Adapted End-to-End Coreference Resolution System for Anaphoric Identities in Dialogues

Liyan Xu
Computer Science
Emory University, Atlanta, GA
liyan.xu@emory.edu

Jinho D. Choi
Computer Science
Emory University, Atlanta, GA
jinho.choi@emory.edu

Abstract

We present an effective system adapted from the end-to-end neural coreference resolution model, targeting on the task of anaphora resolution in dialogues. Three aspects are specifically addressed in our approach, including the support of singletons, encoding speakers and turns throughout dialogue interactions, and knowledge transfer utilizing existing resources. Despite the simplicity of our adaptation strategies, they are shown to bring significant impact to the final performance, with up to 27 F1 improvement over the baseline. Our final system ranks the 1st place on the leaderboard of the anaphora resolution track in the CRAC 2021 shared task, and achieves the best evaluation results on all four datasets.

1 Introduction

Coreference resolution of anaphoric identities (a.k.a. anaphora resolution) is a long-studied Natural Language Processing (NLP) task, and is still considered one of the unsolved problems, as it demands deep semantic understanding as well as world knowledge. Although there is a significant performance boost recently by the neural decoders (Lee et al., 2017, 2018) and deep contextualized encoders such as BERT and SpanBERT (Joshi et al., 2019, 2020), the majority of the experiments are based on *OntoNotes* (Pradhan et al., 2012) from the CoNLL 2012 shared task, which may overestimate the model performance due to two perspectives: the lack of support for harder cases such as singletons and split-antecedents, and the lack of focus on real-world dialogues. In this work, we target on the task of anaphora resolution in the CRAC 2021 shared task (Khosla et al., 2021) that addresses both perspectives, and present an effective coreference resolution system that is adapted from the recent end-to-end coreference model.

All datasets in the CRAC 2021 shared task are in the Universal Anaphora format. For simplicity, we refer to it as the UA format, and refer to the annotation scheme of the CoNLL 2012 shared task as the CoNLL format. The UA format is an extension of the CoNLL format, and further supports bridging references and discourse deixis. For anaphora resolution, the UA format differs from the CoNLL format on three aspects: the support of singletons, split-antecedents, and non-referring expressions (excluded from the current evaluation). Our approach specifically addresses the singleton problem (Section 3.1), which is shown to be a critical component under the UA setting that brings 17-22 F1 improvement on all datasets (Section 5.2). Few recent work has studied the split-antecedent problem (Zhou and Choi, 2018), and we leave the split-antecedents as future work.

In addition to singletons, our approach also emphasizes on the speaker encoding (Section 3.3) and knowledge transfer (Section 3.4) to address the dialogue-domain perspective. Especially, we use a simple strategy of speaker-augmented encoding that captures the speaker interaction and dialogue-turn information, utilizing the strong Transformers encoder. It has been shown by the previous study that conversational metadata such as speakers can be significant for coreference resolution on dialogue documents (Luo et al., 2009), and we do see 2-3 F1 improvement on three datasets by simply applying the speaker encoding strategy (Section 5.3).

Knowledge transfer from other existing resources is also shown to be important in our approach. Two different strategies are experimented, and the domain-adaptation strategy is able to bring large improvement, boosting 8 F1 for *LIGHT* and 6 F1 on *PSUA* (Table 3).

Our final system ranks the 1st place on the leaderboard of the anaphora resolution track in the CRAC 2021 shared task, and achieves the best evaluation results on all four datasets, with 63.96 F1 for *AMI*, 80.33 F1 for *LIGHT*, 78.41 F1 for *PSUA*, 74.49 F1 for *SWBD* (Section 5.1). A brief summary of our final submission is shown in Table 4.

