Distinguishing Address vs. Reference Mentions of Personal Names in Text

Vinodkumar Prabhakaran
Google Research
San Francisco, CA, USA
vinodkpg@google.comn

Aida Mostafazadeh Davani
Google Research
Portland, OR, USA
aidamd@google.com

Melissa J Ferguson
Yale University
New Haven, CT, USA
melissa.ferguson@yale.edu

Stav Atir
University of Wisconsin-Madison
Madison, WI, USA
stav.atir@wisc.edu

Abstract

Detecting named entities in text has long been a core NLP task. However, not much work has gone into distinguishing whether an entity mention is addressing the entity vs. referring to the entity; e.g., *John, would you turn the light off?* vs. *John turned the light off*. While this distinction is marked by a *vocative case* marker in some languages, many modern Indo-European languages such as English do not use such explicit vocative markers, and the distinction is left to be interpreted in context. In this paper, we present a new annotated dataset that captures the *address* vs. *reference* distinction in English,[1] an automatic tagger that performs at 85% accuracy in making this distinction, and demonstrate how this distinction is important in NLP and computational social science applications in English language.

1 Introduction

Named entity recognition (NER) in text has long been a core task in the NLP community (Sundheim, 1995; Yadav and Bethard, 2018). However, not much work has looked into distinguishing whether an entity mention is an instance of addressing the entity or referring to them:

- *John, would you turn the light off?* (Address)
- *John turned the light off.* (Reference)

The address usage is also called a *vocative phrase*: "a noun phrase which does not belong to the thematic grid of a predicate and is used to attract someone's attention" (Moro, 2003). Many languages have explicit morphological *vocative case* markers: e.g., in "Et tu, Brute?", Brute marks the vocative case of the nominative Brutus. However, many modern Indo-European languages, including English, do not have vocative case markers, and the distinction is left to be interpreted based on context.

Distinguishing vocative phrases is important in many NLP tasks, such as sentiment analysis (Karami et al., 2020), offensiveness detection (Mubarak et al., 2020) and information extraction (Makazhanov et al., 2014). For instance, Karami et al. (2020) point out the difference in interpretations between *"Let's eat, Grandma"* and *"Let's eat Grandma"*. The vocative distinction is also important for NLP-aided computational social sciences, since the pragmatics and the patterns of usage vary between these two types of name mentions (Dickey, 1997), and since name mentions capture various societal biases (Prabhakaran et al., 2019). This aspect is especially crucial in studies analyzing political discourse, with the goal of understanding the rhetoric by and about political personalities (Prabhakaran et al., 2014; Gupta, 2022).

Despite the prevalence of NER as a useful task in various NLP applications (Marrero et al., 2013), efforts to make this distinction have largely been limited to languages that have explicit vocative case markers such as Portuguese (Baptista and Mamede, 2017), Hebrew (Tsarfaty et al., 2019), Korean (Nam and Choi, 1997), and Sindhi (Muslim and Bhatti, 2010), and not much work has looked into detecting vocative name mentions in English.

In this paper, we present a dataset of social media text in the political domain in English language, with person mentions annotated with the *address* vs. *reference* distinction. We then build a tagger that is able to make this distinction automatically, with an accuracy of 85%. We use this tagger to demonstrate the importance of this distinction in two large-scale computational socio-linguistic analysis. First,

[1] https://stavatir.com/s/address-vs-reference.xlsx

Findings of the Association for Computational Linguistics: ACL 2023, pages 6801–6809

we demonstrate that female personalities are more likely to be mentioned in the addressing context than male personalities, across three different social medial corpora, which has implications for NLP research on gender bias in data and models. Second, we demonstrate that sentences with address mentions are significantly more likely to be toxic than those with reference mentions. This finding has important implications for the active area of NLP research on detecting online abuse.

2 Address vs. Reference Mentions

How a person is addressed or referenced in language, and its associated pragmatics has long been of interest in sociolinguistics (Brown et al., 1960; Brown and Ford, 1961). While most of this research focused on the different address pronouns and the T/V distinction, much less work has looked into the difference in the social meaning of a mention when used as an address vs. when used as a reference (Dickey, 1997). While this distinction is not limited to persons (for instance, organizations may also be mentioned in an addressing context, as in *Hey Doordash, where is my food?*), person name mentions add additional nuance owing to the social relations. For instance, Dickey (1997) show that the words used to address a person by a speaker may differ from the words used to refer to them depending on the social power relations between the speaker, the referent, and the addressee.

