paper from Cornell and UMD researchers in "It Takes Two To Lie: One to Lie, One to Listen" (Peskov et al., 2020). Through the *Diplomacy* corpus, they used a variety of stacked and base models to classify messages as truths or lies, citing F_1 scores as their primary performance metric. We wanted to take their research a step further by introducing additional architectures that they hadn't explored yet to classify messages. We then wanted to take the results of this classification and pass it into an autoregressive model to generate reasonable responses for diplomatic interactions.

The resulting model shows that, given context, some accurate responses can be generated to navigate negotiation periods. They can be applied to online discussions to detect lies from other people, allowing users of the model to learn what can be trusted online, and generate an appropriate response to online comments. It can also be extended to applications in international relations, given the premise of the game.

1 Introduction

1.1 International Relations

Trust and negotiation are essential for society to move forward, as people need to collaborate and rely on one another in order to grow together. Negotiation is most crucial in policymaking, specifically in terms of international relations where diplomats of all nations meet to solidify international decisions. Trust between nations allow for both to enact alike policies as well foster trust between the citizens of those nations (Torgler, 2008). However, diplomats must determine whether or not to trust one another, as misplaced trust can be detrimental to their nations.

Due to the versatility of natural language processing tasks, such as classification and sentiment analysis, there is a strong desire for its integration in politics (Gigley, 1993). Recent advancements in natural language processing have yielded significant implications for international politics, a prominent example being evidence-based policymaking and policy interpretation (Jin and Mihalcea, 2023).

1.2 Online Communication

Although the average person tends to be more trusting of those physically around them (Weiss et al., 2021), that same trust does not apply to online communication. Lack of trust towards people online stems from deception of scammers, catfishers, and general spread of misinformation, which causes people to be reluctant to participate in online transactions (Wang and Emurian, 2005). Online dishonesty becomes even more nuanced when mixed in with predominantly truthful statements during long term interactions.

There has been a lot of study on this in public statements through misinformation detectors in NLP, but there still exists a large gap for lie detection in real-time. Analyzing real-time human interactions (virtual and in real life) is a new frontier that many models struggle with compared to stand-alone statements due to variability in human communication styles (Bajaj et al., 2023).

In the online version of *Diplomacy*, during what's known as the negotiation period, players engage in private chat rooms. They send messages back and forth with opposing nations, often with the intent to betray them. The corpus we are using describes 250 such interactions (sum of 17,289 messages) where two nations engage in a conversation. Our model hopes to incorporate long term context of such interactions and be able to recognize small lies intricately woven into *mostly* truthful statements between players.

2 *Diplomacy*

Diplomacy is a strategic war game that captures the essence of diplomatic negotiations, alliances, and the art of deception, all set against the backdrop of pre-World War I Europe. In this game, players take on the role of one of the seven Great Powers of the time: England, Germany, Russia, Turkey, Italy, France, and Austria-Hungary. The aim is to gain control over the majority of 18 supply centers spread across a simplified map of Europe. [1]

Collaboration with other players is essential. When it comes to making moves, they're issued as written orders, so there's hardly any room for luck or chance—strategy is key, and the ability to work with (and against) others is what makes or breaks the game.

At its core, the game revolves around two main phases: the Diplomatic phase and the Movement phase. During the Diplomatic phase, players engage in discussions, forming alliances and negotiating deals. Agreements and alliances are not binding which allows for dynamic play and tactical deception.

Diplomacy is not just a board game but a complex exercise in strategic planning and human psychology. It challenges players to balance short-term needs with long-term strategies, all while discerning allies from potential betrayers. The game's depth comes from these strategic elements and the interpersonal interactions it fosters. This combination makes *Diplomacy* a compelling study in both strategic gameplay and the nuances of human communication and relationships.

[1] Rules found here: `https://www.hasbro.com/common/instruct/diplomacy.pdf`

3 Pre-Trained Language Models

The main benefit of fine-tuning an existing, trained language model to execute a task as opposed to building a new model to fit an objective is efficiency. Training a model on a large dataset to evaluate human language takes more computational power than the average person has access too. Additionally, the dataset must be large enough for the model to learn accurately, which can take a long time to process. Pre-trained language models like T5 and BERT are already trained and optimized, so it is much more efficient and accurate to fine-tune those models than create new ones.

