




Neural Anaphora Resolution in Dialogue Revisited

Shengjie Li, Hideo Kobayashi, and Vincent Ng
Human Language Technology Research Institute
The University of Texas at Dallas



Plan for the Talk

- **Anaphora Resolution**
 - Approach
 - Evaluation
- Discourse Deixis Resolution
 - Approach
 - Evaluation
- End-to-End Neural Bridging Resolution
 - Approach
 - Evaluation
- Q&A

Anaphora Resolution: Overview of *UTD_NLP*²⁰²¹

- Baseline: *UTD_NLP*²⁰²¹, which extends *coref-hoi* (Xu and Choi, 2020) as follows:
 - Sentence distance as another feature
 - Mention type prediction model (jointly trained with the coreference model)
 - Loss is a weighted combination of type prediction loss and coreference loss
 - Constraints
 - Span constraint: a span cannot cover more than one speaker's utterance
 - Eight resolution constraints: each one specifies whether two spans can be coreferent depends on the groups they belong to and their speaker(s)
 - E.g. "I" and "me" cannot be coreferent if they have different speakers

3-step Pipelined Approach

- Step 1: Mention Extraction
- Step 2: Coreference Resolution
- Step 3: Non-referring/Non-entity Mention Removal

3-step Pipelined Approach

- Step 1: Mention Extraction

- *UTD_NLP*²⁰²¹
- Type prediction:
 - ENTITY (referring/non-referring spans),
 - NULL (non-entity spans)
- Use a large type prediction coefficient in the model loss so that *UTD_NLP*²⁰²¹ focuses on type prediction rather than anaphora resolution
- Pretrained on OntoNotes 5.0

3-step Pipelined Approach

- Step 1: Mention Extraction
- Step 2: Coreference Resolution
 - Trained on Gold entity mentions; tested on mentions predicted as ENTITY in Step 1
 - Changes to *UTD_NLP*²⁰²¹:
 - Removing the type prediction model
 - Removing the mention scores in the pairwise scores ($s_m(\cdot)$ indicates how likely a span corresponds to an entity mention)

$$s(x, y) = \cancel{s_m(x)} + \cancel{s_m(y)} + s_c(x, y) + s_a(x, y)$$

- Inference-time-only dummy antecedent rescoring
- $s(x, \epsilon) = c \quad (c > 0)$
- Pretrained on OntoNotes 5.0

3-step Pipelined Approach

- Step 1: Mention Extraction
- Step 2: Coreference Resolution
- Step 3: Non-referring/Non-entity Mention Removal
 - *UTD_NLP²⁰²¹*
 - Trained on:
 - gold entity mentions
 - gold non-referring mentions
 - entity mentions in which the underlying word/phrase has appeared at least once as a gold entity mention in the training data
 - Type prediction:
 - REFERRING (referring spans),
 - OTHER (non-referring/non-entity spans)
 - Singletons in the output of Step 2 that are predicted as OTHER are removed.

Evaluation Results

CoNLL scores on the four test sets

	LIGHT	AMI	Persuasion	Switchboard
S1	78.52	59.56	76.43	72.42
S1, S2	79.01	60.64	76.81	71.68
S1, S3	81.40	61.51	78.69	75.81
S1, S2, S3	82.23	62.90	79.20	75.25

- S1: our model without the last two steps.
- S1, S2: our model without the third step.
- S1, S3: output from Step 1 is post-processed by the third-step model to remove non-entity and non-referring mentions.
- S1, S2, S3: our full model.

Evaluation Results

CoNLL scores on the four test sets

	LIGHT	AMI	Persuasion	Switchboard
S1	78.52	59.56	76.43	72.42
S1, S2	79.01	60.64	76.81	71.68
S1, S3	81.40	61.51	78.69	75.81
S1, S2, S3	82.23	62.90	79.20	75.25

- Key takeaways:
 - Mention scores don't have a great impact in generating coreference links, if we have a type prediction model
 - Non-entity/non-referring mentions have a great impact on the performance of our systems
 - Our prelim experiments show that pretraining is helpful for our first-step and second-step model. We did not pretrain our third-step model because OntoNotes covers only a portion of non-referring expressions

Possible Improvements

- Handling long dependencies
 - Documents in the AMI dataset can have more than 8,000 tokens and our system perform significantly worse on AMI than on any other datasets. We hypothesize our systems have a hard time handling long dependencies.
- Handling split antecedents
 - Our system cannot handle cases of plural anaphoric reference in which the antecedents are introduced by separate mentions

Plan for the Talk

- Anaphora Resolution
 - Approach
 - Evaluation
- **Discourse Deixis Resolution**
 - Approach
 - Evaluation
- End-to-End Neural Bridging Resolution
 - Approach
 - Evaluation
- Q&A

Discourse Deixis Resolution: Overview

Three-phase evaluation:

- **Predicted** Phase - Nothing is given; model needs to:
 - Identify antecedent mentions
 - Identify anaphor mentions
 - Perform discourse deixis resolution
- **Gold Mention** Phase - Given gold entity mentions, model needs to:
 - Identify antecedent mentions
 - Identify anaphor mentions **among gold entity mentions**
 - Perform discourse deixis resolution
- **Gold Anaphor** Phase - Given gold anaphor mentions, model needs to:
 - Identify antecedent mentions
 - Perform discourse deixis resolution