Forms of address has been studied in NLP-aided computational sociolinguistics, for instance, in the context of how they relate to social power relations (Prabhakaran et al., 2013). The address vs. references distinction has also been shown to be of value in NLP tasks, for instance, Mubarak et al. (2020) extracts Arabic tweets with the vocative particle "*yA*" as it indicates directing speech to a person or a group, increasing the likelihood of offensiveness. However NLP work on making this distinction is largely limited to languages that have explicit vocative case markers. In the absence of any vocative markers, as in English, this becomes a task that relies on the syntactic context. In this paper, we build resources to perform and evaluate this distinction, and demonstrate its utility in NLP applications.

There is related work in NLP on detecting addressees in multi-party dialog (op den Akker and op den Akker, 2009; Ouchi and Tsuboi, 2016; Le et al., 2019; Ek et al., 2018), which is a substantially different task from ours. First, addressee detection in multi-party dialog takes into account the larger dialog/content context (e.g., prior utterances). For instance, Ouchi and Tsuboi (2016) jointly captures "*who* is talking about *what* at each time step" in order to determine the addressee. Ours is a simple linguistic task that relies on the local syntactic context of named mentions, making it applicable in broader contexts. Second, the above work crucially looks into the implicit cues about addressees. In contrast, our work focuses only on explicit mentions, primarily motivated by the computational social science analyses anchored on them.

2.1 Data

Source: We use the corpus of Facebook comments on politicians' posts released by (Voigt et al., 2018) for this study. Our choice is motivated by three reasons. First, the comments in this corpus are all made in response to a individual's Facebook post and hence it is likely for it to have more instances of comments addressing the person than general social media data with mentions of that person. Second, the corpus captures the individual's name within the metadata, making it easy to detect and disambiguate different mentions referring to the same person. Finally, the corpus also captures the gender information of the person the comments are in response to (unlike most other gender-labeled data that captures the gender of the speaker/writer) as it was originally developed to study gender bias in social media, which is one of our goals too.

Pre-processing: Since the metadata captures the politician's name that each comment is in response to, we use a regex-based approach to determine if that politician is mentioned in the comment or not. We made sure the regex captures different forms of address including full name mentions, first name mentions, and last name mentions. Furthermore, since the corpus contained comments directed at only 402 politicians, we manually coded different common variations and misspellings of their first and last names. For instance, the first name of the politician *Jim Boozman* could be mentioned as *Jim*, *James*, or *Jimmy*, and the common variations of his lastname included *Boozman*, *Boozeman*, and *Bozeman*. While some of these choices may be genuine misspellings, some others may indicate pragmatic connotations: *Jimmy* instead of *Jim* may have been used to evoke familiarity, while *Boozeman* instead of *Boozman* may have been intended to evoke humor or disrespect. We do not analyze

these distinctions in this paper, however, we included them in our regex to ensure that we capture such diverse associated linguistic contexts.

Annotation: We sampled 800 comments with at most 100 words (to avoid exceedingly long comments) from the corpus. We restricted ourselves to only those comments with a single mention of the individual (i.e., removed comments with no or multiple mentions). Multiple mentions were rare in our data (less than 1%), and when they do happen they were almost exclusively all reference mentions, as it is unlikely for someone to address someone by name, and then refer to them in third person in the same sentence itself. We trained two annotators to make the *address* vs. *reference* distinction. The annotators were undergraduate students majoring in Psychology at Yale University. Annotators were provided with the comments, the individual whose post the comment was in response to, as well as the mention of that individual detected in the comment. They were asked to label whether the mention was addressing the individual vs. referencing the individual, along with examples.

Analysis: All comments were double annotated, obtaining an inter-annotator agreement of $\kappa = 0.898$, suggesting that the task is relatively easy for trained humans, and that our annotations capture reliable data. We then performed an adjudication round where both annotators met with one of the authors and arrived at a final label through discussion. While most disagreements were due to misinterpretations, some cases were inherently ambiguous. For instance, in "*Yes!!! Sen. Booker*", it is ambiguous whether the commenter is addressing Sen. Booker or just mentioning him.

The annotation and adjudication process revealed 15 comments where the name mention was not valid; e.g., within a URL contained in the comment, and 11 comments where the comment did not have enough linguistic context to make the distinction; e.g., when the comment was just a name mention. We removed these comments as they will add noise, resulting in 774 comments in the dataset, each with a mention labeled as either *address* or *reference*. There were 250 (32.3%) instances that were the *address* usage compared to 524 (67.7%) instances that were the *reference* usage.

