Anaphora Resolution in Dialogue: Cross-Team Analysis of the DFKI-TalkingRobots Team Submissions for the CODI-CRAC 2021 Shared-Task

Natalia Skachkova, Cennet Oguz, Tatiana Anikina,
Siyu Tao, Sharmila Upadhyaya, Ivana Kruijff-Korbayová
DFKI / Saarland Informatics Campus, Saarbrücken, Germany
ivana.kruijff@dfki.de

Abstract

We compare our team's systems to others submitted for the CODI-CRAC 2021 Shared-Task on anaphora resolution in dialogue. We analyse the architectures and performance, report some problematic cases in gold annotations, and suggest possible improvements of the systems, their evaluation, data annotation, and the organization of the shared task.

1 Introduction

The goal of this paper is to compare the results obtained by the anaphora resolution systems developed by the DFKI-TalkingRobots Team for the *CODI-CRAC 2021 Shared-Task* (CCST) (Khosla et al., 2021) to the results of the other systems in CCST.[1] We submitted two systems for the anaphora resolution task (AR), and one system for discourse deixis resolution (DD). Table 1 shows their main characteristics, for detailed descriptions see Anikina et al. (2021).

Track	AR	AR	DD
System	WCS	M2M	DDR
Setting	pred.	pred.	pred.
Baselines	–	–	–
Learning framework	workspace clustering	mention to mention pairing	heuristics Siamese Net
Markable identification	SpaCy	BiLSTM-CRF	SpaCy
Train. data	CCST	CCST	CCST
Dev. data	CCST	CCST	CCST

Table 1: WCS, M2M and DDR systems summary

2 Workspace Coreference System (WCS)

2.1 System Overview

First we discuss our Worspace Coreference System (WCS) for standard anaphora resolution, which we implemented from scratch for the CCST AR Track. WCS employs a clustering approach explained in (Anikina et al., 2021). WCS differs from other models that participated in the competition in several ways. Both winner models (UTD and lxucs) are based on the implementation described in (Xu and Choi, 2020) that is an adaptation of the end-to-end c2f-coref model introduced by (Lee et al., 2018). The model which achieved the third best result (KU_NLP) uses pointer networks and models pairwise relations between mentions. Unlike these models WCS compares each mention to the clusters available in the workspace and decides whether it should be assigned to one of the existing clusters or initiates a new cluster. Also, WCS did not use any additional data for training and relied only on the annotations provided by the CCST organizers.

WCS achieved an average F1-score of 64.99% on the Light data, 43.93% on AMI, 59.93% on Persuasion and 53.55% on Switchboard. When evaluated on the gold mentions, WCS showed the following results: 73.24% F-score on Light, 56.48% on AMI, 72.06% on Persuasion and 62.85% on Switchboard. Having gold mentions as input gives an improvement of up to 12% on AMI and Persuasion, up to 9% on Switchboard and 6% on Light.

WCS ranked number four on the leaderboard. We see this as an encouraging result for the first prototype of this coreference system.

2.2 Performance Analysis

We analyzed the precision and recall tables provided by the CCST organizers and compared the outputs of different systems. In general, personal pronouns were confused most frequently by all models. For example, in the AMI corpus (*we*, *we*) pairs were annotated incorrectly most of the time, followed by (*I*, *I*) and (*I*, *you*) wrong pairings. On average, there were 5,608 incorrectly resolved pairs of *'we'* pronouns and 3,252 incorrect pairings of *'I'*

[1] URL: competitions.codalab.org/competitions/30312

Proceedings of the CODI-CRAC 2021 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue, pages 63–70
Punta Cana, Dominican Republic, November 10, 2021.

in the AMI test set.[2] This indicates that the methods used by the systems to handle speaker information and switching leave room for improvement.

For example, we noticed that UTD clustered several first person pronouns (*we*, *we*) that were very far apart in the document and not necessarily coreferent. On the contrary, WCS avoided clustering such unrelated pronouns at the expense of missing some cases of long-distance coreference. Our system was unable to annotate these cases because WCS looks only at the recent and salient workspace clusters when making a clustering decision.