Proceedings of the CODI-CRAC 2021 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue, pages 55–62
Punta Cana, Dominican Republic, November 10, 2021. ©2021 Association for Computational Linguistics

2 Related Work

Pretrained Transformers encoders have been successfully adopted by recent coreference resolution models and shown significant improvement (Joshi et al., 2019, 2020). We also adopt the Transformers encoder in our approach because of its superior performance. For the neural decoder, there have been two popular directions from recent work. One is mention-ranking-based, where the model predicts only one antecedent for each mention without focusing on the cluster structure (Wiseman et al., 2015; Lee et al., 2017; Wu et al., 2020). The other is cluster-based, where the model maintains the predicted clusters and performs cluster merging (Clark and Manning, 2015, 2016; Xia et al., 2020; Yu et al., 2020). We adopt the mention-ranking framework in our approach because of its simplicity as well as its state-of-the-art decoding performance.

3 Approach

3.1 Mention-Ranking (MR)

Our baseline model MR adopts the mention-ranking strategy, and follows the architecture of the end-to-end neural coreference resolution model (Lee et al., 2017, 2018) with a Transformer encoder (Joshi et al., 2019, 2020). Given a document with T tokens, the model first enumerates all valid spans, and scores every span for being a likely mention, denoted by the mention score s_m. The model then greedily selects top λT spans by s_m as mention candidates that may appear in the final coreference clusters. Let $\mathcal{X} = (x_1, \dots, x_{\lambda T})$ be the list of all mention candidates in the document, ordered by their appearance. For each mention candidate $x_i \in \mathcal{X}$, the model selects a single coreferent antecedent from all its preceding mention candidates, denoted by $\mathcal{Y}_i = (\epsilon, x_1, \dots, x_{i-1})$, with ϵ being a "dummy" antecedent that may be selected when x_i is not anaphoric (no antecedents).

The antecedent selection is performed by the pairwise scoring process between the current mention candidate x_i and each of its preceding candidate $y \in \mathcal{Y}_i$. The final pairwise score $s(x_i, y)$ consists of three scores: how likely each candidate being a mention, measured by the mention score s_m; and how likely they refer to the same entity, measured by the antecedent score s_a. The final score $s(x_i, y)$ can be denoted as follows:

$$s(x_i, y) = s_m(x_i) + s_m(y) + s_a(x_i, y, \phi(x_i, y))$$

Both s_m and s_a are computed by the FeedForward Neural Network (FFNN), and $\phi(x_i, y)$ represents additional meta features. Unlike previous work, we do not include the specific genre as a feature; instead, we simply use a binary feature on whether the document is dialogue-based or article-based, since dialogues can exhibit quite different traits from written articles (Aktaş and Stede, 2020). We also adopt a speaker feature that indicates whether two candidates are from the same speaker, or whether the speaker information is not available, which is important for written articles or two-party dialogues. In Section 3.3, we further adapt more speaker encoding to benefit multi-speaker dialogues and the personal pronoun issue.

For inference, the selected antecedent is the preceding candidate with the most pairwise score, denoted by $\text{argmax}_{y' \in \mathcal{Y}_i} s(x_i, y')$. For training, the marginal log-likelihood of all gold antecedents $\hat{\mathcal{Y}}_i \subseteq \mathcal{Y}_i$ for each $x_i \in \mathcal{X}$ is optimized, denoted by the loss $\mathcal{L}_c$:

$$P(y) = \frac{e^{s(x_i, y)}}{\sum_{y' \in \mathcal{Y}_i} e^{s(x_i, y')}} \quad (1)$$

$$\mathcal{L}_c = -\log \prod_{x_i \in \mathcal{X}} \sum_{\hat{y} \in \hat{\mathcal{Y}}_i} P(\hat{y}) \quad (2)$$

3.2 Singleton Recognition (SR)

As the UA format does support singletons, MR would fail to predict those singleton clusters, since the antecedent selection can only generate clusters with at least one pair of mentions. Several previous work has addressed the singleton problem from different perspectives (Yu et al., 2020; Zaporojets et al., 2021). Our SR model is built upon MR and further recognizes singletons based on the simple strategy as follows: we make use of the mention score s_m in the antecedent selection process, and create a singleton cluster for any candidates with $s_m > 0$ that have not yet found any antecedents, which poses an additional requirement on the mention score, such that only valid mentions should have $s_m > 0$. Let $\Psi^+ \subseteq \mathcal{X}$ be the set of gold mention candidates, and $\Psi^- = \mathcal{X} \setminus \Psi^+$ be the set of other mention candidates. We optimize the mention score with the binary cross-entropy loss $\mathcal{L}_m$

