Are Word Embedding and Dialogue Act Class-based Features Useful for Coreference Resolution in Dialogue?

Samarth Agrawal[*], **Aditya Joshi**[*†], **Joe Cheri Ross**[*], **Pushpak Bhattacharyya**[*] and **Harshawardhan M. Wabgaonkar**[§]
[*] *Indian Institute of Technology Bombay, India*
[†] *IITB-Monash Research Academy, India*
[§] *Accenture Tech Labs, Bangalore, India*
Email: {samartha, adityaj, joe, pb}@cse.iitb.ac.in, h.wabgaonkar@accenture.com

***Abstract*—Due to the rise in the popularity of chatbots, there is a need to revisit the past work in coreference resolution in dialogue. Dialogues pose unique challenges to coreference resolution. This paper introduces a novel set of features based on word embeddings and dialogue act classes for the task. We show that our system with these novel features gives an improvement 24.8% in F-score over previous work on the same dataset. Additionally, we also evaluate our system using the CoNLL metrics and report the best CoNLL score of 75.93. This paper establishes the importance of these features for coreference resolution in dialogues and points to further work that can be done for the task.**

***Keywords*-Coreference resolution; Dialogue; Word embeddings; Dialogue act classes;**

I. Introduction

Coreference resolution is the task of finding entities that refer to each other in a discourse. Due to additional challenges in doing so for a dialogue, approaches to coreference resolution in dialogue have been reported [1] [2]. With the rise of chatbot deployments, we believe that it is critical to look back at this crucial problem. This is because conversational agents or chatbots need to resolve coreferences in order to sustain a natural conversation with a user. Consider the following conversation with Google Assistant[1]:

Human: Who was Alan Turing ?
Bot: Here's his profile. (Results)
Human: Where was he born ?
Bot: He was born in Maida Vale on June 23,1912
Human: Where is that ?
Bot: Heres the top search result. (Results for Alan Turing)

In the conversation above, the chatbot resolves 'he' correctly as 'Alan Turing', but fails to resolve 'that' as 'Maida Vale'. Similar outputs are also obtained from Siri[2] and Cortana[3], at the time of writing this paper. Therefore, better coreference resolution in dialogue would improve the user experience, thereby increasing the adoption of such chatbots.

A dialogue is a discourse where multiple speakers take turns to participate in a conversation. A dialogue typically exhibits following properties:

1) It contains dis-fluencies like pauses and filler words like 'uh', 'um', etc.
2) It may contain noisy text and grammatical errors. The noise may also be in the form of sentence fragments. For example[4],
   A: I, uh, my sister has a, she just had a baby.
   A: He's about five months old'
3) It may have personal and demonstrative pronouns with non-noun phrase antecedents or no antecedent [1]. For example in the dialogue,
   A: I work off and on just temporarily and usually find friends to babysit
   B: I don't envy anybody who's in that situation to find day care'
   the phrase '*that situation*' refers to '*work on and off*' which is a non-NP antecedent.

These properties elucidate that semantics play an important role in coreference resolution in spoken dialogue. Word embeddings and dialogue acts have been shown to be useful techniques to incorporate such semantic information for other NLP tasks. Therefore, in this paper, we address the question:

Can state-of-the-art for coreference resolution in dialogues be enhanced with the help of features based on word embeddings and dialogue acts?

II. Related work

A detailed analysis of spoken and written text corpora done by [3] reveals that different coreference strategies are required for coreference resolution in spoken and written

[1]https://assistant.google.com/
[2]https://www.apple.com/ios/siri/
[3]https://www.microsoft.com/en-us/windows/cortana

