Bridging Resolution: A Journey Towards Modelling Referential Discourse Entities

Yufang Hou

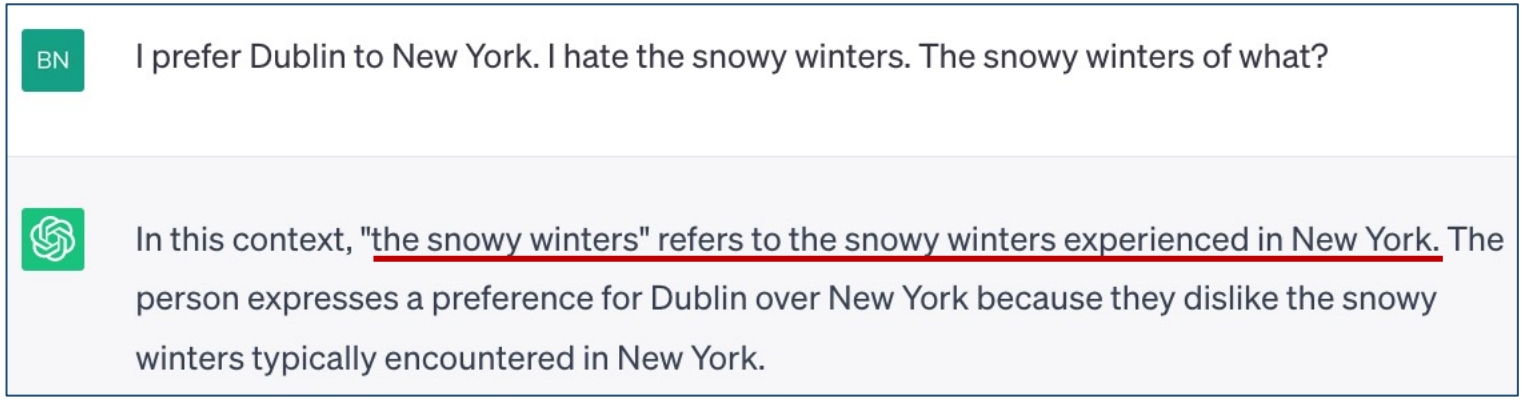
@yufanghou

https://yufanghou.github.io/

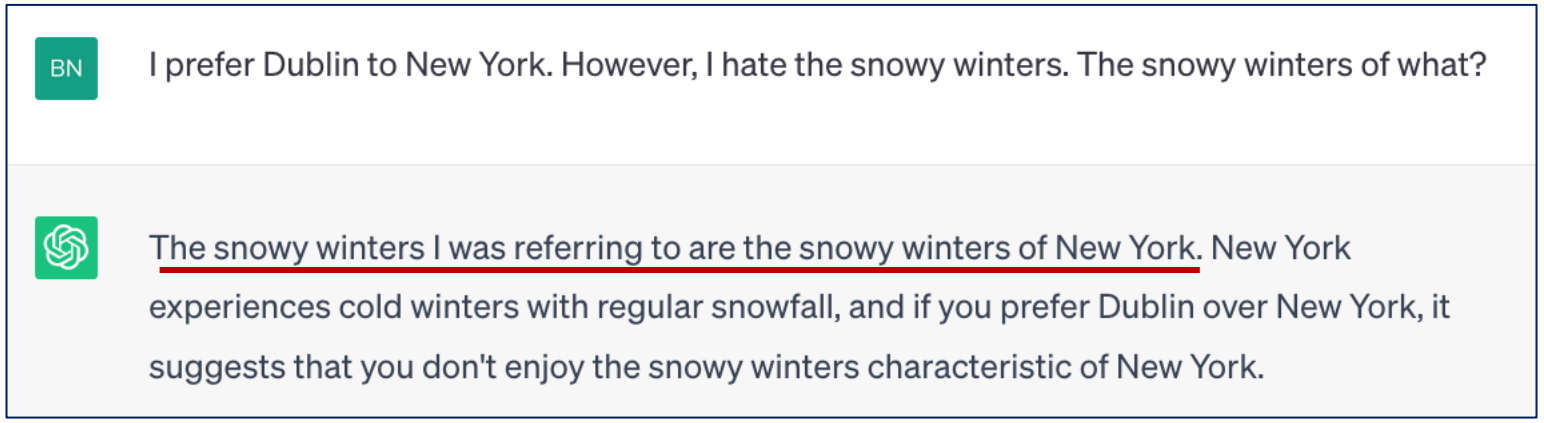
July 13th, 2023

Bridging Examples

- I prefer Dublin to **New York**. I hate **the snowy winters**.



- I prefer **Dublin** to New York. However, I hate **the snowy winters**.