3.1 Text-to-Text Transfer Transformer (T5)

Text-to-Text Transfer Transformer, or T5, is a Transformer-based model that, at its core, treats every natural language processing task as a 'text-to-text' problem. In more technical terms, both its input and output are strings. T5 is first pre-trained through self-supervised learning where the model creates labels from a generic, unlabeled corpus. Then, it can be fine-tuned on a more specialized dataset in order to complete a specific task, such as text generation or machine translation (Mastropaolo et al., 2021).

3.2 Grounded Open Dialogue Language Model (GODEL)

Grounded Open Dialogue Language Model, or GODEL, is a Transformer-based model that is trained specifically for dialog tasks like text generation. It includes grounded pre-training, meaning it is also conditioned on a corpora external to the training dataset (Peng et al., 2022). This allows the model to more effectively generate responses that require information outside the current context.

3.3 BERT and FROZEN-BERT

Bidirectional Encoder Representations from Transformers (BERT) is another pre-trained language model similar to T5; however, it is trained on masked language modeling, meaning the model predicts a word given its left and right context instead of predicting the succeeding word (Wang and Cho, 2019).

FROZEN-BERT is another derivative of BERT where its parameters are frozen after word embeddings are created, so no further updates occur. As a result, it does not require back-propagation. We can use the embeddings created by BERT in its

Message	Suspected Meaning by Opponent	Actual Meaning
Here's the deal: I like you better than England	Truth	Lie
I am not planning to keep Vienna. And yeah I've asked France for support to the Channel. Do you think he's on board?	Truth	Truth

Table 1: Samples of a lie and truth from the dataset. In the first message, the opponent they were talking to perceived it as a truth, but it was actually a lie.

frozen model and pass them into other models like Logistic Regression and Random Forest. As opposed to base BERT, FROZEN-BERT updates less layers, so it's computationally more efficient to use the embeddings in more lightweight models than to train BERT (Lee et al., 2019).

4 "It Takes Two to Lie" 2020 Research

This project was originally inspired by the 2020 research paper, "It Takes Two to Lie: One to Lie, One to Listen," conducted by Cornell University and University of Maryland researchers (shortened to IT2TL for the remainder of this paper) (Peskov et al., 2020). They explored *Diplomacy* in the context of online conversations and trust in this setting to determine that a well-trained model using various dynamics from a dataset can be trained to predict lies as well as a human can.

4.1 *Diplomacy* Dataset

Our research uses the same *Diplomacy* corpus they created to classify player statements as a truth or lie. This corpus is extremely detailed and has a lot of great information to explore binary sentence classification.

The researchers moderated 12 different games of *Diplomacy* and had players mark each message they sent and received as TRUE or FALSE. They compiled all of this data into 12 transcripts: nine are used for training, one is used for validation of tuning parameters, and the remaining two are test data. Each game's data is split into the messages exchanged between each pair of players (7 countries total, so $(C(7,2) = 21)$, yielding 189 transcripts for training, 21 transcripts for validation, and 42 transcripts for testing (242 total).The data is presented in JSON Lines format. Each game contains the following data:

- Messages: A list of strings
- Sender Labels: List of booleans in which each value corresponds to the sender's *intended* truth of the message of the same index. Can be TRUE or FALSE.
- Receiver Labels: List of booleans in which each value corresponds to the receiver's perceived belief of the message of the same index. Can be TRUE, FALSE, or NO ANNOTATION.
- Speakers: A list of strings of the country which sent each message
- Relative Message Index: A list of integers which keeps track of the index numbers of each message
- Seasons and Years: Lists of strings that correspond to the passage of time within the game
- Game Score and Game Score Delta: Numerical data that keeps track of power dynamics between the players
- Players: A list of participating countries
- Game ID: A unique integer identifier of the game

Each message in the corpus was tagged in real-time with a corresponding boolean label from both the sender and the receiver. If the sender label is TRUE, it means that the sender was sincere in the message, and if it's FALSE, the sender lied in it. The receiver's label describes how the receiver perceived the message, though some of these have NO ANNOTATION.