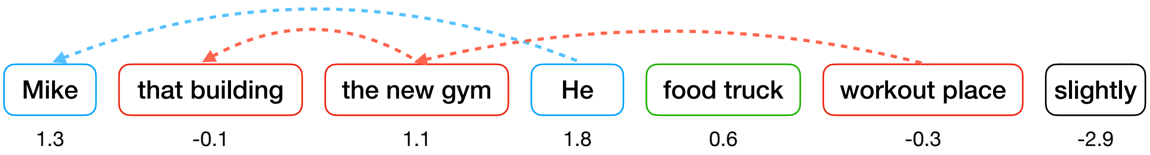

Figure 1: Example of the new antecedent selection process that support singletons (Section 3.2). Each arrow indicates the selected antecedent (the dummy antecedent is excluded), and the mention score s_m is shown below each mention. Mentions of the same predicted clusters are marked in the same color. Although no antecedent is selected for "food truck", it will still be assigned as a singleton cluster because of $s_m = 0.6 > 0$. "that building" and "workout place" are still assigned to the corresponding cluster even though their $s_m < 0$, to allow some slacks on the mention score prediction. "slightly" will not be assigned to any clusters.

and joinly train with the coreference loss $\mathcal{L}_c$:

$$\mathcal{L}_m = - \sum_{x_i \in \Psi^+} \log \sigma(s_m(x_i)) - \sum_{x_j \in \Psi^-} \log(1 - \sigma(s_m(x_j))) \quad (3)$$

$$\mathcal{L} = \mathcal{L}_c + \alpha_m \cdot \mathcal{L}_m \quad (4)$$

σ is the sigmoid function, and α_m is a hyperparameter. $\mathcal{L}$ is the final loss composed of two tasks. In practice, we also perform negative sampling on Ψ^- dynamically, so that Ψ^+ and Ψ^- are of similar sizes ($|\Psi^+| \approx |\Psi^-|$), to alleviate the negative effects from the skewed class distribution.

In the new selection process, we still regard the selected non-dummy antecedent y to be valid by $y = \text{argmax}_{y' \in \mathcal{Y}_i} s(x_i, y')$, even though the mention score of either candidate can be negative ($s_m(x_i) < 0$ or $s_m(y) < 0$). This is to allow certain slacks on the mention score prediction which could help with the mention recall. Figure 1 shows three different cases of the predicted clusters by the SR model.

3.3 Speaker Encoding (SE)

Our SE model is further adapted upon SR model and aims to strengthen the speaker encoding for each candidate representation. As we are targeting on the coreference resolution in dialogues, encoding speaker interactions becomes more critical, especially for the correct understanding of the speaker-grounded personal pronouns that are more frequent in dialogues than other non-dialogue genres (Aktaş and Stede, 2020).

The speaker feature introduced in Section 3.1 provides shallow distinction on whether two mentions are from the same speaker. However, the speaker interactions across dialogue turns are not presented in the document encoding; therefore, the representation of each candidate has no awareness on the speaker interactions at all. To provide deeper knowledge on the interactions, we adopt a simple but effective strategy that is similar to some other work in speaker encoding (Le et al., 2019; Wu et al., 2020): a special speaker token is prepended to each sentence, and we feed the new speaker-augmented document to the encoder directly.

Table 1 shows an example on this speaker augmentation. Each speaker is indexed by the order of the first appearance in the dialogue. All the special speaker tokens are added to the tokenizer vocabulary, and will be picked up in the tokenization and encoding process. Therefore, all encoded candidate representation in the SE model is conditioned on the entire speaker interactions, and automatically learns to fuse the information of speakers and turns in the training process.

John: Do you know Mike? Mary: He is my best friend! Paul: I like him too! Mary: We should meet together!
[SPK1] Do you know Mike ? **[SPK2]** He is my best friend ! **[SPK3]** I like him too ! **[SPK2]** We should meet together !