Discourse Deixis Resolution: Predicted Phase

- Approach: *coref-hoi* (Xu and Choi, 2020) extended with:
 - Candidate Anaphor Extraction: words/phrases that appeared at least once as anaphor in the training data
 - Anaphor Prediction: ANAPHOR and NULL
 - Candidate Antecedent Extraction: 10 Closest utterances as antecedents
 - Dummy Antecedent Elimination
 - Additional features
 - Anaphor-based features
 - Antecedent-based features
 - Pairwise features
 - Inference-Time-Only Distance-Based Candidate Antecedent Filtering: n closest utterances ($1 \leq n \leq 10$) during inference

Discourse Deixis Resolution: Gold Phases

- Gold Mention Phase
 - Candidate Anaphor Extraction: Extract from the set of given gold mentions.

- Gold Anaphor Phase
 - Candidate Anaphors: The given set of gold anaphors.
 - We removed the Anaphor Prediction model since there's no need.

Evaluation Results and Discussion

CoNLL scores on the four test sets

	Light	AMI	Persuasion	Switchboard
Predicted Phase	37.09	53.31	54.59	49.76
Gold Mention Phase	38.38	55.12	54.89	49.83
Gold Anaphor Phase	52.40	72.50	69.61	72.11

- Only a small performance gain is achieved in Gold Mention phase
- The provision of gold anaphors has brought huge improvements (14%-22% CoNLL score)
- Our system performs much worse on LIGHT than on other datasets.

Evaluation Results and Discussion

CoNLL scores on the four test sets

	Light	AMI	Persuasion	Switchboard
Predicted Phase	37.09	53.31	54.59	49.76
Gold Mention Phase	38.38	55.12	54.89	49.83
Gold Anaphor Phase	52.40	72.50	69.61	72.11

- Key takeaways:
 - One of the key weaknesses of our system is anaphor identification.
 - We ran some ablation experiments which shows that these factors play an important role in DD resolution:
 - Recency between antecedent and anaphor
 - Dummy Antecedent Elimination

Plan for the Talk

- Anaphora Resolution
 - Approach
 - Evaluation
- Discourse Deixis Resolution
 - Approach
 - Evaluation
- **End-to-End Neural Bridging Resolution**
 - Approach
 - Evaluation
- Q&A

Bridging Resolution: Predicted Phase

- Approach: Yu and Poesio's (Y&P) span-based model (2020) extended with:
 - Using SpanBERT as encoder
 - Y&P uses bi-LSTM and frozen BERT/GloVe embeddings
 - Adding Turn Distance as a feature
 - Y&P is not designed for the dialogue domain. It has only two features: the length of a mention and the mention-pair distance
 - We add the turn distance between mentions as a feature, where a turn is defined as a set of contiguous sentences by the same speaker
 - Using our S1 system in the AR track as a mention extractor
 - Y&P performs bridging resolution on gold mentions

Bridging Resolution: Gold Phases

- Gold Mention Phase
 - We applied our models trained in the Predicted phase to the given set of gold mentions

- Gold Anaphor Phase
 - We constraint our models so that only gold anaphors can be resolved to other gold mentions

Evaluation Results and Discussion

Resolution F-scores on the four test sets

	Light	AMI	Persuasion	Switchboard
Predicted Phase	23.25	13.42	27.75	19.72
Gold Mention Phase	26.77	19.65	34.59	22.74
Gold Anaphor Phase	46.80	39.35	56.91	44.40

- The performance gains we achieve in the Gold Mention phase over the Predicted phase can be attributed solely to the difference between using predicted mentions and using gold mentions.
- Although gold mentions are given in the Gold Mention phase, Identifying bridging anaphors is still a non-trivial task.
- Our system performs much better in the Gold Anaphor phase.

Evaluation Results and Discussion

Resolution F-scores on the four test sets

	Light	AMI	Persuasion	Switchboard
Predicted Phase	23.25	13.42	27.75	19.72
Gold Mention Phase	26.77	19.65	34.59	22.74
Gold Anaphor Phase	46.80	39.35	56.91	44.40

- Key takeaways:
 - The number of training epochs has a large impact on the performance of our bridging resolver
 - Different setups of training data have an even large impact:
 - Trained on all shared-task datasets
 - Pretrained on datasets outside of target domain, then fine tuned on target datasets
 - Pretrained on datasets outside of target domain, then fine tuned on a specific target dataset
 - Pretrained on all shared-task datasets, then fine tuned on a specific target dataset

Concluding Remarks

- We participated and ranked first in all three tracks of the shared task
- Our models are built upon state-of-the-art span-based neural models
 - Anaphora Resolution: *UTD_NLP²⁰²¹*, which extends Xu and Choi's (2020) *coref-hoi* model
 - Discourse Deixis Resolution: Xu and Choi's (2020) *coref-hoi* model
 - Bridging Resolution: Yu and Poesio's (2020) MTL approach

Thank You!