The most pressing issue to handle with the dataset was the skewed TRUE and FALSE classes. This plagued both the original researchers and ourselves. After deliberation and testing on various methods to fix this, we decided to implement the Synthetic Minority Oversampling Technique, or SMOTE (Chawla et al., 2002). SMOTE does well to analyze a minority class's feature space and create new synthetic data points based on the K-Nearest Neighbors algorithm. The parameters can be changed including how many k neighbors to analyze and what percentage of the minority class you want in relation to the majority class.

4.2 IT2TL Experiments

In IT2TL, they explored various stacked ensemble models and methods including LSTM+Power+BERT, Context LSTM+Power, Context LSTM+BERT, and Bag of Words+Power. The primary metrics they analyzed were MACRO (average) F_1 and FALSE class F_1 scores. They analyzed the standalone FALSE class because predicting messages in an extremely skewed dataset like this is often met with a high F_1 for the majority class and a low F_1 for the minority class. As a result, it's important to gauge the minority class score and see if there is any, even marginal, improvement. In this case, it means accuracy isn't a great metric since a bad model can still have an accuracy greater than 90%.

Using human analysis as their benchmark, the researchers observed a MACRO F_1 of 0.581 and FALSE F_1 of 0.225. This essentially tells us what we already know– picking out a small number of lies in a large set of messages is hard, even for humans. With this 0.225 FALSE classification as a baseline, they presented their results for each model. They had the most success with the Context LSTM+Power model, surpassing the human benchmark for FALSE F_1 at 0.27, but falling just short in the MACRO F_1 at 0.572.

These results were very promising and showcased a potentially successful lie-detector model to help users decide in the moment the ideal course of action. It did as well, if not better, than the human benchmark, especially in identifying FALSE statements.

4.3 Other Related Work

As previously mentioned, new research by Jin and Mihalcea 2023 illustrates the advanced role that natural language processing plays in international relations. Their model extracts data to collect online opinions on certain events and political leaders as well as facts on those subjects. Using that data, it creates informed policies.

That model is also capable of interpreting political decisions after policies are made. It learns of a policy's political agendas and compares how much that policy aligns with the public sentiment, which evaluates its effectiveness and reception. In general, most work directly related to policy, government, and international relations is relevant to sentiment analysis rather than using that to engage in discourse. This is where we hope our project could fit in.

5 Approach

Given a context of prior messages, we intended to create a model that could classify the most recent message as either a truth or a lie. From this, we intended to use this result in separate models to generate strategic responses to further the conversation. Examples of those messages are shown in Table 1.

5.1 Classification

Similar to IT2TL, we first wanted to explore the prospect of a lie classifier in *Diplomacy*. We experimented with new stacked and base models with FROZEN-BERT embeddings and including combinations of Logistic Regression, Random Forest, and Support Vector Machines (SVM). The first 3 are all well suited for binary classification and have various methods to help handle imbalance class data (`class_weight='balanced'`). We attempted to experiment with DISTILBERT but ran into problems with GPU resources, so we decided to forego this. In addition to time constraints, we felt that IT2TL had already explored BERT in various ways, so it wouldn't add enough value to work through.

The experiments we ran with multiple models utilized either the Stacked Ensemble method or the Voting Classifier method. Both are effective ways of congregating results from base models, but have variations in how the embeddings are combined. Thus, it was good to try both. The models we tested were Logistic Regression, Random Forest, FROZEN-BERT+Logistic Regression, FROZEN-BERT+Logistic Regression+SVM, and Logistic Regression+SVM.

5.2 Generation

Additionally, we wished to explore a text generation task which aids *Diplomacy* players in responding to their opponents. We explored two generative transformer models for this task, including Huggingface's T5 encoder-decoder and Microsoft's GODEL.

As opposed to their T5-BASE model, we opted for the Huggingface's T5-SMALL, a scaled-down version, to tokenize corpora inputs. Much like our reason for using FROZEN-BERT over BERT, T5-SMALL is much faster to fine-tune while also producing similar results to the T5-BASE (Raffel et al., 2020).

Additionally, we opted to use GODEL because the dataset's messages were a collection of chat messages between two players. This message format is especially conducive for a dialogue-based model such as GODEL. Furthermore, we wanted to include the output of our classification task as additional knowledge for the model when training, which is another task GODEL is useful for.

5.3 Evaluation Metrics

For the classification task, as mentioned earlier, we used F_1 scores for the FALSE class and the MACRO F_1 to assess performance. This worked best considering the class imbalance in the dataset.