Table 1: Example for the speaker-augmented encoding. A special speaker token is assigned to each speaker and prepended to each corresponding sentence.

3.4 Knowledge Transfer

We also emphasize on the knowledge transfer in this task, as the training resources of dialogue corpora annotated in the UA format are limit and expensive to obtain, while there already exist larger-scale training corpora for other domains in different annotation schemes, e.g. the *OntoNotes* dataset in the CoNLL format that mainly consists of non-dialogue genres. For clarity, we denote the provided data annotated in the UA format as *UAD*, and other existing data in non-UA format as *OD*. We investigate two common ways to make use of *OD* in the training for SE, denoted as follows:

- SE$^{+\text{M}}$: **M**ix *OD* and *UAD* together as a larger dataset, regarding *OD* as data augmentation that provides more knowledge.

- SE$^{+\text{P}}$: **P**retrain the model on *OD* first, then further train the model on *UAD* only, regarding training on *UAD* as domain adaptation.

Above two choices are plausible in our approach, because we only use data in the CoNLL format for *OD*, which is still largely similar to the UA format, despite the difference on the singletons, non-referring expressions, and split-antecedents.

Similarly, we denote the model as SR$^{+\text{M/P}}$ if the SR model is used instead of SE.

4 Experiments

4.1 Datasets

For data in the UA format (*UAD*), we use the ARRAU corpus (Poesio et al., 2018) from the CRAC 2018/2021 shared task. Four sub-corpora are used as the training set for *UAD*, namely *TRAINS-93*, *PEAR*, *RST*, *GNOME*. One sub-corpus named *TRAINS-91* is used as one of the development (dev) set. In addition, four other corpora from the CRAC 2021 shared task are also used as the development set as well as the final test set, namely *AMI*, *LIGHT*, *Persuasion for Good* (*PSUA*), *Switchboard* (*SWBD*). All above datasets are of the dialogue domain except for *RST* and *GNOME*. Table 2 shows the detailed statistics of all *UAD* datasets. Note that certain datasets do not provide speaker information, therefore their averaged numbers of speakers per document are shown as 0.

For non-UA format data (*OD*), we use two datasets in the CoNLL format: *OntoNotes* (*ON*) (Pradhan et al., 2012) and *BOLT* (Li et al., 2016). *OntoNotes* consists of documents in six genres, where only two genres "Telephone Conversation" and "Broadcast Conversation" are of the dialogue domain; we use the same provided train/dev/test split for *OntoNotes*. *BOLT* has the same annotation scheme as *OntoNotes* and consists of documents from discussion forums, instant messages and telephone conversations. We perform a random 80/10/10 split for the train/dev/test set of *BOLT*. Detailed statistics of both datasets are shown in the bottom of Table 2.

	#D	#M	#C	#S
TRAINS-93	98	12148	4523	0.0
PEAR	20	3401	1168	0.0
RST	413	62409	38724	0.0
GNOME	5	5499	2598	0.0
TRAINS-91	16	2501	828	0.0
AMI (DEV)	7	7441	3120	4.0
LIGHT (DEV)	20	3448	1357	2.0
PSUA (DEV)	21	2437	1273	2.0
SWBD (DEV)	11	3421	1771	0.0
AMI (TST)	3	4139	1883	4.0
LIGHT (TST)	21	3501	1359	2.0
PSUA (TST)	28	3446	1857	2.0
SWBD (TST)	22	7847	3897	2.0
ON (TRN)	2802	155558	35142	0.6
ON (DEV)	343	19155	4545	0.8
ON (TST)	348	19764	4532	0.8
BOLT (TRN)	1110	58146	12854	2.5
BOLT (DEV)	137	8029	1649	2.5
BOLT (TST)	137	7599	1610	2.5

Table 2: Statistics for all datasets (Section 4.1), excluding non-referring expressions. *UAD* datasets are shown in the upper part, and *OD* datasets shown in the bottom part. TRN/DEV/TST: the train/dev/test split. #D: total number of documents; #M: total number of mentions; #C: total number of clusters; #S: averaged number of speakers per document (excluding unknown speakers).