In general, WCS avoided clustering *'we'* and *'they'* if they were uttered by the same speaker while the output of other systems includes such pairings. To encode speaker information WCS uses a randomly generated embedding of fixed dimensionality, concatenated with other mention features to create input for the clustering network. However, the random nature of such speaker embeddings might result in some instability during training. We believe that the speaker augmentation approach implemented in the lxucs model suits the coreference task better because it handles speaker information in a more uniform way by prepending a special token to each utterance in dialogue.

Merged tokens such as *'Ill'* posed another challenge. E.g., the AMI test set had 16 *'Ill'* pronouns and WCS was unable to handle them properly because it requires more strict tokenization and uses special features for personal pronouns. Although markables extracted by other models were often more flexible and not dependent on the SpaCy output, they were also incorrect in quite a few cases. For example, we believe that UTD_NLP could achieve even better results by filtering markables since some pairs annotated by them as coreferent included phrases such as *'gon na'* (paired with *'I'*) or *'D _'* (paired with *'it'*).

WCS did not mix up *'we'* and *'well'* or *'were'* and *'we'* and did not annotate different occurrences of *'two'* as coreferent (this was a problem for both UTD and lxucs). However, WCS suffered from the wrong head matching. Especially, when there was an overlap between the two spans WCS tended to cluster both mentions together. For example, *'we'* was clustered with a long noun phrase that contained the same pronoun: *'a power supply which **we** d probably get its probably gon na be the battery'*. Also, noun phrase extraction of WCS that was based on the SpaCy chunking module contributed to some errors, e.g., phrases like *'Yeah i'* and *'Great and you'* were considered as valid markables by WCS.

WCS often fails to recognize coreference between concrete and metaphoric expressions: *'the royal dog'*/*'you'* or *'poor soul'*/*'you'*. In some cases WCS puts too much emphasis on the head features and in other cases it suffers from the hard constrains introduced by filters at the post-processing stage (especially number and animacy filters). Gold annotations include mentions of the same entity that may differ in number, e.g., in the following pair of sentences: *"Well, then you will be interested to know that the **STC** website is already pledging to help these kids out!"* and *"Have you ever been on **their** website?" 'STC'* is coreferent with *'their'* but the number feature is not the same, hence these two mentions were not clustered by WCS.

Some errors were common among all systems. For instance, partial NP- or head-overlap was often treated as an indication of coreference. Phrases like *'speech **recognition**'* and *'sound **recognition**'* were considered coreferent by 4 out of 5 systems. Deictic expressions such as *'here'*, *'a week'* or *'today'* were also annotated as coreferent by most systems even when they referred to different entities.

Other hard cases include coreference between contextually related mentions (e.g., *'your pay for this task'* and *'the whole thing'*, or *'this sword'* and *'your gift to us, the soldiers'*). Some abbreviations were also not resolved correctly by the participating systems: *'R_S_I'* and *'repetitive strain injury'*. Moreover, anaphoric cases that require additional inference to establish coreference were missing in the system outputs (e.g., *'the remote control'* and *'the product'*, or *'here'* and *'Texas'*).

Our observations suggest that the span errors and wrong clustering of personal pronouns are among the most frequent errors that are common for all participating systems. We observed that WCS makes less extra mention mistakes compared to other systems while it has more missing mentions in the output. This can be explained by the fact that the workspace in WCS has a limited size and if some mentions refer to the same entity but are very far apart in the document they are not clustered by WCS (e.g., in the Switchboard data there were occurrences of entity *'capital punishment'* separated by more than 380 tokens).

[2]This number represents the total number of wrong pairings divided by the number of participating systems.

2.3 Observations on Gold Annotations

We found some unclear and problematic cases in the gold annotation. For instance, *'although the temple is full , no one is speaking and all you can hear are muted scuffling feet'* and *'I'* were marked as the same entity in the Light data (episode_8093). Another example comes from the Persuasion test set that has two different entity IDs for *'I'* and *'Tommy'* in the sentence *"I m Tommy, by the way"* (Persuasion, 20180826-180314-10). We would expect these mentions to have the same entity IDs because both can be used to refer to the same entity interchangeably. Besides, there were some valid markables missing in the gold annotation, e.g. *'my university'* did not have a markable/entity annotation in *"And when I cant give monetarily donate food, my time, and go to blood drives on* ***my university****"* (Persuasion, 20180826-180314-24).

