"Alexa, I don't know who she is": Discourse and Knowledge Driven Coreference Resolution

Angela Ramirez, Cecilia Li, Eduardo Zamora, Phillip Lee, Jeshwanth Bheemanpally, Marilyn Walker, and Adwait Ratnaparkhi
University of California Santa Cruz, Santa Cruz, CA
aramir62, yli331,ezamora9,plee35,jbheeman,
mawalker, adratnap@ucsc.edu

Abstract

Coreference resolution is the task of identifying mentions that refer to the same entity in a given text (Stylianou and Vlahavas, 2019). It is also a key part of improving interpretation in a discourse for spoken dialog (Stent and Bangalore, 2010). We curate a dataset that captures coreference based on conversations with Athena, a bot participating in the Amazon Alexa Prize Social Bot Grand Challenge 4, and users that prompt into the challenge. Then we use this dataset to fine-tune a model and compare against the not fine-tuned version. The performance of the models are compared across a set of pronominal types: it, they, that, she, he, her, him, his, and hers, and how they perform across the topics of Movies, Sports, Television, and Music. Our findings suggests that our model fine-tuned on curated dataset outperforms the out-of-the-box AllenAI model (Lee et al., 2018) that uses SpanBERT embeddings (Joshi et al., 2020) on all types and topics.

1 Introduction

Amazon hosts an open domain dialogue social bot competition aimed to have universities competing to engineer a bot that can interact with a user for longer than 20 minutes (Ram et al., 2018). The University of California, Santa Cruz has competed for the past five years, and in the past two years, with our system known as Athena (Harrison et al., 2020). In a discourse, a key part of the interpretation process for advanced spoken dialog requires a coreference module to be present in applications (Stent and Bangalore, 2010), so to further improve Athena's ability to interpret the user we focus on resolving coreference in a user's dialogue.

The task of coreference resolution in our case involves resolving entities that are referred by the user using ambiguous terms. For instance, the system could say "I love Adam Driver in Star War's a New Hope!", and the user could respond "oh I love him", where in this case we have a pronoun "him" to resolve. This would be an example of pronominal anaphora (Carnie, 2002), which is considered to be one of the most common and prevalent types to occur in day-to-day speech (Sukthanker et al., 2020). It can also, be described as when an antecedent is being referenced prior to the ambiguous term (King and Lewis, 2018).

To capture this case and other types of anaphora, we built a dataset based off of user interactions between Athena and users. Our data pertains to the topics Movies, Music, TV, and Sports because we use knowledge-base response generators that know that the entities are being referenced. Therefore, the system is aware of the entity it has brought up, and we can store that information. To further narrow our scope, we segment our data to focus in on a set of pronoun types: it, they, that, she, him, her, his, hers, and he. An analysis of our dataset found that less than 3 percent did not have cases that required considering more than 1 turn of context out of the total 5 turns that were pulled, so we only trained and tested on 1 turn to further simplify our scope.

This curated dataset was used to fine-tune the out-of-the-box AllenAI model (Lee et al., 2018) that uses SpanBERT embeddings (Joshi et al., 2020), and we compare it's performance on the unfine-tuned version that was solely trained on CoNLL 2012 (Pradhan et al., 2012).

Topic/Pronoun	Movies	Music	TV	Sports	he	her	hers	him	his	it	she	that	they
Counts	3056	2044	1471	1317	886	190	1	300	62	3122	292	2651	384

Table 1: Topic and Pronoun Distribution

2 Dataset

From conversations [1] between our system and users of the Amazon Alexa Social Bot, we collected the examples where a pronoun was present. For each sample, we gathered the following information for our task: the unique conversation session ID, entities mentioned by the system and user, the last system utterance (at timestamp $t-1$), the user response utterance (at timestamp t), the knowledge graph topic used in this conversation (Movies, Music, TV, Sports), and up to five past utterances. We were able to annotate coreference clusters for 7888 data samples, where we refer to Table 1 for the topic and pronoun distribution. Our annotators focused on identifying spans for entity types (actor, film, television show, athletes, etc.) that could be discussed by the system and prolong the conversation. For example, given the system utterance in Figure 1, we have the gold label as **it** referring to **Scandal** and having the entity type of television show. [1]

The annotators, who consists of subject matter experts as well as the researchers involved in this project, each annotated portions of the data across different topics without overlaps. To ensure consistency, we also required the reasoning behind each annotator's choices, which were used for our multiple quality assessments.