4.2 Preprocessing

We only perform one trivial preprocessing step specific to the training set of *UAD* datasets: remove all non-referring expressions and regard them as non-mentions, as they will not be counted in the final evaluation (Section 5). In addition, our current approach does not consider split-antecedents, which we will leave as future work.

4.3 Implementation

Our system is based on the PyTorch implementation of the end-to-end coreference resolution model from Xu and Choi (2020), and we follow the similar hyperparameter settings. Specifically, SpanBERT$_{\text{Large}}$ (Joshi et al., 2020) is used as the Transformers encoder with maximum sequence length of 512. Long documents are split into multiple sequences, and each sequence is encoded by SpanBERT$_{\text{Large}}$ independently, as suggested by (Joshi et al., 2019). During training, we limit the maximum sequences to be 3 due to the GPU memory constraints, and a long document will be truncated into multiple documents if it exceeds the maximum sequences.

Hyperparameters For all datasets, nested mentions are always enabled. We set the $\lambda = 0.5$ and maximum span width to be 30 in the span enumer-

	AMI				*LIGHT*			
	MUC	B^3	CEAF$_{\phi_4}$	Avg F1	MUC	B^3	CEAF$_{\phi_4}$	Avg F1
MR	46.06	31.28	17.87	31.73	79.47	48.61	24.06	50.71
SR	54.66	53.32	53.64	53.87	79.12	63.34	65.59	69.35
SR^{+P}	50.08	52.63	52.60	51.77	86.39	74.68	**69.57**	76.88
SE	57.01	53.91	53.65	54.86	77.38	63.91	65.46	68.92
SE^{+M}	**59.36**	48.49	40.05	49.30	86.92	68.49	41.66	65.69
SE^{+P}	56.93	**55.27**	**53.92**	**55.37**	**87.23**	**74.91**	68.86	**77.00**
SE^{+P}+DEV	70.70	61.37	59.81	63.96	90.03	79.19	71.77	80.33

(a) Evaluation results on the test set of *AMI* and *LIGHT*.

	PSUA				*SWBD*			
	MUC	B^3	CEAF$_{\phi_4}$	Avg F1	MUC	B^3	CEAF$_{\phi_4}$	Avg F1
MR	75.15	50.80	21.78	49.24	72.70	46.65	22.54	47.30
SR	73.36	67.97	64.15	68.49	73.92	62.38	58.01	64.77
SR^{+P}	78.96	73.83	65.27	72.69	75.30	65.16	57.69	66.05
SE	72.99	68.56	64.37	68.64	74.47	63.32	58.77	65.52
SE^{+M}	81.63	65.82	44.83	64.10	76.54	61.43	43.97	60.65
SE^{+P}	**82.19**	**76.50**	**67.46**	**75.38**	**77.56**	**67.56**	**59.36**	**68.16**
SE^{+P}+DEV	84.04	79.57	71.63	78.41	80.63	74.39	68.45	74.49

(b) Evaluation results on the test set of *Persuasion for Good* (*PSUA*) and *Switchboard* (*SWBD*).

Table 3: Evaluation results on the test set of four datasets (Section 4.1). The macro-averaged F1 of MUC, B^3, and CEAF$_{\phi_4}$ is the main evaluation metric. Section 3 describes the details of all listed approaches. SE^{+P}+DEV is the setting of our final submission to the CRAC 2021 shared task, where all available development sets are also added in the training process for SE^{+P} (Section 5.1).

