Anaphora Resolution in Dialogue

System Description – CODI-CRAC 2022 Shared Task

Tatiana Anikina[1], Natalia Skachkova[1],
Joseph Renner[2], Priyansh Trivedi[2]

[1]DFKI / Saarland Informatics Campus, Saarbrücken, Germany

[2]INRIA, Nancy, France

Introduction

Introduction

Our team (DFKI/INRIA) participated in all three tracks of the CODI-CRAC 2022 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue.

- For the **Anaphora track** we combined the outputs of the Workspace Coreference System [Anikina et al., 2021] and the *coref-hoi* [Xu and Choi, 2020] model.
- For the **Discourse Deixis track** we implemented a multi-task learning system that combines the learning objective of *coref-hoi* with the anaphor type classification.
- For the **Bridging track** we trained a model that is based on the coreference architecture introduced in [Joshi et al., 2019].

Identity Anaphora

Identity Anaphora

Resolving identity anaphora means finding mentions that refer to the same entity (e.g., *an apple* and *it*):

[John]$_i$ took [an apple]$_j$ and gave [it]$_j$ to [Sarah]$_k$.

Dialogue data make anaphora resolution challenging due to switching speakers, deictic references and various disfluencies:

- *Did [you]$_i$ take a ... [I]$_k$ mean the- [the apple]$_j$?*

- *Yes, [I]$_i$ did.*

Identity Anaphora

Workspace Coreference System (WCS) incrementally clusters mentions based on embedding-based semantic similarity.

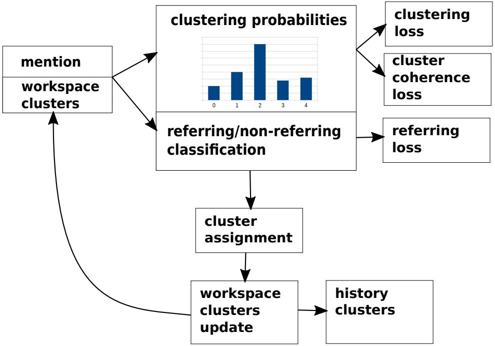
WCS uses SpaCy for mention detection and a combination of BERT [Devlin et al., 2018], GloVe [Pennington et al., 2014], Numberbatch [Speer et al., 2017] and feature-based embeddings (e.g., for animacy and speaker) to represent mention spans.

Each type of embeddings is processed by the neural network through a separate set of layers. WCS outputs a clustering score for each mention-cluster pair.

Identity Anaphora

WCS combines three different loss functions:

- **clustering loss** measures an overlap between the gold cluster and the assigned one;
- **cluster coherence loss** checks how many mentions in the assigned cluster belong to the same cluster in the gold data;
- **referring loss** checks whether mention is a referring expression;



Identity Anaphora

The motivation behind WCS is an incremental processing of mentions. However, WCS has two important limitations:

- It relies on the mention extraction by SpaCy;
- It does not support long distance coreference because if some cluster has not been updated in 100 steps it is removed from the workspace and stored in the history;

Hence, we decided to experiment with a combination of WCS and other coreference models.

Identity Anaphora

Setting	Light	AMI	Persuasion	Swbd.
Vanilla WCS	65.96	46.04	59.54	50.63
WCS + CCS	67.27	46.68	63.46	53.92
WCS + CCS + filter	67.46	46.70	63.51	54.07
WCS + coref-hoi	**72.06**	**51.41**	**69.87**	**60.61**

Table 1: Evaluation of WCS in combination with other models on the CODI-CRAC test set. CCS refers to Crosslingual Coreference System based on AllenNLP and SpaCy pipelines. Filter means the incompatibility check (e.g., number agreement between the mentions).

Identity Anaphora

We trained a “cluster merging” variant of the *coref-hoi* model. The model was developed using the CoNLL 2012 dataset without singletons, hence it does not output singleton clusters.

We evaluated *coref-hoi* and WCS on the CODI-CRAC data (dev set):

Setting	Light	Light NS	AMI	AMI NS	Persuasion	Persuasion NS
WCS	**65.39**	61.48	**43.33**	35.85	**61.23**	56.55
coref-hoi	59.84	**76.89**	43.30	**54.70**	60.60	**81.00**

Table 2: NS (No Singletons) refers to annotations without singleton clusters. Scores represent CoNLL F1 scores.

The results show that WCS performs better with singletons included and the opposite trend is observed for *coref-hoi*.

Identity Anaphora

We combined the strengths of the two systems by adding the singleton predictions of WCS to the cluster predictions of *coref-hoi*. This resulted in the highest test set scores:

	Light	AMI	Persuasion	Switchboard
Winner	82.23	62.90	79.20	75.81
Ours	72.06	51.41	69.87	60.61
Baseline	54.23	34.14	53.16	49.30

Table 3: Evaluation of the combined approach (coref-hoi and WCS) on the CODI-CRAC test set (CoNLL F1 scores).