Speaker	**Utterance**	**Entities**	**Entity Types**
SYSTEM	**Scandal** is considered both a **television drama and a thriller television series**. what's your opinion of those genres?	Scandal, television drama and a thriller television series	television show, genre
USER	is **it** on netflix	it: Scandal	television show

Figure 1: Conversation with Pronominal Resolution

Lastly, for 10-fold cross-validation, 10% of our data is set aside as our development set. The training data is converted to CoNLL format with information in the coreference resolution column and speaker column (other columns are set to default). To ensure adequate testing data per fold, we produce two sets of 5 folds with different random seeds.

[1] The data used in this task will not be released due to privacy reasons.

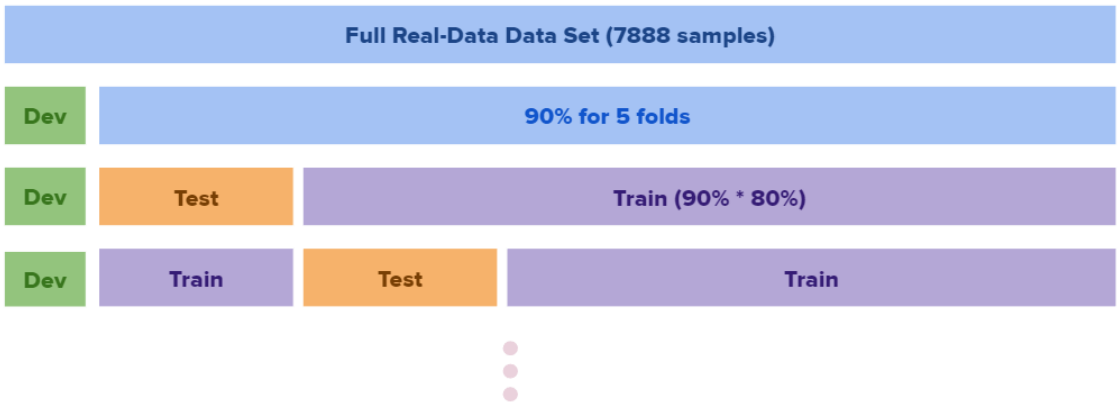

Figure 2: Dataset splits.

3 Design and Experiments

The discourse model utilizes the state table to pass and receive information from the rest of the system. Entities are stored based on their importance or saliency. For the Athena system, entities such as songs, bands, and athletes are tracked. The entities fall within four different topics: music, television, sports and movies. Abstract entity types are not included in the discourse model. We also consider the amount of dialogue context to include when tracking entities. Therefore, the discourse model is a function of entity types and context: $D(t, n)$. With T being the set of all entity types found within Wikidata, $t \subset T$ where t is in the Athena gazetteers. $1 \leq n \leq N$ where N is total number of conversation turns.

The neural model utilizes the AllenNLP Coreference Resolution model (Gardner et al., 2018) which consists of multiple LSTM layers and antecedent span pruning (Lee et al., 2018). Their current off-the-shelf model uses SpanBERT embeddings (Joshi et al., 2020) rather than the GloVe embeddings (Lee et al., 2017) mentioned in the paper. The model is trained on the CoNLL 2012 (Pradhan et al., 2012) English data, and is used for our baseline comparisons. In our experiments, this model is also fine-tuned and evaluated on the annotated data collected from the Athena conversation logs as well as synthetically generated datasets using 10-fold validation. The neural model can use a variable amount of context in the input, or just the raw utterances. In our case of using a context where $n = 1$, the input for our baseline testing only contains a concatenation of the last system utterance with the current user utterance. The target label is the corresponding antecedent for a given pronoun. This label is extracted from the coreference cluster predicted around the pronoun.