[4]The examples in this section are from the dataset used

Basic features	
distance_words, distance_sent, zero_distance_sent	Distance between mentions
same_words	Levenshtein distance between mentions
len_ante, len_ana	#words in mentions
defNP_ante, defNP_ana	Definite or indefinite Noun Phrase
cat_ante, cat_ana	Constituency parse category
tfidf_ante	Average TF-IDF of antecedent
imp_ante, imp_ana	Average of inverse document frequency
isSubj_ante, isSubj_ana	Syntactic function of mention in Dependency parse
type_ana	Type of pronoun
is_num_agree	Numeric agreement between mentions
gender_agree	Gender agreement between mentions
Semantic features	
semantic_agree	Semantic class agreement
semantic_distance_ment, semantic_distance_sent	Cosine similarity between average word vectors
Dialog-specific features	
is_same_speaker	Speaker agreement
dialAct_ante, dialAct_ana	Dialog act class of the utterances

Table I
FEATURE SETS. HERE 'ANTE' IS SHORT FOR ANTECEDENT AND 'ANA' IS SHORT FOR ANAPHORA

texts. Coreference resolution for dialogues has been reported for different domains. [4] present a method to resolve demonstrative pronouns, giving emphasis to resolution of pronouns referring to abstract entities on TRAINS corpus. On the other hand, the impact of prosodic features for coreference resolution is analyzed by [5] on spoken dialogue in German. [6] discusses a coreference resolution system for a virtual patient dialogue system. In addition, multimodal approaches like [7] [8] use hand gestures correlated with coreference to resolve anaphoric references. The work closest to ours is by [1] who investigate different spoken dialogue specific features for coreference resolution on Switchboard corpus. We use their features as basic features. However, the feature values in their case are determined manually while we use lexical resources to do so.

III. FEATURES

Table I shows three types of features used for our experiments. The basic features have been used in past work, while the next two sets have been introduced.

A. Basic features

These features are based on past work in coreference resolution in dialogues by [1]. They obtain the values for is_num_agree and $gender\_agree$ features from human annotators. This approach is not scalable in case of large datasets or in real-world applications like chatbots. We instead use a corpus with word frequencies of words and named entities [9]. The corpus has frequency information of nouns for gender and number (singular or plural). We extract the head word of the mention and use the class with maximum frequency of the head word for gender and number agreement features.

B. Semantic features

The first set of features that we introduce are semantic features. The idea here is to harness the semantic similarity between corefering mentions and also the similarity between the utterance in which they occur. For $semantic\_agree$ feature, we define five semantic classes: person, group, location, object and entity. Mentions are classified according to what class is the nearest to the head word in the WordNet hierarchy [10]. The feature is set if both mentions have the same semantic class or there is a compatible pronoun in case of a pronominal mention. For $semantic\_distance\_ment$ feature, we take average of word vectors of all words in a mention and then take cosine similarity between them. The feature $semantic\_distance\_sent$ works similarly by averaging the word vectors for the entire utterance in which the mention occurs. Word embedding based similarity features have recently been used by [11] for coreference resolution on general text documents.

	Pairwise			B^3			CEAFe			MUC			
Model	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**CoNLL**
Strube	56.74	40.72	47.42	-	-	-	-	-	-	-	-	-	-
CORT	-	-	-	13.30	84.21	22.97	64.90	8.86	15.60	67.23	86.10	75.50	38.02
C4.5	49.80	73.37	59.33	21.98	85.78	34.99	60.51	13.94	22.65	68.58	88.85	77.41	45.02
SVM(linear)	57.16	68.14	62.17	63.21	62.40	62.80	71.85	47.52	57.21	72.39	67.18	69.69	63.23
SVM(rbf)	65.01	68.44	66.68	74.64	68.14	71.24	78.65	61.13	68.79	81.17	73.54	77.16	72.40
NN	63.71	77.74	70.03	75.21	71.61	73.36	81.84	63.03	71.21	83.20	77.07	80.02	74.86
RF	67.12	78.13	**72.21**	78.03	71.62	74.69	81.86	64.23	71.98	85.09	77.50	81.11	**75.93**

Table II
PERFORMANCE OF CLASSIFIERS WITH ALL FEATURES COMPARED WITH BASELINE OF [1] AND CORT TOOL; SOME VALUES ARE NOT REPORTED BECAUSE THEY ARE NOT AVAILABLE FOR PAST WORK; LAST COLUMN INDICATES AVERAGE OF F-SCORES OF B^3, CEAFE AND MUC

C. Dialogue-specific features

The next set of features are based on speaker and dialogue act classes. $is\_same\_speaker$ is a mention-pair level feature that encodes whether both mentions in a pair come from the same or different speakers.