Identity Anaphora

We also analyzed the output of the combined model and found the following types of mistakes:

- Split mention spans: *“half”* and *“hour”* in *“see you in half and hour”*;
- Overextended mention spans: *“Of course, good Monk”* or *“this realm, stories, population”*;
- Semantic mismatches: e.g., *“some”* and *“they”* in *“**Some** don’t give the money out like they are suppose to. Did you heard that **they** now do every payment taken from people transparent?”*;

Identity Anaphora

We would like to perform a more **fine-grained error analysis** and investigate whether **adding coreference signal from the pre-trained coreference models** can help to improve the performance and reduce training time of WCS.

For this shared task we combined the outputs of two different models using simple heuristics and we would like to **experiment with a coreference editor** model that can learn optimal combinations.

Discourse Deixis

Discourse Deixis

Discourse deixis resolution is a process of linking abstract anaphors (usually NPs) to discourse entities, such as propositions, facts, descriptions of events, situations, etc.:

A: [Although... if you help me leave, I can pay you handsomely for your troubles...]$_i$ B: [That]$_i$ would be a violation of my duty.

Discourse Deixis

Automatic detection of discourse deictic anaphors is challenging:

- They are difficult to differentiate from ‘standard’ anaphors
- Interpretation of a mention as a discourse deictic anaphor often depends on the antecedent and/or context
- Abstract character of such anaphors is difficult to model
- Discourse deictic anaphors are less frequent than ‘standard’ anaphors

Discourse Deixis

For this Shared Task we focused on the resolution of *this*, *that*, *it* and *which* only. These pronouns make 72.3% of all discourse deictic anaphors in the data:

this	*that*	*it*	*which*	other
8.2%	52.9%	8.3%	2.9%	27.7%

Table 4: Distribution of discourse deictic anaphors in the training data

Discourse Deixis

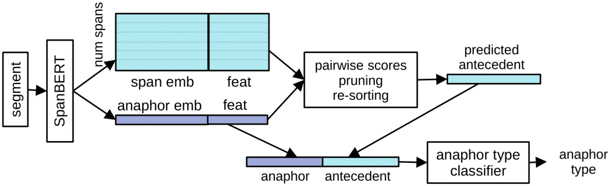
Challenge: target anaphor candidates may be of 3 different types:

- Discourse deictic (14.8%): *A: [Although... if you help me leave, I can pay you handsomely for your troubles...]$_i$ B: [That]$_i$ would be a violation of my duty.*
- Anaphoric (41.2%): *I am starvin - even [that] hay$_j$ is looking tasty.*
- Non-referential (44%): *He is so sure [that] he is right.*

Discourse Deixis

We implement discourse deixis resolution as a multi-task learning model:

1. Consider all spans in the given segment potential antecedents
2. Represent both anaphor and antecedent candidates as embeddings with additional shallow linguistic features
3. Calculate pairwise anaphor-antecedent scores similar to the *coref-hoi* model and choose the antecedent based on the largest score
4. Use anaphor-antecedent pair representation to classify the anaphor type and discard non-discourse deictic anaphors



Discourse Deixis

Representation of mentions:

Anaphor	Antecedent	Pair
token emb.	start emb.	sentence dist. emb.
parent emb.	end emb.	token dist. emb.
local context emb.	weighted avg. emb.	similarity emb.
POS tag emb.	span width emb.	
DEP tag emb.	span type emb.	
	end token POS emb.	
	end token DEP emb.	

Table 5: Representations of anaphor and antecedent candidates, and pairwise features

Discourse Deixis

The model was trained using a combination of several loss functions:

- marginal log-likelihood loss of possibly correct antecedents
- anaphor type loss checking how well the model distinguishes between discourse deixis, identity and non-referential anaphors
- label loss that punishes the model if it tends to reject all antecedent candidates while having a referential anaphor
- span type loss checking how well the model can differentiate between valid (verbal and nominal) and invalid (various fragments) antecedents

Discourse Deixis

Results (2nd place):

Track	Light			AMI			Persuasion			Swbd.		
	Baseline	Ours	Winner	Baseline	Ours	Winner	Baseline	Ours	Winner	Baseline	Ours	Winner
Eval-DD (Pred)	10.94	36.82	37.09	17.39	50.09	53.31	16.61	47.04	54.59	13.30	n/a	49.76
Eval-DD (Gold M)	18.14	35.91	38.38	22.95	47.13	55.12	30.15	48.24	54.89	21.37	n/a	49.83
Eval-DD (Gold A)	40.07	44.95	52.40	39.89	56.54	72.50	51.43	62.79	69.61	37.72	n/a	72.11

Table 6: CoNLL F1 scores on the official test sets

Discourse Deixis

Error analysis:

- The model struggles with the anaphor type identification: out of 292 true discourse deictic *this*, *that*, *it* and *which* 62 (21.25%) are classified as anaphoric, and 18 (6.16%) as non-referential ones
- The model successfully finds antecedents for only 144 (67.92%) out of 212 correctly identified discourse deictic anaphors
- The model also has difficulties finding split antecedents: 41 anaphors (14.04%) out of 292 refer to them, but our model resolves only 7.