Track	Resolution of anaphoric identities
Setting	Predicted mentions
Baseline	MR (§3.1). The end-to-end coreference resolution model with the SpanBERT encoder (Joshi et al., 2020; Xu and Choi, 2020) is used as the baseline.
Approach	SE^{+P}+DEV (§3.4). The final model is built upon baseline with three key adaptations: 1) An updated antecedent selection process is used to support singletons, with an additional optimization on the mention scores. 2) Speaker-augmentation strategy is used to encode the speakers and dialogue-turns. 3) Knowledge transfer is employed that pretrains the model on CoNLL datasets, then further trains on the UA datasets as a domain adaptation step. The final submission includes the dev data into training.
Train Data	*TRAINS-93, PEAR, RST, GNOME, ON, BOLT* (§4.1)
Dev Data	*TRAINS-91, AMI, LIGHT, PSUA, SWBD, ON, BOLT* (§4.1)

Table 4: Summary of our final submission to the CRAC 2021 shared task. Train/Dev Data: all datasets we use for the training set and development set.

ation stage, and limit the maximum antecedents to be 50 in the pair scoring process. Adam optimizer is used for the optimization, with the weight decay rate of 10^{-2} and gradient clipping norm of 1. We employ the learning rate of 1×10^{-5} for Transformers parameters, and 3×10^{-4} for task parameters. $\alpha_m = 0.1$ is used for Eq (4). In particular, we do not apply any higher-order inferences, as their benefits are shown trivial (Xu and Choi, 2020).

Training When training *UAD* or *OD* alone, we concatenate and mix all its corresponding corpora together as the training data. For SE^{+M}, we concatenate and mix all available training corpora together regardless of *UAD* or *OD*. All experiments are conducted on a Nvidia A100 GPU. 20 training epochs are used for all the settings, and the training takes around 1-2 hours for *UAD* and 3-4 hours for *OD*.

In particular, development sets are not added to the training data, except for our final submission to the shared task, where the best-performed model has been identified, then we train the final model with the same setting but adding all development sets in the training (Section 5.1).

5 Results and Analysis

The Universal Anaphora Scorer[1] is used in the official evaluation process. For the task of anaphora resolution, the main evaluation metric is the averaged F1 score of MUC, B^3 and CEAF$_{\phi_4}$, same as the CoNLL 2012 shared task. Singletons and split-antecedents are included in the evaluation, while non-referring expressions are excluded.

5.1 Results

Table 3 shows the evaluation results on the test set of four datasets using different approaches. Among all approaches without adding the dev sets into training, SE^{+P} achieves the best results on all four datasets. Another SE^{+P} model is then trained with adding the dev sets as our final submission, dentoed by SE^{+P}+DEV, which further yields the best results, and ranks the 1st place at the "anaphoric identity" track in the CRAC 2021 shared task.

Final Submission Table 4 lists the summary of our final submission to the shared task.

5.2 Analysis: Singleton Recognition

One of the main differences between the UA and CoNLL format is that UA supports singletons, as UA annotates all noun phrases. The left side of Table 6 shows the total number and percentage of the singleton clusters on the test set of four datasets. Singletons are indeed prevalent, and all four datasets have at least 73% of their gold clusters as singletons. Therefore, recognizing singletons can become critical for coreference resolution on the UA formatted data.

	Mentions			*Singletons*		
	P	R	F	P	R	F
MR	**90.3**	40.6	56.1	-	-	-
SR	84.7	76.4	80.3	44.7	54.8	49.2
SR^{+P}	83.1	74.3	78.5	42.8	54.0	47.8
SE	83.4	77.7	80.4	43.1	**55.9**	48.7
SE^{+M}	81.5	69.5	75.0	29.8	35.7	32.5
SE^{+P}	84.1	**77.9**	**80.9**	**45.8**	53.5	**49.4**

(a) Statistics on the test set of *AMI*.

	Mentions			*Singletons*		
	P	R	F	P	R	F
MR	**97.1**	60.0	74.2	-	-	-
SR	87.7	86.8	87.2	56.3	72.1	63.2
SR^{+P}	90.1	89.4	89.7	**65.1**	68.3	**66.6**
SE	87.6	86.3	87.0	55.0	**73.6**	62.9
SE^{+M}	91.4	71.7	80.3	43.1	23.9	30.8
SE^{+P}	90.0	**89.6**	**89.8**	62.1	68.4	65.1

(b) Statistics on the test set of *LIGHT*.