2.4 Outlook

We would like to perform an error-driven analysis as described in (Kummerfeld and Klein, 2013) to gain more insight in the types of mistakes made by different systems (e.g., missing vs. extra entities or divided vs. conflated entities). Kummerfeld and Klein (2013) implemented a tool for automated evaluation of different types of mistakes but it does not handle the CONLLUA format and, importantly, requires syntactic parse in the gold annotation. Since we do not have this information in the gold data we cannot perform a fine-grained analysis at this time.

As for the performance of WCS, we would like to improve it based on the analysis results and experiment with different features and embeddings. In particular, we are interested in trying out SpanBERT embeddings (Joshi et al., 2020) that were used by the winner models and combine WCS with the M2M system that was also implemented by the DFKI-TalkingRobots team (cf. Section 3).

The analysis revealed that the recall of our model was impaired by the hard animacy and number constraints. Hence, we would like to model these constraints differently, i.e., not as binary filters but as additional features provided as input to the clustering model. Another interesting venue to explore is contextual knowledge, because for each markable that needs to be resolved we need to identify relevant context that goes beyond the utterance and this context should be used for further inference. For example, coreference between such pairs as *'a child in the US'* and *'a child in need'* (Persuasion, 20180826-044626) proved difficult to resolve without reasoning over broader context since both NPs start with an indefinite article and have different PPs modifying the head noun. In fact, none of the participating systems annotated this pair correctly.

More work needs to be done on the dialogue-specific issues, especially personal pronouns and temporal/spatial deixis such as *'today'* and *'here'*. Our analysis showed that all systems had difficulties with such cases and a rule-based approach is not sufficient to handle them. Perhaps deictic mentions could be detected and resolved using a different module within the coreference system.

Finally, we would like to further investigate different ways of representing mentions and clusters in dialogue and apply our model to other data, e.g., dialogues from the emergency response domain (Kruijff-Korbayová et al., 2015). Some of the contextual reference challenges observed in these dialogues are discussed in (Skachkova and Kruijff-Korbayová, 2020).

3 Mention-to-Mention System (M2M)

3.1 System Overview

Our second submission to the AR track was the Mention-to-Mention (M2M), cf. system (Anikina et al., 2021). M2M differs from the other participating systems in that it resolves anaphora in a pairwise fashion: For each mention, M2M performs a similarity comparison with all preceding mentions and pairs the current mention with the most similar preceding one (or none). After processing all mentions in a document M2M creates clusters based on transitive closure over the mention-mention pairs, i.e., pairs which have a common mention are clustered together. To compute similarity, M2M uses the head, speaker, and span features of mentions as well as a distance vector generated with several distance functions, namely minkowski, euclidean, manhattan, chebyshev, cityblock, and braycurtis.

M2M consists of three different models. The *Self Model* is built for personal pronouns with average embeddings of the speaker and head features of the mentions. The *Pronoun Model* is designed to pair third person and place pronouns to the corresponding candidate by exploiting the head average embeddings of mentions. The *Noun Model* attempts to pair mentions by using head and span embeddings with contextualized and average embeddings when they are noun phrases. All three models are imple-

mented with linear layers with a ReLU activation function to use the pair of mention's head, speaker, and span features. The extracted features from the linear layers are passed to a sigmoid function for obtaining output probabilities.

The candidate mentions for M2M were extracted using a BiLSTM-CRF sequence tagging model with an ELMo embedding layer. The model is trained using the data provided by the CCST organizers. The data were preprocessed into IOB2 format, which means that, in cases of nested markables, only the markable with the widest scope is used in training. Consequently, we use SpaCy (*en_core_web_trf* model) to extract all noun chunks as nested markables in a post-processing step. See (Anikina et al., 2021) for more details.

3.2 Performance Analysis

We first evaluate the performance of the BiLSTM-CRF model in extracting the candidate mentions. We observe that the system predicts many more candidate mentions than are in the gold annotations, which implies a high false positive rate. This is indeed further confirmed in our analysis, as we note poor precision scores across the four test sets (25.8% on Light, 23.7% on Persuasion, 23.5% on Switchboard, and 23.1% on AMI).