2.2 Automatic Tagger

We now investigate automatically distinguishing *address* vs. *reference*, given a text and a name mention in it. Since contextualized embeddings such as BERT (Devlin et al., 2019) are proven to capture syntactic information (Clark et al., 2019), we expect the positional embedding of the name mention to capture its syntactic context and hence help make this distinction. Further, we use the intuition that *reference* mentions are more likely to occur in syntactic contexts where third person pronouns could fit, while *address* mentions are more likely to fit second person pronouns or address terms. We consider three settings, each with two sets of words that fit with the *address* vs. *reference* contexts:

S1: you/your vs. he/him/his/she/her
S2: you/your vs. he/him/his/she/her/they/them
S3: you/your/hey/hi vs. he/him/his/she/her

S1 uses singular pronouns, S2 includes the (usually) plural pronouns *they*/*them*, S3 includes addressing terms (hey/hi). For each setting, we use a contextual embedding, replace the mention with [MASK] and calculate the score for each word in the list to fit the masked slot. If the top scored word from the list is of the *address* category, we predict the mention as *address*, otherwise, as *reference*. To illustrate, the top candidate from S3 above for the input "*[MASK], would you turn the light off?*" as per BERT is *hey*, while the top candidate for "*[MASK] turned the light off*" is *he*, then *she*.

This approach is not entirely foolproof, but as Table 1 shows, this simple approach yielded good performance of 85% accuracy. We report results using BERT and DistillBERT models across all three settings outlined above. Adding addressing terms *hey* and *hi* increased the accuracy, while adding the third person pronouns *they* and *them* that are usually used in plural context (but also has singular usage) resulted in reducing the accuracy.

Most errors happen when the sentence is not well-formed or uses non-standard language. An approach to circumvent this issue is to fine-tune a pre-trained model using our data. In our preliminary experiments, fine-tuning a BERT model only yields marginal (∼1%) improvement in accuracy at sentence level. Using more advanced models and hyper parameter tuning may yield better performance. However, our goal in this paper is not to build the best tagger possible for this task, rather to demonstrate the utility of this task in NLP and computational social science applications. Given the high performance of the Slot-filling model, we use it for all analyses in the rest of this paper.

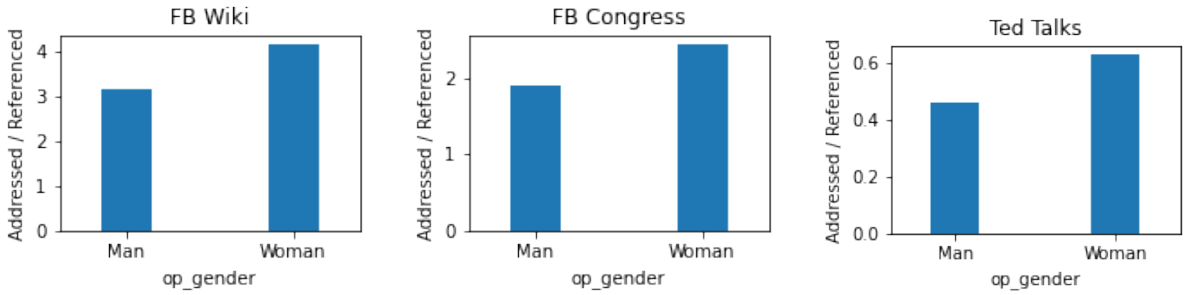

Figure 1: Gender bias in Address vs. Reference mentions (Politicians' Facebook page comments)

	Address			Reference			
Model	**P**	**R**	**F**	**P**	**R**	**F**	**Acc.**
RND	30.5	46.8	36.9	65.9	49.0	56.2	48.3
BERT Slot-filling							
S1	79.6	68.8	73.8	86.0	91.6	88.7	84.2
S2	76.3	70.8	73.4	86.5	89.5	88.0	83.5
S3	82.2	68.4	74.7	86.0	92.9	89.4	85.0
DistilBERT Slot-filling							
S1	81.5	65.2	72.4	84.8	92.9	88.7	84.0
S2	77.1	67.2	71.8	85.3	90.5	87.8	82.9
S3	83.0	66.4	73.8	85.4	93.5	89.3	84.8

Table 1: Results on predicting *address* vs. *reference* distinction using Random (RND), BERT based Slot-filling, and DistillBERT based Slot-filling approaches.