	Mentions			*Singletons*		
	P	R	F	P	R	F
MR	**94.9**	56.8	71.1	-	-	-
SR	89.5	85.4	87.4	66.2	55.4	60.3
SR^{+P}	91.8	86.4	89.0	74.3	51.5	60.8
SE	88.8	86.0	87.4	64.6	**57.2**	60.7
SE^{+M}	90.5	69.9	78.9	53.6	26.0	35.1
SE^{+P}	91.9	**87.4**	**89.6**	**74.8**	54.4	**63.0**

(c) Statistics on the test set of *Persuasion for Good* (*PSUA*).

	Mentions			*Singletons*		
	P	R	F	P	R	F
MR	**92.0**	54.0	68.1	-	-	-
SR	85.7	80.1	82.8	54.0	51.8	52.9
SR^{+P}	86.3	80.3	83.2	52.9	50.5	51.7
SE	85.0	80.6	82.7	53.3	**54.0**	**53.7**
SE^{+M}	86.3	68.9	76.6	39.2	31.7	35.1
SE^{+P}	87.4	**81.0**	**84.1**	**56.9**	50.5	53.5

(d) Statistics on the test set of *Switchboard* (*SWBD*).

Table 5: Statistics of different approaches on the test set of four datasets. The left side shows the Precision/Recall/F1 (P/R/F) of the predicted mentions over gold mentions, and the right side shows the predicted singletons over gold singletons.

Comparing MR and SR in Table 3, it is clear that singleton recognition plays a pivotal role in the final performance, with SR outperforming MR by a huge margin of 17-22 Avg F1 on all four datasets. To further examine the performance of SR, we collect the precision/recall of the predicted mentions by different models, as well as the precision/recall of predicted singletons over gold singletons, as shown in Table 5. Compared with MR, all models that support singletons receive huge gains on the mention recall with 26-36% improvement, with relatively small 5-10% degradation on the mention precision.

[1] https://github.com/juntaoy/universal-anaphora-scorer

	#AC	#SC	#AM	#PM
AMI	1883	1383 (73.5%)	4139	1566 (37.8%)
LIGHT	1359	1024 (75.4%)	3501	1676 (**47.9%**)
PSUA	1857	1525 (**82.1%**)	3446	1464 (42.5%)
SWBD	3897	2968 (76.2%)	7847	3746 (47.7%)

Table 6: Statistics on the test set of all four datasets. `#AC`: total number of all clusters. `#SC`: total number of singleton clusters, with the corresponding percentage indicated inside parentheses. `#AM`: total number of all mentions. `#PM`: total number of personal pronoun mentions, with the percentage inside parentheses. All statistics exclude non-referring expressions.

More interestingly, most `SR`/`SE`-related models are able to recover the majority of gold singletons on all four datasets, up to 73% recall on *LIGHT*, demonstrating the effectiveness of the mention score optimization in Eq (3) and the new antecedent selection process. Nevertheless, the best F1 for singletons is still below 67 out of four datasets, suggesting that resolving singletons alone can be a challenging aspect already.

5.3 Analysis: Speaker Encoding

Despite the simple strategy of speaker-augmented encoding described in Section 3.3, $\texttt{SE}^{\texttt{+P}}$ shows decent improvement over its counterpart $\texttt{SR}^{\texttt{+P}}$, with 2-3% enhancement on Avg F1 on all datasets, except for *LIGHT* that has only trivial improvement, confirming that stronger speaker encoding is indeed important for the dialogue domain.

Meanwhile, `SE` does not show advantages over `SR` due to the fact that the current training corpora of all ARRAU datasets do not provide the speakers (Table 2); consequently, neither models could learn to use the speaker information, resulting in similar performance. This on the other side also demonstrates the significance of knowledge transfer that utilizes other existing resources.

5.4 Analysis: Knowledge Transfer

Comparing the two knowledge transfer strategies, the pretraining paradigm $\texttt{SE}^{\texttt{+P}}$ performs significantly better than the mixing paradigm $\texttt{SE}^{\texttt{+M}}$. In fact, while the pretraining brings improvement over `SE`, the mixing paradigm even performs worse than without knowledge transfer, likely because of the domain mismatch and the annotation format mismatch, showing that the pretraining strategy should always be preferred in this case.