Although the high false positive rates are not ideal and preclude meaningful manual inspection of the results, they are not wholly unexpected, as our model was meant to extract as many markable candidates as possible, with the possibility of false positives taken into account in further procesteps. Thus, the recall is perhaps the better metric for our purposes and indeed we observe better recall scores than precision across the board (48.7% on Light, 47.3% on Persuasion, 56.9% on Switchboard, and 54.7% on AMI). Nevertheless, as the mention extraction has a lot of room for improvement, we expect that the M2M system will perform better if evaluated on the gold mentions instead.

The M2M system achieved F1-score of 61.26% on Light, 59.20% on Persuasion and 51.24% on Switchboard. Since only one model of each team is included in the CCST Official Ranking document provided by the organizers, we do not have details for a performance comparison of M2M against the other systems. An analysis of M2M's predictions against the gold annotations reveals that the *Self Model* shows high performance. Thus, the utilized distance vector of the speaker names increases the similarity of the mentions even though the pronouns for the entities are changing, e.g., *'I' - 'you'*. On the other hand, the performance of both *Pronoun* and *Noun Model* decreases, because of the lexical difference between an anaphor mention and a candidate antecedent mention, e.g. *'he'* and *'father'*, and *'father'* and *'Paul'*, respectively.

3.3 Outlook

The M2M system relies on lexical similarity between pairs of mentions instead of the similarity between mentions and clusters. Thus, the system is fragile to lexical variations because of the generated distance vector. However, the same vector increases the performance of the *Self Model* because of the speaker names. Therefore, in order to exploit the distance vector for personal pronoun resolution in the *Self Model*, we need to also apply clustering, to avoid the variations between mentions in *Pronoun Model* and *Noun Model*. Therefore, we plan to combine M2M and WCS (Section 2) to benefit from their complementary strengths.

4 Discourse Deixic Resolution (DDR)

Next we present a cross-team comparison of the discourse deixis resolution systems submitted to CCST, as well as some difficulties we had with the gold discourse deixis annotations. We discuss the *Eval-DD(Pred)* track, as we did not participate in *Eval-DD(Gold)*. There were only two submissions, from us and from the UTD_NLP team.

4.1 System overview

At first sight the two approaches look very different. We utilize machine learning using a Siamese network together with hand-crafted rules (Anikina et al., 2021). The UTD_NLP system is purely machine learning based. It extends the model by Xu and Choi (2020), originally designed for entity coreference resolution. While we perform discourse deixis mention detection and resolution separately, the UTD_NLP system does it jointly.

Despite these differences, both systems actually exploit the same idea – given an anaphor to resolve, they rank several antecedent candidates using neural networks. The two realizations of this idea are more similar than it may seem. We check antecedent-anaphor 'compatibility' by replacing the anaphor with an antecedent candidate and comparing the encoding of the resulting utterance with the encoding of the original utterance

containing the anaphor. The UTD_NLP system encodes anaphors and antecedent candidates separately from the utterances they are part of (incorporating context with SpanBERT instead) and afterwards ranks their joint 'compatibility' using only distance as a feature. Nevertheless, the UTD_NLP system achieves twice as high F-scores.

4.2 Performance Analysis

Table 2 shows the results of discourse deixis mention extraction of both systems. We use a simple *SpaCy*-based method to find anaphors. This gives us moderate recall; according to our rules many anaphors (e.g., all noun phrases and personal pronoun *it*) are simply omitted. Because the approach is greedy, the precision is really low, which leads to low F-scores. In our case the choice of antecedent candidates for each anaphor is also rule-based. Because of erroneous or missed anaphors, as well as mistakes made by our Siamese Net-based scorer and sometimes strange gold annotations (see examples in Section 4.4) the scores for antecedents are even worse than for anaphors.

Table 2 shows that the anaphor identification method employed by the UTD_NLP team achieves better recall for the Light and Switchboard test sets, but for the other two it is lower than ours. Having analysed the recall tables prepared by the CCST organizers we noticed that while our system simply ignores all the anaphors that have a *SpaCy* POS tag different from DET, the UTD_NLP system also has difficulties recognizing anaphors other than *'this'* and *'that'*. E.g., both systems miss absolutely all anaphors represented by NPs (with an exception of one case of *'the same'* recognized by UTD_NLP). Additionally, the winning system misses many anaphors represented by *'it'*, all cases of *'which'*, as well as all capitalized anaphors. This behavior can be caused by the insufficient amount of such markables in the training data.