3 Gender Effects in Addressing

We first look into the RtGender dataset (Voigt et al., 2018) built to study differential responses to gender. They found that responses to female posters or speakers were more likely to be about the individuals (e.g., their appearance) rather than about the content they posted or talked about. As a complementary analysis, we analyze whether these responses were addressed to the speaker or poster, or referring to them. We apply the tagger to 5K comments each, chosen at random, from three different sub-corpora in the RtGender corpus: comments in response to (1) Facebook posts by politicians (FB Congress), (2) Facebook posts by celebrities (FB Wiki), and (3) TED talk videos (Ted Talks). We ensured that the tagger does not introduce systematic gender bias; t-test revealed no association between gender and error ($p = 0.166$).

Across board, mentions of female personalities were more likely to be in the *address* rather than *reference* contexts (Figure 1). This difference was statistically significant in all three cases: $t(4999) = 3.51$, $p < .001$ (FB Congress); $t(4999) = 3.87$, $p < .001$ (FB Wiki); and $t(4999) = 4.41$, $p < .001$ (TED Talks). For the congress dataset, we also have access to the political party they belong to; we added it as a control causing the effect size to decrease (2.72) suggesting that political party affiliation plays an important role. In fact, Figure 2 shows that the gender disparity is present only for the Republican party politicians.

Addressing someone directly could be an expression of friendliness or familiarity, and its prevalence in comments directed at female personalities is notable. These insights enable adding nuance to many NLP-aided studies of gender and power. Moreover, this finding adds to research on gender influences on communication with and about professionals (Atir and Ferguson, 2018).

4 Address vs. Reference and Toxicity

We now turn to online abuse detection, an NLP task where address vs. reference distinction is important. Prior work has shown that 2nd person pronouns are spuriously associated with toxic comments (Hede et al., 2021). In languages such as Arabic that has explicit vocative markers, researchers have used vocative markers to curate comments with higher likelihood of offensiveness (Mubarak et al., 2020). In this section, we use our tagger to analyze the tox-

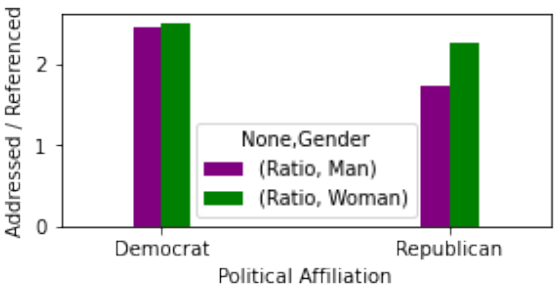

Figure 2: Address vs. Reference mentions across gender and party affiliation of the politician.

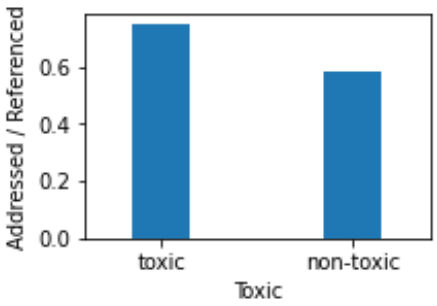

Figure 3: Ratio of names being addressed vs. referenced in the Jigsaw dataset across toxic vs. non-toxic.

icity dataset annotated by Jigsaw (Jigsaw, 2018) to see if this pattern holds true. In the Jigsaw dataset, we do not have access to the mentions of people in text. Hence, we created a tagger for the Jigsaw dataset by first using the SpaCy python package to detect person mentions, then used the BERT Slot-filling (S3) tagger to detect whether each person is addressed or referenced in the message.

We find significant difference in address vs. reference in toxic vs. non-toxic tweets. The average toxicity score of sentences with address mentions were 0.088, compared to 0.070 for those without; this difference is statistically significant using the standard Student's t-test ($p < .001$) and a permutation test ($p < .001$). Figure 3 shows differences in the ratios of address to reference mentions in toxic and non-toxic texts. This finding is important for NLP-aided content moderation, especially in detecting targets of abuse.

5 Discussion/Conclusion

In this paper, we introduced the basic NLP task of distinguishing a name mention to be *address* or *reference*, annotated a new dataset for it in the English language, and presented a simple tagger using contextual word embeddings. Our annotation and tagging experiments reveal this to be a relatively easy task, however our accuracy being only at 85% suggests room to improve. We also demonstrate the utility of this capability in computational social science work anchored on name mentions through two analyses: first, on gender bias in mention patterns, and second, in toxic comments online.