Consider the following example:

A: Well is [that] a good indicator? (Question)

B: That well [it] can be. (Answer)

Here question - answer relationship between utterances can be harnessed for resolution of 'it' with 'that'. This motivates the need to add dialogue act class based features for our task. The features $dialact\_ante$, $dialact\_ana$ encode the dialogue act class of the utterance in which the mention occurs. This class can take one of 5 values: '*Statement*', '*Opinion*', '*Question*', '*Answer*', and '*Other*'. Since our dataset contains 43 dialogue act classes, we map them to these 5 coarse classes. 'Statement' and 'Opinion' classes are obtained as is from the original dataset while fine classes for questions and answers are clubbed into 'Question' and 'Answer' respectively. Remaining classes are clubbed into 'Others'.

IV. EXPERIMENT SETUP

We use the Switchboard corpus [12] for our experiments. Switchboard is a long-standing corpus consisting of telephone conversations, consisting of 642 conversations between speakers of American English. We use NXT Switchboard annotations which have coreference annotations for 147 dialogues [13]. Each dialogue contains approximately 200 utterances on an average. In addition to coreference information, we also use the annotation to get dialogue act classes and part of speech tags. We use this corpus for our task as it is a large general domain dataset with coreference annotations available for spoken dialogues.

We take all mentions that are part of some coreference chain in a dialogue and create all possible valid mention pairs. We use the gold mention boundaries in this task. A mention pair is deemed valid if the anaphora comes after the antecedent and the number of utterances between the two mentions is less than 10. The value of 10 is experimentally determined based on experimentation over a range of values.

This results in 210,000 valid mention pairs out of which 32,000 are coreferent. We then generate the 28 features and the binary target for the valid mention pairs. The dataset is split into 100 dialogues for training and rest 47 for testing purpose. Since there is a skew in the dataset i.e. positive cases consist only 15% of the total data, we undersample the negative cases during training. For the semantic features, we use Google word vectors [14]. Since these word vectors are trained on news articles, it is a good match for the Switchboard corpus which contains general domain dialogues. For the dialogue act features, as stated earlier, we map 43 classes to 5 dialogue act classes. The following classifiers are used in experiments:

1) Random Forest (RF) with Gini impurity as splitting criteria and 5000 trees
2) Neural networks (NN) with single hidden layer of 20 neurons. Sigmoid function was used in hidden layer and softmax was used in the output layer
3) SVM with linear and rbf kernels
4) C4.5 with entropy as splitting criteria

The classification step is followed by best-first clustering to get coreference chains from the predictions [15]. We report precision, recall and F-score for standard CoNLL metrics like B^3, CEAFe and MUC. These are based on past work by [16]. As **baselines**, we consider two systems:

1) Reported values in [1]
2) CORT system [17] trained and tested on the same dataset

V. RESULTS

Table II gives the results of our models in terms of pairwise and CoNLL (B^3, CEAFe, MUC) metrics. The last column indicates average F-score, as per convention. We observe that the best performance is obtained in case of Random Forest (CoNLL score of 75.93). Random Forest

Feature sets	C4.5	SVM	NN	RF
Basic	42.26	68.00	72.34	72.84
Basic +semantic	45.49	72.73	74.58	75.15
Basic +semantic +dialogue	45.02	72.40	74.86	**75.93**