Interestingly, UTD_NLP's recall of antecedent identification is worse than ours for all the files. One possible explanation of this can be the fact that the UTD_NLP system seems to omit all split antecedent cases, i.e. they probably do not consider antecedents consisting of several utterances, and thus miss a certain number of markables. This hypothesis is supported by very low F-scores for singleton clusters reported by the team in their system description paper. As far as we can judge, gold annotations use singleton clusters as containers for parts of split antecedents, and low F-scores may mean that split antecedents are disregarded. Still, due to a much better precision of both anaphor and antecedent recognition, UTD_NLP achieves higher F-scores for discourse deixis mention extraction.

To compare clustering results of both teams we refer to the recall and precision files provided by the CCST organizers. Table 3 compares the number of (in)correct pairs identified by each system. The numbers are disjoint, e.g., the Light test set contains 67 antecedent-anaphor pairs, of which 6 pairs were correctly predicted by both systems, other 19 pairs only by the UTD_NLP system, and different 5 ones only by our system. At the same time, of 133 incorrectly predicted pairs in the Light test set 10 were found in the output of both systems, 25 pairs were predicted only by the UTD_NLP system, and 98 pairs only by our system.

One can notice rather poor recall of both systems, with ours being often (but not always) worse than by UTD_NLP's. Our system has especially low recall on the Light test set. This probably happens, because Light contains quite a number of cases, where an anaphor is parts of its antecedent, and our approach does not consider this possible. Table 3 also shows that our system, in contrast to the UTD_NLP's approach, creates really many wrong antecedent-anaphor pairs. This is probably due to the fact that our system has very low precision for anaphor identification, and thus tries to resolve too many false anaphors. Also, having taken a closer look at our output files and gold annotations, we saw that our approach was able to correctly recognize only a small number of split antecedents. As far as we can judge, no anaphors referring to split antecedents are present in the recall/precision files prepared by the CCST organizers. Still, it is obvious that detecting split antecedents is very challenging for our system.

4.3 Outlook

The UTD_NLP system demonstrates that a model originally designed for entity coreference resolution can be rather successfully adapted for discourse deixis resolution without adding any sophisticated features. It is difficult to tell what exactly helps their model to achieve better scores in comparison with our approach. It can be the usage of SpanBERT, more refined scoring method (it includes three components), longer training with better chosen parameters, a combination of these

		Light			AMI			Persuasion			Swbd		
		P	R	F	P	R	F	P	R	F	P	R	F
Anaphor	UTD_NLP	71.4	68.8	70.1	58.0	64.4	61.0	76.7	64.2	69.9	65.7	70.7	68.1
	DFKI_TR	24.2	57.8	34.1	17.3	79.8	28.5	37.0	69.4	48.3	25.2	66.5	36.6
Antecedent	UTD_NLP	50.8	27.7	35.8	66.0	20.5	31.3	59.6	21.2	31.3	60.8	21.5	31.7
	DFKI_TR	19.1	37.5	25.3	11.5	37.3	17.5	21.8	49.3	30.2	20.9	39.4	27.3

Table 2: Markable extraction results (UTD_NLP's scores come from Kobayashi et al. (2021))

	Corpus	Light	AMI	Persuasion	Swbd
Recall	UTD_NLP	19	13	18	31
	DFKI_TR	5	14	13	36
	Both	6	12	7	26
	Total	67	100	113	224
Precision	UTD_NLP	25	36	27	142
	DFKI_TR	98	316	108	358
	Both	10	4	8	16
	Total	133	356	143	516

Table 3: Discourse deixis resolution results

reasons, or something else. We also think that both systems can be improved at least in terms of markable identification. It is clear that in our case the most obvious step is the improvement of precision. The UTD_NLP team could, on the other hand, rather easily incorporate split antecedents. Currently they limit the width of a span by 70 tokens, so that spans consisting of several utterances are not considered. Allowing larger spans would help identifying split antecedents, and probably improve the recall for antecedent identification.

4.4 Observations on Gold Annotations

We highly appreciate the efforts that were invested into the discourse reference annotation of the data sets provided for the CCST. However, some aspects could be improved.