The impact of the pretraining on OD can be dataset-specific, as shown by Table 3. $\texttt{SE}^{\texttt{+P}}$ is able to boost performance upon `SE` by a good margin on *AMI*/*SWBD* with 0.5/2.6 F1 respectively, while *LIGHT*/*PSUA* can benefit significantly, with 6.7/8.1 F1 improvement. Encouraged by the results, we suggest to further explore the utilization of existing resources as a future direction.

6 Conclusion

In this work, we present an adapted end-to-end coreference resolution system for anaphoric identities in dialogues, specifically addressing three aspects: the support for singletons, stronger speaker and turn encoding through the dialogue interactions, as well as the knowledge transfer utilizing other existing resources. Our final system achieves the best results on all four datasets on the leaderboard of the CRAC 2021 shared task, and further analysis is performed to show the effectiveness of our proposed adaptation strategies.

References

Berfin Aktaş and Manfred Stede. 2020. Variation in coreference strategies across genres and production media. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5774–5785, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Kevin Clark and Christopher D. Manning. 2015. Entity-centric coreference resolution with model stacking. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1405–1415, Beijing, China. Association for Computational Linguistics.

Kevin Clark and Christopher D. Manning. 2016. Improving coreference resolution by learning entity-level distributed representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653, Berlin, Germany. Association for Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. 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.

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.

Ran Le, Wenpeng Hu, Mingyue Shang, Zhenjun You, Lidong Bing, Dongyan Zhao, and Rui Yan. 2019. Who is speaking to whom? learning to identify utterance addressee in multi-party conversations. 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 1909–1919, Hong Kong, China. Association for Computational Linguistics.

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. Association for Computational Linguistics.

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, Volume 2 (Short Papers)*, pages 687–692, New Orleans, Louisiana. Association for Computational Linguistics.

Xuansong Li, Martha Palmer, Nianwen Xue, Lance Ramshaw, Mohamed Maamouri, Ann Bies, Kathryn Conger, Stephen Grimes, and Stephanie Strassel. 2016. Large multi-lingual, multi-level and multi-genre annotation corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 906–913, Portorož, Slovenia. European Language Resources Association (ELRA).

Xiaoqiang Luo, Radu Florian, and Todd Ward. 2009. Improving coreference resolution by using conversational metadata. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 201–204, Boulder, Colorado. Association for Computational Linguistics.

Massimo Poesio, Yulia Grishina, Varada Kolhatkar, Nafise Moosavi, Ina Roesiger, Adam Roussel, Fabian Simonjetz, Alexandra Uma, Olga Uryupina, Juntao Yu, and Heike Zinsmeister. 2018. Anaphora resolution with the ARRAU corpus. In *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 11–22, New Orleans, Louisiana. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Sam Wiseman, Alexander M. Rush, Stuart Shieber, and Jason Weston. 2015. Learning anaphoricity and antecedent ranking features for coreference resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1416–1426, Beijing, China. Association for Computational Linguistics.

Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li. 2020. CorefQA: Coreference resolution as query-based span prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6953–6963, Online. Association for Computational Linguistics.

Patrick Xia, João Sedoc, and Benjamin Van Durme. 2020. Incremental neural coreference resolution in constant memory. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8617–8624, Online. 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.

Juntao Yu, Alexandra Uma, and Massimo Poesio. 2020. A cluster ranking model for full anaphora resolution. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 11–20, Marseille, France. European Language Resources Association.

Klim Zaporojets, Johannes Deleu, Chris Develder, and Thomas Demeester. 2021. Dwie: An entity-centric dataset for multi-task document-level information extraction. *Information Processing & Management*, 58(4):102563.

Ethan Zhou and Jinho D. Choi. 2018. They exist! introducing plural mentions to coreference resolution and entity linking. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 24–34, Santa Fe, New Mexico, USA. Association for Computational Linguistics.