Table III
CONLL SCORES BY FEATURE SETS

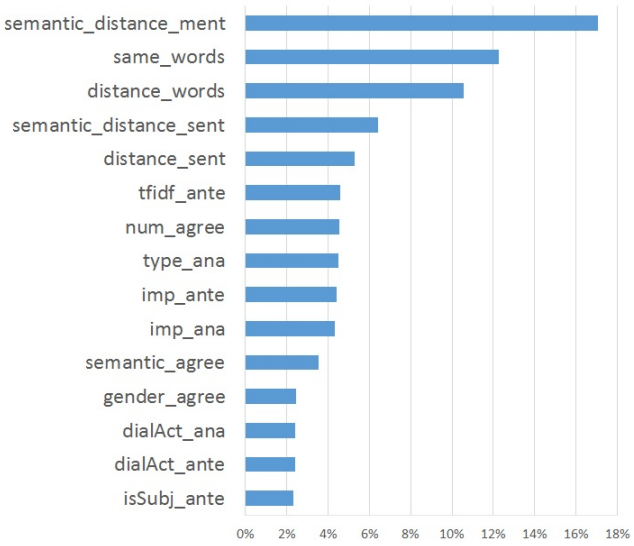

Figure 1. Feature importances of Top 15 features

also has the highest mention-pair F score of 72.2 which is a huge improvement over 47.42 reported by [1].

To understand the benefit of our three classes of features, we show results for the combinations in Table III. We obtain 72.84 CoNLL score when we run RF on basic features. Adding semantic features gives an improvement of 2.31%. Further addition of dialogue act-based features gives an improvement of 0.78%. Figure 1 lists the top 15 features as estimated by the random forest learner. We observe that **our word embedding and dialogue act based features appear** in this list. The total feature importance of semantic features is 27% and that of dialogue-specific features is 5.1%.

VI. DISCUSSION

The high importance of word embeddings and dialogue act class based features along with significant increase in performance as reported in Table III shows that they play a very important role in coreference resolution in dialogues. In addition to the reported results in the Tables II and III, we also run our classifiers after removing features with low importance and observe a decline in performance.

We observe that semantic features help when the meaning of mentions in a pair is important for resolution, as in the example:
A: I have [four kids].
B: Well, I have [four little boys].

Dialogue-specific features are helpful in cases like:

1) where speaker information is important. For example, in the following dialogue, 'we' and 'us' are resolved because they are from the same speaker:
A: uh, [we]'ve been married for ten years
B: so and it's worked out pretty well.
A: It's uh it's helped [us] ...

2) where question-answer relationship is important, as discussed in the following example in Section III:
A: Well is [that] a good indicator? (Question)
B: That well [it] can be. (Answer)

Table IV shows the distribution of all possible pairs dialog act classes for all coreferent mention pairs in our dataset. For example, out of all anaphoric mentions in 'Statement' class, 84% have antecedent from 'Statement', 6.4% have antecedents in 'Opinion' and so on. This shows that some combination of dialog act classes are more likely than others, which motivates the need for dialog act class based features for coreference resolution.

	Antecedent				
Anaphora	Answer	Opinion	Other	Question	Statement
Answer	10.9	37.3	11.5	13.6	26.7
Opinion	01.0	67.9	05.0	03.8	22.3
Other	00.9	15.5	24.6	08.9	50.1
Question	01.1	11.1	07.5	24.2	56.1
Statement	00.5	06.4	05.1	04.1	84.0

Table IV
PERCENTAGE LIKELIHOOD OF ANTECEDENT FROM A DIALOGUE ACT CLASS GIVEN THE CLASS OF ANAPHORA

VII. ERROR ANALYSIS

An analysis of errors shows the following types of errors:

A. Incorrect head word

The system is unable to extract the correct head word in some cases which leads to incorrect gender and number agreement feature values. For example, in the dialogue
A: [The average person out on the farm], at least now [they] have tractors ...,
the head word of antecedent is selected as 'farm' instead of 'person'. As 'farm' is not semantically compatible with 'they' our system returns this pair as non corefering.