This capability is important, but often ignored, for tasks that assume entity mentions to be part of the expressed propositional meaning; e.g., belief modeling (Prabhakaran et al., 2015), and social relation extraction (Massey et al., 2015). It will also aid in tasks that model relationships between interactants, such as power (Prabhakaran and Rambow, 2014) and influence (Rosenthal and Mckeown, 2017). The vocative usage is arguably already being implicitly modeled in tasks such as dialog act tagging. However, it may be important to model it explicitly in certain cases, e.g., our work could contribute to ongoing efforts in detecting addressees in multi-party dialog (Ouchi and Tsuboi, 2016; Le et al., 2019). Future work should look into these applications, and more advanced modeling techniques such as few-shot training for this task.

6 Limitations

Our work is not without its limitations. First of all, our annotated data is relatively small. However, given the relatively straightforward task (as reflected in high IAA), and since we are using this data only for evaluations, we believe that this amount of data is sufficient for the research questions we are asking/answering in this paper. Second, our data entirely comes from the politics domain and social media, situated in the US context. This choice was driven by our downstream use case of a large scale social science analysis in the US political domain. While we have not established how well our tagger performs in domains other than politics, given that our tagger relies on contextualized language models trained on web data and since it is performing a basic linguistic task, we believe that the performance is robust across domains used in Section 3 and 4. However, we expect performance degradation with genre or dialectal shifts with substantial differences in syntactic patterns. Third, we have not fully exploited the utility of the dataset in this work. As mentioned in Section 2.2, our aim in this paper is not to build the best tagger possible, and hence we did not explore state of the art modeling techniques such as few-shot learning. Finally, our work is done entirely on English language data. While we believe that similar approach could work in other languages without vocative markers, more research need to be performed to verify that. While we acknowledge these limitations, we reiterate that these are outside the scope of what could be meaningfully done within this short paper.

7 Ethical Considerations

Like any technology, our work also has the potential for misuse. For instance, using the tagger for social science analyses in contexts where it was not

trained or tested for might result in erroneous insights. Hence, we will be releasing a data card and model card along with the publication to document the intended use cases and various analysis results. Furthermore, although we ensured our tagger do not have gender bias in error rates, it may vary across other socio-demographic groups. However, the likelihood of this is rather low since we mask the identity of the name in the slot-filling approach, and hence any biases captured by person names are avoided in our current scheme. Finally, our gender bias analysis is limited to the binary gender, as all the RtGender corpus captured only binary gender.

Acknowledgements

We thank Jacob Eisenstein, Emily Reif, Kathy Meier-Hellstern, and the anonymous reviewers for helpful feedback. We also thank our research assistants Sevi Burget-Foster and Julia Sanderson who annotated the comments in our data.

References

Stav Atir and Melissa J Ferguson. 2018. How gender determines the way we speak about professionals. *Proceedings of the National Academy of Sciences*, 115(28):7278–7283.

Jorge Baptista and Nuno Mamede. 2017. Vocatives in portuguese: Identification and processing. In *6th Symposium on Languages, Applications and Technologies (SLATE 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Roger Brown and Marguerite Ford. 1961. Address in american english. *The Journal of Abnormal and Social Psychology*, 62(2):375.

Roger Brown, Albert Gilman, et al. 1960. The pronouns of power and solidarity. *Style in language*, pages 252–281.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Eleanor Dickey. 1997. Forms of address and terms of reference. *Journal of linguistics*, 33(2):255–274.

Adam Ek, Mats Wirén, Robert Östling, Kristina N. Björkenstam, Gintarė Grigonytė, and Sofia Gustafson Capková. 2018. Identifying speakers and addressees in dialogues extracted from literary fiction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Akshat Gupta. 2022. On building spoken language understanding systems for low resourced languages. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 1–11, Seattle, Washington. Association for Computational Linguistics.

Anushree Hede, Oshin Agarwal, Linda Lu, Diana C Mutz, and Ani Nenkova. 2021. From toxicity in online comments to incivility in american news: Proceed with caution. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2620–2630.

Jigsaw. 2018. Toxic comment classification challenge. `https://www.kaggle.com/c/\jigsaw-toxic-comment-classification-challenge/data`. Accessed: 2021-05-01.

Mansooreh Karami, Ahmadreza Mosallanezhad, Michelle V Mancenido, and Huan Liu. 2020. "let's eat grandma": When punctuation matters in sentence representation for sentiment analysis. *arXiv e-prints*, pages arXiv–2101.