Outline

- Information Status: A Model to Understand Discourse Entities
- Models for Bridging Resolution
- Probing LLMs for Bridging Inference
- Some Thoughts on Future Work

From Text to Discourse Entities

Information Status: reflects the accessibility of a discourse entity based on the evolving discourse context and the speaker's assumption about the hearer's knowledge and beliefs.

S1: The Bakersfield Supermarket went bankrupt last May.

S2: The business located in northern Manhattan closed when its owner was murdered.

S3: Friends expressed outrage at the murder.

S4: Unfortunately, such crimes are not unusual.

From Text to Discourse Entities

Discourse introduces new entities.

S1: **[The Bakersfield Supermarket]**$_{new}$ went bankrupt last May.

S2: The business located in northern Manhattan closed when its owner was murdered.

S3: Friends expressed outrage at the murder.

S4: Unfortunately, such crimes are not unusual.

From Text to Discourse Entities

Discourse refers back to already known entities.

S1: [**The Bakersfield Supermarket**]$_{new}$ went bankrupt last May.

S2: [**The business located in northern Manhattan**]$_{old}$ closed when [**its**]$_{old}$ owner was murdered.

S3: Friends expressed outrage at the murder.

S4: Unfortunately, such crimes are not unusual.

From Text to Discourse Entities

Some entities are accessible via other entities introduced before.

S1: [**The Bakersfield Supermarket**]$_{new}$ went bankrupt last May.

S2: [**The business located in northern Manhattan**]$_{old}$ closed when [**its**]$_{old}$ owner was murdered.

S3: [**Friends**]$_{bridging}$ expressed outrage at the murder.

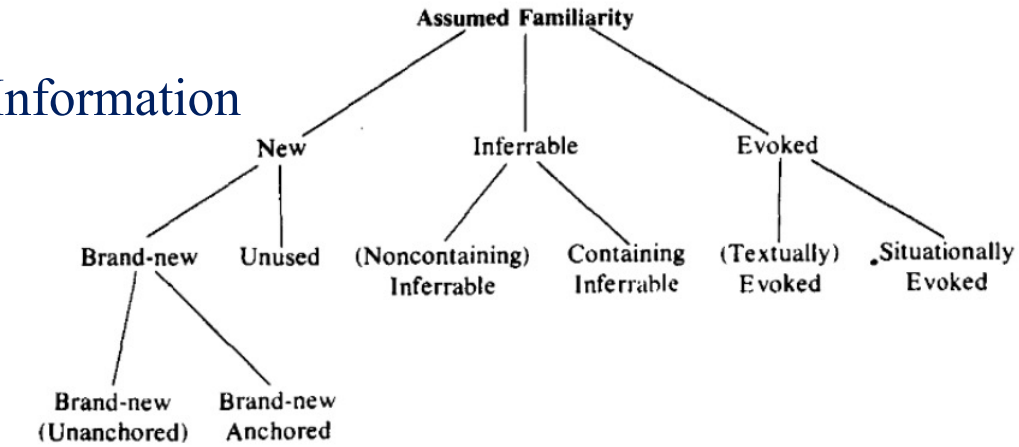
S4: Unfortunately, such crimes are not unusual.

Bridging: Discourse new, hearer old

From Text to Discourse Entities

Information Status Schemes

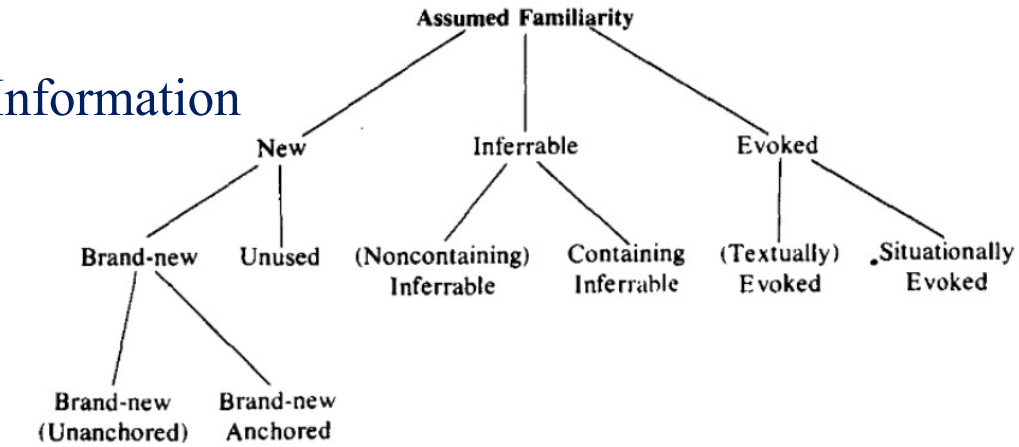
- Prince (1981; 1992): Towards a Taxonomy of Given-New Information



