HHAI2022: Augmenting Human Intellect
S. Schlobach et al. (Eds.)

doi:10.3233/FAIA220220

Evaluation of a Coreference Resolution Model on Mention Patterns with Common Ground

HHAI Poster Submission

Jaap KRUIJT [a,1]
[a] *Vrije Universiteit Amsterdam, the Netherlands*

Keywords. coreference resolution, common ground, Human-AI interaction, social dialogue

Abstract

Efficient communication between humans and AI requires the AI to understand references that are made during the conversation. For this, the AI needs to establish *common ground* with the human in order to make correct judgments and take the right actions [1]. Common ground is the shared information that speakers rely on during a conversation, which is built up over time as the speakers share more interactions [2]. In our research, we focus on the role of common ground in resolving third-person references in human-robot social interaction. In social dialogue, these references are often vague and context-dependent. Especially in a conversation between two well-acquainted individuals, the references to people they both know can become highly ambiguous. Through shared interactions, these references become increasingly efficient (i.e. shorter). This comes at the cost of intelligibility for outsiders who do not share the common ground [3]. For machines that have no understanding of the common ground then, these references are difficult to interpret and relate to other references. In future work, we aim to approach this problem by building a reference resolution model which utilises a knowledge-rich approach and builds up common ground with a human in an interactive setting, where the robot and the human can coordinate to form the common ground together. In preparation for this, here we first investigate the limitations of existing reference resolution models in social interaction scenarios by evaluating to what extent they utilise common ground in resolving vague references in social dialogue. Our expectation is that these models do not fare well with references that require long-distance common ground knowledge, but that providing them with the relevant background knowledge will improve performance.

Machine learning-based coreference resolution models can achieve impressive performance (e.g. [4,5]). However, most datasets used in coreference resolution tasks consist of snippets of formal text, e.g. from news articles. These datasets are not useful for

[1]Corresponding Author: Jaap Kruijt, E-mail: j.m.kruijt@vu.nl.

the analysis of complex social references for a number of reasons. First of all, most social conversation is not formal, so the language used in these datasets does not reflect the way that humans talk to each other in social dialogue. Secondly, the snippets of data usually consist of one single context or discourse, with no long-distance links to mentions in other (discourse) contexts. Lastly, the data does not contain a temporal aspect, and as a result it does not reflect the gradual emergence of more ambiguous or context-specific references as common ground builds up between conversation partners.

To test existing machine performance on social dialogue which does contain long-distance links and a temporal relation between discourse contexts, we took an existing dataset consisting of episodes of the *Friends* TV show used for a combined entity linking and coreference resolution task [6] and modified its structure to use it for a coreference resolution task specifically aimed at resolving third-person reference chains. In this show the characters have a shared social history which grows throughout the series. Discourse context and social common ground are key in the resolution of these references. While some references, such as pronouns, can be solved within the discourse context, other references rely on background information and common ground which was established in previous episodes. Once this common ground knowledge is established, individuals which are in the shared knowledge base can more readily be referenced in social conversation, using vague or ambiguous references, whereas individuals which are not in the knowledge base need to be referenced first using an explicit reference before they can be referenced with ambiguous references such as pronouns for the duration of the discourse. The mention patterns and coreference chains for a character will thus be different depending on the level of shared interactions that this character was mentioned in or a part of.

We divide the entities mentioned by third person references in our dataset into 'inner circle' and 'outer circle' based on a set of rules that determine whether a character is of importance to the characters and the story throughout the series. For each episode, we calculate the ratio of inner circle/outer circle mentions. We then test the current state-of-the-art end-to-end coreference resolution model by [5] as implemented by [7] on two episodes with a 1/1 ratio and two with a 4/1 ratio of inner circle and outer circle mentions (a 4/1 ratio means that the episode has 4 times as many references to people in the inner circle as references to people in the outer circle). We hypothesize that the pre-trained model, which has not been finetuned on this dataset, will perform worse on the episode with a 4/1 ratio than on the episode with a 1/1 ratio, since it does not have the required common ground to resolve the inner circle mentions. Afterwards, we finetune the model on preceding conversations. We then expect the model performance on the 4/1 ratio episode to improve significantly more than the performance on the 1/1 ratio episode due to the increase in background knowledge on the inner circle mentions. Crucially, we only finetune the model on the episodes that chronologically precede the test episode. Since we aim to simulate the natural way in which humans establish social common ground, the system should not have any knowledge of individuals and events presented later in time.

Our results show some indications that the chosen model does have more trouble with the vague inner circle references, but also that the model did not learn common ground knowledge through training. Thus, a more knowledge-rich approach to resolving references is desirable for social dialogue. Our future work will address this in interactive real-world scenarios, with the buildup of common ground as key to our approach.

References

[1] Chai JY, She L, Fang R, Ottarson S, Littley C, Liu C, et al. Collaborative Effort towards Common Ground in Situated Human-Robot Dialogue. In: Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction. HRI '14. New York, NY, USA: Association for Computing Machinery; 2014. p. 33–40. Available from: https://doi.org/10.1145/2559636.2559677.

[2] Stalnaker R. Common ground. Linguistics and philosophy. 2002;25(5/6):701-21.

[3] Hawkins RD, Franke M, Frank MC, Goldberg AE, Smith K, Griffiths TL, et al.. From partners to populations: A hierarchical Bayesian account of coordination and convention. arXiv; 2021. Available from: https://arxiv.org/abs/2104.05857.

[4] Joshi M, Levy O, Weld DS, Zettlemoyer L. BERT for Coreference Resolution: Baselines and Analysis. In: Empirical Methods in Natural Language Processing (EMNLP); 2019. .

[5] Joshi M, Chen D, Liu Y, Weld DS, Zettlemoyer L, Levy O. SpanBERT: Improving Pre-training by Representing and Predicting Spans. Transactions of the Association for Computational Linguistics. 2020 01;8:64-77. Available from: https://doi.org/10.1162/tacl_a_00300.

[6] Choi JD, Chen HY. SemEval 2018 Task 4: Character Identification on Multiparty Dialogues. In: Proceedings of The 12th International Workshop on Semantic Evaluation. New Orleans, Louisiana: Association for Computational Linguistics; 2018. p. 57-64. Available from: https://aclanthology.org/S18-1007.

[7] Xu L, Choi JD. Revealing the Myth of Higher-Order Inference in Coreference Resolution. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics; 2020. p. 8527-33. Available from: https://www.aclweb.org/anthology/2020.emnlp-main.686.