Ran Le, Wenpeng Hu, Mingyue Shang, Zhenjun You, Lidong Bing, Dongyan Zhao, and Rui Yan. 2019. Who is speaking to whom? learning to identify utterance addressee in multi-party conversations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1909–1919, Hong Kong, China. Association for Computational Linguistics.

Aibek Makazhanov, Denilson Barbosa, and Grzegorz Kondrak. 2014. Extracting family relationship networks from novels. *arXiv preprint arXiv:1405.0603*.

Mónica Marrero, Julián Urbano, Sonia Sánchez-Cuadrado, Jorge Morato, and Juan Miguel Gómez-Berbís. 2013. Named entity recognition: fallacies, challenges and opportunities. *Computer Standards & Interfaces*, 35(5):482–489.

Philip Massey, Patrick Xia, David Bamman, and Noah A Smith. 2015. Annotating character relationships in literary texts. *arXiv preprint arXiv:1512.00728*.

Andrea Moro. 2003. Notes on vocative case. a case study in clause structure. *Amsterdam Studies in the Theory and History of Linguistic Science Series 4*, pages 247–262.

Hamdy Mubarak, Kareem Darwish, Walid Magdy, Tamer Elsayed, and Hend Al-Khalifa. 2020. Overview of OSACT4 arabic offensive language detection shared task. In *Proceedings of the 4th Workshop on open-source arabic corpora and processing tools, with a shared task on offensive language detection*, pages 48–52.

Mutee U Rahman Muslim and Mohammad Iqbal Bhatti. 2010. Finite state morphology and sindhi noun inflections. In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, pages 669–676.

Jee-sun Nam and Key-Sun Choi. 1997. A local grammar-based approach to recognizing of proper names in korean texts. In *Fifth Workshop on Very Large Corpora*.

Harm op den Akker and Rieks op den Akker. 2009. Are you being addressed? - real-time addressee detection to support remote participants in hybrid meetings. In *Proceedings of the SIGDIAL 2009 Conference*, pages 21–28, London, UK. Association for Computational Linguistics.

Hiroki Ouchi and Yuta Tsuboi. 2016. Addressee and response selection for multi-party conversation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2133–2143, Austin, Texas. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Ashima Arora, and Owen Rambow. 2014. Staying on topic: An indicator of power in political debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1481–1486.

Vinodkumar Prabhakaran, Tomas By, Julia Hirschberg, Owen Rambow, Samira Shaikh, Tomek Strzalkowski, Jennifer Tracey, Michael Arrigo, Rupayan Basu, Micah Clark, et al. 2015. A new dataset and evaluation for belief/factuality. In *4th Joint Conference on Lexical and Computational Semantics,* SEM 2015*, pages 82–91. Association for Computational Linguistics (ACL).

Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. 2019. Perturbation sensitivity analysis to detect unintended model biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5740–5745.

Vinodkumar Prabhakaran, Ajita John, and Dorée D Seligmann. 2013. Who had the upper hand? ranking participants of interactions based on their relative power. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 365–373.

Vinodkumar Prabhakaran and Owen Rambow. 2014. Predicting power relations between participants in written dialog from a single thread. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 339–344.

Sara Rosenthal and Kathleen Mckeown. 2017. Detecting influencers in multiple online genres. *ACM Transactions on Internet Technology (TOIT)*, 17(2):1–22.

Beth M Sundheim. 1995. Overview of results of the MUC-6 evaluation. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*.

Reut Tsarfaty, Shoval Sadde, Stav Klein, and Amit Seker. 2019. What's wrong with hebrew NLP? and how to make it right. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 259–264.

Rob Voigt, David Jurgens, Vinodkumar Prabhakaran, Dan Jurafsky, and Yulia Tsvetkov. 2018. RtGender: A corpus for studying differential responses to gender. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Vikas Yadav and Steven Bethard. 2018. A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Left blank.

☑ A2. Did you discuss any potential risks of your work?
Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 2.1

☑ B1. Did you cite the creators of artifacts you used?
Section 2.2

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 7

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Section 2

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 2

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 2 and

C ☑ Did you run computational experiments?

Section 2.2

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Not applicable. We used checkpoints of pre-trained models and discussed their size and parameters (and refer to respective papers). We do not train any new models.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Not applicable. No hyper-parameter tuning was performed

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 2.2

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 2

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Section 2.1

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. We trained expert annotators on the topic based on the information presented in the paper.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. We used data published in another publication, no new data was collected from users.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. We did not perform any data collections from users

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. This will be discussed in the camera-ready, since annotators were students from a specific university.