For the generation task, we used a variety of metrics to evaluate the quality of responses of the three models. We first used BLEU (Bilingual Evaluation Understudy) as an overall assessment of output precision with the target sentences (Papineni et al., 2002). We also used BERTSCORE, a more robust evaluation metric that computes the token similarity between texts using contextual embeddings. This metric is more effective in capturing the semantic similarities between words as well as a better understanding of its context, making this a more meaningful evaluation metric for the back-and-forth dialogue of this dataset (Zhang et al., 2020).

Apart from these two metrics, we also wanted to apply automated linguistic analysis techniques to understand how well the generated outputs continued the conversation between two players. To do this, we computed the Jaccard index of each output. This can be used as a simple method of determining the lexical cohesion of a generation output, specifically computing how well the output continues the context (Roemmele et al., 2017).

Despite the acceleration of text generation analysis technology, human evaluation still remains as the most effective method of evaluating generated text (van der Lee et al., 2021). We created a survey which presented the results of the text generation task for all three of the models. For each model, we showed three sample outputs and their corresponding contexts. To ensure consistency, we used the same context for each model. We then asked the respondents to rate the generated response's coherence and relevance on a scale of 1 to 5. The participants of this survey were primarily university students in Model United Nations, as these students regularly participate in conferences where they must negotiate with their peers to pass mock resolutions.

6 Putting It Together

We also wanted to use the predictions of our most successful classifier model as an additional input when fine-tuning the T5 and GODEL. This input, along with the context of the conversation, aims to influence generation and help users navigate the intricacies of diplomatic conversations.

We anticipate that accurate predictions of TRUE and FALSE messages can help generate reasonable responses between nations in *Diplomacy* and beyond. Overall, we wanted to see if there's potentially a useful application for our model in the context of International Relations and diplomatic mediation.

7 Experimentation and Results

7.1 Classification

Table 2: Classification Model Performance Comparison

Model	FALSE F_1	MACRO F_1
Human Benchmark	$\mathbf{0.225}$	$\mathbf{0.581}$
IT2TL Best Model	$\mathbf{0.27}$	$\mathbf{0.572}$
LR (Logistic Regression)	$\mathbf{0.22}$	$\mathbf{0.55}$
Random Forest (and others)	0.00	0.48
LR+SVM	0.11	0.53
FROZEN-BERT+LR+SVM	0.16	0.51
FROZEN-BERT+LR	0.20	0.52

In the classification models, we iterated through all messages in the dataset and stored the current with its preceding 14 messages as context as an index. We also added an instruction to the query. The result was: "[INSTRUCTION] Given a dialog context and a message, you need to classify the message as TRUE or FALSE. [CONTEXT] {context} [CLASSIFY] {message}." We also explored modifying the input by adding various inputs from the corpus, such as `speaker: speaker` or `game score: game_score`. For classification, this didn't provide much value and often resulted in lower scores.

Our first classification attempt was with FROZEN-BERT embeddings using Logistic Regres-

Model	BLEU	BERTSCORE	Jaccard	Human Evaluation (Coherence)	Human Evaluation (Relevance)
3-GRAM	0.216	0.619	0.037	1.29	1.29
T5	**0.224**	0.814	0.036	3.41	2.88
T5 + Predictions	0.165	0.817	0.036	2.54	2.13
GODEL	0.143	0.832	**0.038**	3.17	**3.08**
GODEL + Predictions	0.171	**0.833**	**0.038**	**3.63**	2.96

Table 3: Scores achieved by the generation models against various evaluation metrics. Human evaluation scores are the average of survey responses between 1-5 while all other metrics are between 0-1.

sion from the `sklearn` library as the classifier. This achieved *some* success when implemented with SMOTE. We explored other ensemble methods with FROZEN-BERT and without, and realized our most successful model was a standalone Logistic Regression model. The embeddings we used in lieu of FROZEN-BERT were from `TfidfVectorizer` in `sklearn` resulting in much simpler word embeddings. It's likely FROZEN-BERT wasn't super effective here because of the size of the dataset and instructions– we simply didn't need complex vectors to capture the tokens effectively.