4 Results and Conclusion

Each model is evaluated using F1, Precision, and Recall scores. The predicted antecedent is compared to the gold antecedent for the detected pronoun, with an exact match.

In table 2, we find that our fine-tuned model outperforms the baseline model. Overall, our fine-tuned model does better despite pronoun and topic. From performing a paired t-test, we find that our results between the baseline and fine-tuned model are extremely significant with p-values less than 0.0001, so we reject the null hypotheses.

We hypothesize that the baseline model's poor performance could be due to the data it was trained on being less conversational and containing cases requiring more context. To test this further, analysis would need to be completed to test the robustness of our model and the baseline against other conversational datasets that contain different styles and structures. In addition, for future work the data set annotation process could be expanded to potentially harness more information like the type of anaphora occurring to better handle edge cases to increase our F1 scores. Our overall take away is that a curated dataset based on a bot's conversation with users can produce a fine-tuned model that outperforms a baseline that is state-of-the art on known datasets.

Model	Pronoun/Topic	all	it	that	they	he/she	his/her	movies	music	sports	tv
AllenNLP Baseline	F1	47.55	47.42	15.71	46.34	68.35	51.52	47.14	38.37	67.19	33.75
	Precision	41.17	39.31	20.63	35.33	56.32	38.82	41.37	33.21	61.75	27.63
	Recall	55.52	59.90	12.72	68.60	87.09	77.03	54.9	45.53	73.88	43.68
AllenNLP Fine-Tuned	F1	**78.55**	**75.16**	**66.13**	**79.14**	**88.76**	**90.18**	**83.43**	**74.96**	**85.79**	**75.83**
	Precision	91.24	88.18	92.25	92.57	93.08	94.46	92.70	93.32	95.88	83.91
	Recall	68.99	65.88	51.78	64.66	85.08	86.35	75.85	62.93	64.15	69.50

Table 2: Average of Ten Folds for Each Model

References

Andrew Carnie. 2002. *Syntax: a generative introduction*. Blackwell.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. Allennlp: A deep semantic natural language processing platform.

Vrindavan Harrison, Juraj Juraska, Wen Cui, Lena Reed, Kevin K. Bowden, JiaQi Wu, Brian Schwarzmann, Abteen Ebrahimi, Rishi Rajasekaran, Nikhil Varghese, Max Wechsler-Azen, Steve Whittaker, Jeffrey Flanigan, and Marilyn A. Walker. 2020. Athena: Constructing dialogues dynamically with discourse constraints. *CoRR*, abs/2011.10683.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans.

Jeffrey C. King and Karen S. Lewis. 2018. Anaphora. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2018 edition.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark, September. Association for Computational Linguistics.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. *CoRR*, abs/1804.05392.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea, July. Association for Computational Linguistics.

Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, Eric King, Kate Bland, Amanda Wartick, Yi Pan, Han Song, Sk Jayadevan, Gene Hwang, and Art Pettigrue. 2018. Conversational AI: the science behind the alexa prize. *CoRR*, abs/1801.03604.

Amanda J. Stent and Srinivas Bangalore. 2010. Interaction between dialog structure and coreference resolution. In *2010 IEEE Spoken Language Technology Workshop*, pages 342–347.

Nikolaos Stylianou and Ioannis P. Vlahavas. 2019. A neural entity coreference resolution review. *CoRR*, abs/1910.09329.

Rhea Sukthanker, Soujanya Poria, Erik Cambria, and Ramkumar Thirunavukarasu. 2020. Anaphora and coreference resolution: A review. *Information Fusion*, 59:139–162.