First, the available brief description of the corpus format,[3] is very useful, but it does not provide an annotation manual with enough explanations of the annotation principles (e.g., no clear definition of a markable) and examples. The instructions that were used by the annotators have not been published. So we needed to infer these principles from data/examples themselves. In some cases it was difficult to correctly determine markables and their spans. Especially, when the data contains inconsistent examples. E.g., in the *Light_dev* file in the utterance *"I will do my best to spread your kind words around , my god ."* the noun phrase *'your kind words'* is annotated as a discourse deixis markable, but the phrase *'your words'* in *"You are the wisest merchants in these lands , and I shall heed your words - ..."* is not, despite the fact that it has an antecedent.

Also, some antecedents' spans seem to be too long. E.g., according to the gold annotations of the same *Light_dev* file, *'That'* in *"That certainly scared the eternal daylights out of me !"* has the whole utterance *"I myself happened to look out from the kitchen and see his dangling legs in front of the window ."* as the antecedent, while logically only the part *'see his dangling legs in front of the window'* should be considered as such. Sometimes the anaphor is annotated as a part of the antecedent, like the pronoun *'this'* in *"You are a good man , I can recognize this ."* from the same file.

Second, some files with gold annotations have markables with mismatched closing brackets. Thus, in the file *Pear_Stories* some markables span over several utterances, and may even contain nested markables. E.g., *"Then a kid came along on a bicycle"* (which is actually one of the fragments of a larger utterance) is a separate discourse deixis markable, and at the same time it is the beginning of a larger markable whose closing bracket can be found three fragments later. At the same time both these markables (together with eight other ones that follow) are split antecedents of an anaphor that occurs later in the dialogue.

5 Conclusions and CCST Outlook

We compared our systems and their results to the systems and results by other teams. We also presented some ideas that this analysis gave us for the further development of our systems. Besides improvements of mention detection we also plan to combine WCS and M2M to benefit from their complementary strengths. As the analysis showed, M2M did well on resolving personal pronoun coref-

[3] See https://github.com/UniversalAnaphora/UniversalAnaphora/blob/main/UA_CONLL_U_Plus_proposal_0.1.md

erence, whereas WCS' clustering approach is better than M2M at noun coreference cases. Another interesting challenge to consider is to also account for discourse deixis in one integrated system or even by the same approach, similarly to the UTD_NLP system.

Admittedly, the difficult examples that we have pointed out are ones that every state-of-the-art system finds hard. In the scope of the analysis that we were able to carry out given the time and resources, we could not identify properties of examples that different system configurations are able to handle versus not, in order to draw generalizations on what kinds of examples are handled well by what approaches, and why. Such analysis is also complicated by the fact that the systems differ in markable detection. A systematic deeper analysis is an interesting topic for future research.

To conclude, we would like to make a few suggestions for the future of the shared task. We think that the shared task is very useful and it would be good to continue running it, because anaphora resolution is far from being a solved task.

First, we suggest to have separate tracks for the evaluation of systems trained only on the provided data vs. systems (pre-)trained on external/additional data. This would help not only to achieve a fair comparison but also to see whether improvements are due to (pre-)training with more data and transfer or due to better feature engineering and modeling. In order to study what specifically was learned from the external data that could not be learned by the training data provided in CCST it would be useful to compare the results of models trained with the same approach on only the external data vs. only the CCST data vs. using all data.

Second, we would like to see a separate comparison of mention detection and anaphor identification results for the different systems. Our experience showed that finding correct mentions is crucial for the overall performance and we believe that this stage should be evaluated separately. It would also be interesting to compare results on both dialogues and narrative texts.

Third, future competitions might consider using more fine-grained evaluation metrics because different cases of coreference have different difficulty levels. E.g., '*a black cat'* and *'the black cat*' are much easier to resolve compared to *'a black cat'* and *'the dark and furry creature'*. We believe that a more differentiated evaluation could be done (at least partially) in an automated way. For instance, similarly to the sieve-based approach (Raghunathan et al., 2010) one could compare a referent and its true antecedent using exact string matching vs. head matching, compute similarity between the corresponding mention embeddings or count the number of candidates (valid markables) between the anaphor and its antecedent. Our analysis showed that some tricky cases of coreference, e.g., *'this sword'* and *'your gift to us, the soldiers'* were missed by all participating systems and we expect that weighting each pair w.r.t. the difficulty level and/or distinguishing difficulty levels in the presentation of the results would help to shift the focus towards resolving 'harder' cases. The evaluation could also report results on pronominal and nominal anaphora separately, and distinguish the different syntactic types of antecedents.