Another note is that we often ran into the problem of a model optimizing accuracy and classifying every message as true in testing, meaning the FALSE F_1 score would be 0 and the MACRO F_1 would be approximately 0.48. Obviously, these are not results representative of actual predictions even if the accuracy ends up being over 90%. In general, models that were more complex faced this issue. Some examples include Random Forest, XGBoost, Random Forest + XGBoost, and Random Forest + Logistic Regression. While these models struggled with overfitting and were unable to generalize well, simpler models like Logistic Regression provided a more balanced approach.

In relation to the human benchmark and IT2TL's Context LSTM+Power model, our Logistic Regression with `TfidfVectorizer` embeddings fell just short of desired performance. In general, the FALSE F_1 scores for all 3 indicate most messages are believable and predicting lies is a very difficult task. While our dataset may have been well suited for simpler models like Logistic Regression to generalize inputs more effectively, we likely should have utilized more combinations of inputs and methods to overcome class imbalance and potentially increase our performance. Despite this, we felt the model still performed fairly well given our resources, and we chose it to proceed in our generation task.

7.2 Generation

We first created a 3-GRAM language model to use as a baseline for the task of generating a response to continue the conversation between two parties in the game of *Diplomacy*. We created word probability distributions using the training data and generated results at each iteration based on the previous iterations' final two words as the prefix sequence.

We then trained T5 and GODEL by fine-tuning on these hyperparameters: 3e-4 learning rate, 3 training epochs, 500 warm-up steps, and 0.01 weight decay. For each input, we created a specialized prompt to give the model information about its task. For each iteration, we used the previous four messages exchanged between the players as context for the conversation, separated by an "EOS" token. We also gave the instruction, "Given a dialog context, you need to respond strategically." As a follow up, for both T5 and GODEL, we also used the outputs of the logistic regression model for prior knowledge about the truthfulness of the last message in the context. For these tasks, we altered the instruction to read, "Given a dialog context and prior knowledge about the truthfulness of the most recent message, you need to respond strategically."

After the model was finished training, we generated responses with a max length of 100 words with top-p sampling of $p = 0.95$, meaning that tokens are sampled only from a group of most probable tokens, where the collective probability is 0.95 or more. This allows for the model to choose from a more diverse range of tokens while still prioritizing those with higher probabilities (Holtzman et al., 2020).

The results of the different models' generations can be seen in Table 3. Using the 3-GRAM eval-

Context	**Player 1:** Long term how should be dispose of the remaining powers? I think Germany is the clear next target for me and Italy for you? **Player 2:** Then I'll be strong enough to push England off the mainland and hit them by sea from the north and France won't be much of a problem? The only thing we have to fear is any strong alliances between the remaining players. **Player 1:** Position is key: you need to get well pushed over the stalemate line, and I'll be focussing on getting a bit past Italy to France, in fact (hence my new fleets) **Player 2:** Makes sense. Munich is my main goal for next year, Kie as well ideally.
3-GRAM	Cleverness lining NRG cordially tapping leading squeezed pushing impressions giving
T5	But we have enough options to make sure we protect France from these attacks
T5 + Predictions	Oraccounts would be wonderful. You didn't see this or was you going out at all?
GODEL	I think you should take Serbia! I'll take ruh and lead you to ruh, then you and I can secure stp next season, taking back Spl and Carr. You can hold Serbia, if you want. The truth is that I'll get 2 builds and Turkey will get 2 builds, and I will get 3.
GODEL + Predictions	Might make the decision against you. If you want to start a war on Austria, you can go all in against Germany, and you will bounce them off of me.

Table 4: An example of responses generated from each generation model given a conversation context between two players. The response is in the viewpoint of Player 1 replying to Player 2's last message.

uation scores as a baseline, our models generally performed better, indicating that we were able to produce meaningful responses to an extent. One notable exception is BLEU score, in which the 3-GRAM outperformed all of the transformer-based models, except for T5. This may be because BLEU score measures the precision of the n-grams from the target and generated responses, meaning a 3-GRAM is likely to be more similar to the target than a model that uses random sampling. The BERTSCORES are much higher than the BLEU scores overall, which may be explained by the ability of BERT to capture contextual information better. Our GODEL model that was trained with the results of logistic regression for classification performed the best, with a score of 0.833. The Jaccard scores are similar overall, but they achieve a better score than the human gold standard achieved in previous works (Roemmele et al., 2017).