B. Named Entities

The system has data for gender and numerical attributes of named entities but is unable to assign semantic classes to them. This is observed in the case of
A: [Lexus] is a Toyota sub-brand.
B: [It]'s kind of their ...

C. Nested dialogues

A speaker might quote a dialogue of a third person. This creates outlier cases like the following example which are incompatible in our model.
A: there was [a woman]
B: she said [my] best friends are lawyers ...

VIII. Conclusion & Future Work

Coreference resolution in dialogues is challenging due to a variety of reasons. In this paper, we presented two novel sets of features for coreference resolution in dialogues. These features are based on word embeddings and dialogue act classes. We evaluate our results on the Switchboard corpus. Our features result in an improvement of 24.79% over previous reported work. We show that when augmented with word embedding and dialogue act class-based features, the basic set of features demonstrate an improvement. Our word embedding and dialogue act class-based features also turn out to be among the most important features.

Our error analysis points to several directions for future work. This includes correct resolution of named entities, nested dialogues, etc. This paper sets up the promise of word embedding and dialogue act-based features for coreference resolution in dialogues.

References

[1] M. Strube and C. Müller, "A machine learning approach to pronoun resolution in spoken dialogue," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*. Association for Computational Linguistics, 2003, pp. 168–175.

[2] D. K. Byron, "Resolving pronominal reference to abstract entities," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2002, pp. 80–87.

[3] M. Amoia, K. Kunz, and E. Lapshinova-Koltunski, "Coreference in spoken vs. written texts: a corpus-based analysis." in *LREC*, 2012, pp. 158–164.

[4] D. K. Byron and J. F. Allen, "Resolving demonstrative anaphora in the trains93 corpus," in *Proceedings of DAARC2Discourse, Anaphora and Reference Resolution Colloquium*, 1998.

[5] I. Rösiger and A. Riester, "Using prosodic annotations to improve coreference resolution of spoken text." in *ACL (2)*, 2015, pp. 83–88.

[6] C.-J. Lin, C.-W. Pao, Y.-H. Chen, C.-T. Liu, and H.-H. Hsu, "Ellipsis and coreference resolution in a computerized virtual patient dialogue system," *Journal of medical systems*, vol. 40, no. 9, p. 206, 2016.

[7] J. Eisenstein and R. Davis, "Gesture improves coreference resolution," in *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*. Association for Computational Linguistics, 2006, pp. 37–40.

[8] L. Chen, A. Wang, and B. Di Eugenio, "Improving pronominal and deictic co-reference resolution with multi-modal features," in *Proceedings of the SIGDIAL 2011 Conference*. Association for Computational Linguistics, 2011, pp. 307–311.

[9] S. Bergsma and D. Lin, "Bootstrapping path-based pronoun resolution," in *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. Sydney, Australia: Association for Computational Linguistics, July 2006, pp. 33–40.

[10] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[11] H. Lee, M. Surdeanu, and D. Jurafsky, "A scaffolding approach to coreference resolution integrating statistical and rule-based models," *Natural Language Engineering*, pp. 1–30, 2017.

[12] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, vol. 1. IEEE, 1992, pp. 517–520.

[13] S. Calhoun, J. Carletta, J. M. Brenier, N. Mayo, D. Jurafsky, M. Steedman, and D. Beaver, "The nxt-format switchboard corpus: A rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue," *Language resources and evaluation*, vol. 44, no. 4, pp. 387–419, 2010.

[14] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[15] V. Ng and C. Cardie, "Improving machine learning approaches to coreference resolution," in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002, pp. 104–111.

[16] J. Cai and M. Strube, "Evaluation metrics for end-to-end coreference resolution systems," in *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, 2010, pp. 28–36.

[17] S. Martschat, P. Claus, and M. Strube, "Plug latent structures and play coreference resolution." in *ACL (System Demonstrations)*, 2015, pp. 61–66.