Discourse Deixis

Future plans:

- Make the model more computationally efficient
- Check the influence of linguistic features on a larger training set
- Expand the set of potential anaphors
- Adapt the model for the resolution of identity anaphora

Bridging

Bridging

Bridging includes cases of anaphoric mentions linked to their antecedents by various associative (non-identity) relations:

[I]$_i$ m not much fer fightin, after that arrow ta [the knee]$_i$ in the last war.

Challenges:

- Difficult to differentiate between bridging and ‘standard’ anaphors
- Interpretation of a mention as a bridging anaphor depends on the antecedent
- Low inter-annotator agreement and lack of annotated data

Bridging

We submit our model to the Eval-Br (Gold A) track, in which gold anaphors are given.

Our approach is based on the ‘independent’ variant of the higher-order coreference architecture introduced in Joshi et al. (2019) with some modifications [Renner et al., 2021]:

- No calculation of the mention score $s_m(x)$, and coarse part of the coarse-to-fine pairwise score $s_p(x, y)$ is removed, as gold anaphors and mentions are given
- Passing only one anaphor at a time into the model (together with the whole document text) to decrease memory usage
- No span pruning
- Cross entropy loss instead of marginal log-likelihood

Bridging

Results:

	Light	AMI	Persuasion	Switchboard
Winner	46.80	39.35	56.91	44.40
Ours*	37.68	35.23	50.99	35.78
Baseline	29.93	22.69	37.89	30.39

Table 7: Test set results for the bridging task (gold anaphors)

*The model was trained on the Shared Task data (AMI, Switchboard, Light, Persuasion) plus BASHI [Rösiger, 2018] and ISNotes [Markert et al., 2012] corpora

Conclusion

Conclusion

Our system for the **anaphora track** combines the outputs of WCS and *coref-hoi* trained with “cluster merging”. It ranked second in the shared task competition.

Our system for the **discourse deixis track** is based on a novel idea that it is possible to combine the tasks of discourse deixis resolution and anaphor type classification. It ranked second in the shared task.

Our implementation for the **gold bridging track** is based on a higher order coreference system [Joshi et al., 2019] adapted for the shared task setting. It also ranked second in the competition.

Anaphora (WCS): tatiana.anikina@dfki.de

Discourse Deixis: natalia.skachkova@dfki.de

Anaphora & Bridging: joseph.renner@inria.fr

Anaphora & Bridging: priyansh.trivedi@inria.fr

Thank you for your attention!

References i

Anikina, T., Oguz, C., Skachkova, N., Tao, S., Upadhyaya, S., and Kruijff-Korbayova, I. (2021).
Anaphora resolution in dialogue: Description of the DFKI-TalkingRobots system for the CODI-CRAC 2021 shared-task.
In *Proceedings of the CODI-CRAC 2021 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue*, pages 32–42, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018).
BERT: Pre-training of deep bidirectional transformers for language understanding.
arXiv preprint arXiv:1810.04805.

References ii

- Joshi, M., Levy, O., Zettlemoyer, L., and Weld, D. (2019).
  BERT for coreference resolution: Baselines and analysis.
  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 5803–5808, Hong Kong, China. Association for Computational Linguistics.

- Markert, K., Hou, Y., and Strube, M. (2012).
  Collective classification for fine-grained information status.
  In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 795–804, Jeju Island, Korea. Association for Computational Linguistics.

References iii

Pennington, J., Socher, R., and Manning, C. D. (2014).
Glove: Global vectors for word representation.
In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Renner, J., Trivedi, P., Maheshwari, G., Gilleron, R., and Denis, P. (2021).
An end-to-end approach for full bridging resolution.
In *Proceedings of the CODI-CRAC 2021 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue*, pages 48–54, Punta Cana, Dominican Republic. Association for Computational Linguistics.

References iv

Rösiger, I. (2018).
BASHI: A corpus of Wall Street Journal articles annotated with bridging links.
In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Speer, R., Chin, J., and Havasi, C. (2017).
ConceptNet 5.5: An Open Multilingual Graph of General Knowledge.
In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4444–4451. AAAI Press.

References v

Xu, L. and Choi, J. D. (2020).
Revealing the myth of higher-order inference in coreference resolution.
In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8527–8533, Online. Association for Computational Linguistics.