The human evaluation scores indicate that our models were able to produce relatively coherent and relevant responses. While there is not much difference between the T5 models and GODEL models for the coherence scores, the GODEL models produce, on average, more relevant responses. This can be explained by GODEL being a dialogue-specific model which is able to understand the context of the conversation better than T5.

Furthermore, although adding the classification labels increased the BERTSCORE for both models, it did not yield a significant improvement in either coherence or relevance scores. This means that there was no consistent effect of adding the prediction labels. One possible explanation for this is that adding the labels confused the model, and changing the prompt is something that is left to be desired. Another explanation of this is that the labels themselves were not completely accurate, meaning they did not give the model any meaningful information about the context. One thing to note about the human evaluation scores is that the survey used three random outputs from each model which may not have been the strongest examples to convey the overall effectiveness of the models.

8 Conclusion

IT2TL did well to set a baseline by examining this dataset purely in terms of truths or lies, and their conclusions had significant implications for *Diplomacy* gameplay and other industries like health, cybersecurity.

Building upon the foundation laid by IT2TL, our research takes a step forward in not just identifying but also facilitating real diplomatic communications. We've presented preliminary success in training our T5 and GODEL models to generate coherent and strategic responses in *Diplomacy*, and our research could extend further into policymaking, government, and negotiation. This leap from detection to generation offers profound possibilities for application in the field of international relations.

More work is needed in optimizing the performance of our model, most notably in GODEL where few-shot prompting may improve the coherence and/or relevance of generated responses. Furthermore, more successful classification models like the one presented in IT2TL could yield more promising results. There exists a whole new avenue of automated political strategy that could be used to inform diplomats in formal and informal settings. Not only does our classification-generation model begin to understand the situations, it partakes in them.

References

Nikesh Bajaj, Marvin Rajwadi, Tracy Goodluck Constance, Julie Wall, Mansour Moniri, Thea Laird, Chris Woodruff, James Laird, Cornelius Glackin, and Nigel Cannings. 2023. Deception detection in conversations using the proximity of linguistic markers. *Knowledge-Based Systems*, 267:110422.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. Smote: Synthetic minority over-sampling technique. *J. Artif. Int. Res.*, 16(1):321–357.

Helen M. Gigley. 1993. Projected government needs in human language technology and the role of researchers in meeting them. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration.

Zhijing Jin and Rada Mihalcea. 2023. *Natural Language Processing for Policymaking*, pages 141–162. Springer International Publishing, Cham.

Jaejun Lee, Raphael Tang, and Jimmy Lin. 2019. What would elsa do? freezing layers during transformer fine-tuning.

Antonio Mastropaolo, Simone Scalabrino, Nathan Cooper, David Nader Palacio, Denys Poshyvanyk, Rocco Oliveto, and Gabriele Bavota. 2021. Studying the usage of text-to-text transfer transformer to support code-related tasks.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Baolin Peng, Michel Galley, Pengcheng He, Chris Brockett, Lars Liden, Elnaz Nouri, Zhou Yu, Bill Dolan, and Jianfeng Gao. 2022. Godel: Large-scale pre-training for goal-directed dialog.

Denis Peskov, Benny Cheng, Ahmed Elgohary, Joe Barrow, Cristian Danescu-Niculescu-Mizil, and Jordan Boyd-Graber. 2020. It takes two to lie: One to lie and one to listen. In *Association for Computational Linguistics*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Melissa Roemmele, Andrew S. Gordon, and Reid Swanson. 2017. Evaluating story generation systems using automated linguistic analyses.

Benno Torgler. 2008. Trust in international organizations: An empirical investigation focusing on the united nations. *The Review of International Organizations*, 18(3):65–83.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:101151.

Alex Wang and Kyunghyun Cho. 2019. Bert has a mouth, and it must speak: Bert as a markov random field language model.

Ye Wang and Henry Emurian. 2005. An overview of online trust: Concepts, elements, and implications. *Computers in Human Behavior*, 21:105–125.

Alexa Weiss, Corinna Michels, Pascal Burgmer, Thomas Mussweiler, Axel Ockenfels, and Wilhelm Hofmann. 2021. Trust in everyday life. *Journal of Personality and Social Psychology*, 121(1):95–114.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.