Finally, we have some suggestions for the organization of the shared task. Namely, it would be helpful to have a more detailed description of the task, its tracks, test data format (with a couple of input-output examples). For instance, it was unclear that for the *Eval-DD(Gold)* track we would only get anaphors in the IDENTITY column and would need to distinguish between discourse deixis and 'standard' anaphora markables. More information regarding the submission procedure would also be useful, for example for the case that a team has more than one system for one track. Quality control of the annotations should be improved, or gold vs. silver (vs. bronze) data should be distinguished. Data quality is very important, because we are making conclusions about the correctness of models based on the annotations. One possibility could be to ask participating teams to provide annotations or additional data sets annotated according to uniform guidelines.

As for new challenges to include in the shared task in the future, there is no shortage of smaller steps and giant leaps. The easiest ones were already mentioned above, i.e., evaluation of the contribution of external training data, separate evaluation of mention detection and anaphor identification, as well as a more finegrained evaluation scheme, in order to obtain more insight into the capabilities of the different models. Other potential extensions depend on the availability of annotated data, and even agreed annotation standards. Thanks to existing datasets it would probably be most feasible to

extend the task, on the one hand, to zero pronouns and other forms of implicit entity reference and/or event references, and on the other hand, to other languages. Multimodal reference would also be interesting, but more challenging, from reference to visual objects embedded in text to reference to physical situations, for example in cooking or physical activity videos, which are used in the language-and-vision community. Finally, non-verbal forms or reference, such as pointing, are in our view the most challenging, due to lack of data as well as annotation standards, and the difficulty of annotation.

Acknowledgements

The authors have been supported by the German Ministry of Education and Research (BMBF): T. Anikina and I. Kruijff-Korbayová through project CORA4NLP (grant Nr. 01IW20010); C. Oguz through project IMPRESS (grant Nr. 01IS20076) and N. Skachkova through project A-DRZ (grant No. I3N14856).

References

Tatiana Anikina, Cennet Oguz, Natalia Skachkova, Siyu Tao, Sharmila Upadhyaya, and Ivana Kruijff-Korbayová. 2021. Anaphora Resolution in Dialogue: Description of the DFKI-TalkingRobots System for the CODI-CRAC 2021 Shared-Task.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Trans. Assoc. Comput. Linguistics*, 8:64–77.

Sopan Khosla, Juntao Yu, Ramesh Manuvinakurike, Vincent NG, Massimo Poesio, Michael Strube, and Carolyn Rosé. 2021. The CODI-CRAC 2021 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue. In *Proceedings of the CODI-CRAC 2021 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue*. Association for Computational Linguistics.

Hideo Kobayashi, Shengjie Li, and Vincent Ng. 2021. Neural Anaphora Resolution in Dialogues.

Ivana Kruijff-Korbayová, Francis Colas, Mario Gianni, Fiora Pirri, Joachim de Greeff, Koen V. Hindriks, Mark A. Neerincx, Petter Ögren, Tomás Svoboda, and Rainer Worst. 2015. TRADR project: Long-term human-robot teaming for robot assisted disaster response. *Künstliche Intell.*, 29(2):193–201.

Jonathan K. Kummerfeld and Dan Klein. 2013. Error-driven analysis of challenges in coreference resolution. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 265–277.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 687–692.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nate Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher D. Manning. 2010. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP 2010, 9-11 October 2010, MIT Stata Center, Massachusetts, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 492–501.

Natalia Skachkova and Ivana Kruijff-Korbayová. 2020. Reference in team communication for robot-assisted disaster response: An initial analysis. In *Proceedings of the 3rd Workshop on Computational Models of Reference, Anaphora and Coreference. (CRAC-2020), COLING'2020, Virtual, Spain*. Association for Computational Linguistics.

Liyan Xu and Jinho D. Choi. 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.