From Text to Discourse Entities

Information Status Schemes

- Prince (1981; 1992): Towards a Taxonomy of Given-New Information



- Markert et al. (2012): An IS scheme for written text
  - Based on Prince (1981; 1992) and Nissim et al. (2004)

Old	Old
Mediated	World Knowledge
	Syntactic
	Comparative
	Bridging
	Aggregation
	Function
New	New

From Text to Discourse Entities

Information Status Scheme [Markert et al. (2012)]

Old	Old
Mediated	World Knowledge
	Syntactic
	Comparative
	Bridging
	Aggregation
	Function
New	New

S1: [The Bakersfield Supermarket]$_{new}$ went bankrupt [last May]$_{med/knowledge}$.

S2: [The business located in [northern Manhattan]$_{med/knowledge}$]$_{old}$ closed when [[its]$_{old}$ owner]$_{med/syntactic}$ was murdered.

S3: [Friends]$_{med/bridging}$ expressed [outrage]$_{new}$ at [the murder]$_{old}$.

S4: Unfortunately, [such crimes]$_{med/comparative}$ are not unusual.

From Text to Discourse Entities

Information Status Scheme [Markert et al. (2012)]

Old	Old
Mediated	World Knowledge
	Syntactic
	Comparative
	Bridging
	Aggregation
	Function
New	New

S1: [The Bakersfield Supermarket]$_{new}$ went bankrupt [last May]$_{med/knowledge}$.

S2: [The business located in [northern Manhattan]$_{med/knowledge}$]$_{old}$ closed when [[its]$_{old}$ owner]$_{med/syntactic}$ was murdered.

S3: [Friends]$_{med/bridging}$ expressed [outrage]$_{new}$ at [the murder]$_{old}$.

S4: Unfortunately, [such crimes]$_{med/comparative}$ are not unusual.

Bridging Anaphora

- ✓ Establishes entity coherence in a text by linking anaphors and antecedents via various non-identity relations
- ✓ Tasks
  - Bridging Anaphora Recognition
  - Bridging Anaphora Resolution
  - Full Bridging Resolution

Outline

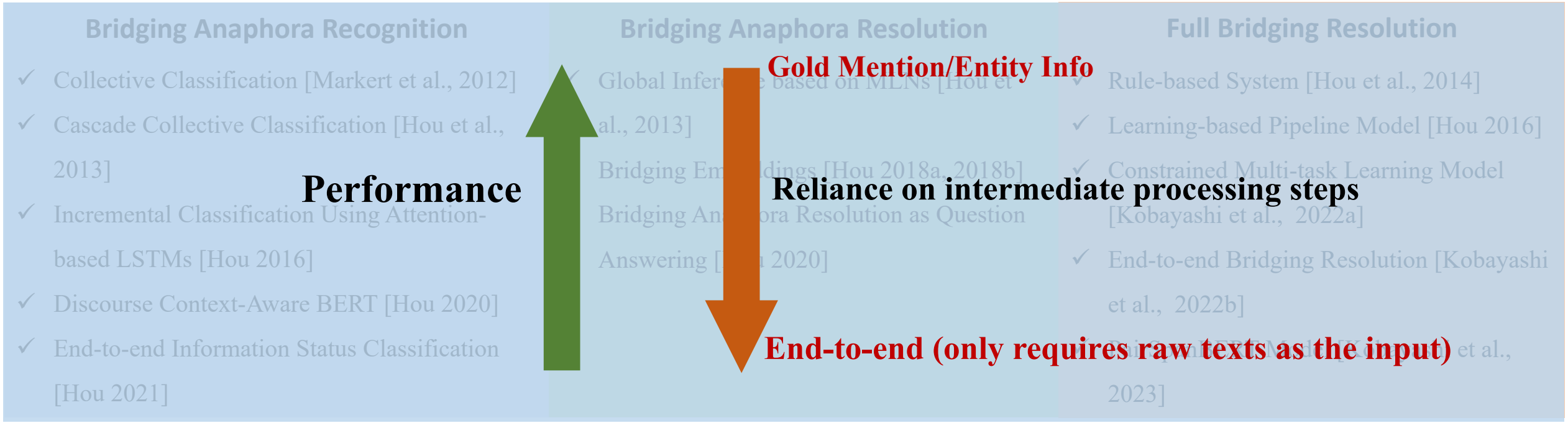
- Information Status: A Model to Understand Discourse Entities
- Models for Bridging Resolution
- Probing LLMs for Bridging Inference
- Some Thoughts on Future Work

Models for Bridging Resolution

Bridging Anaphora Recognition	Bridging Anaphora Resolution	Full Bridging Resolution
✓ Collective Classification [Markert et al., 2012]	✓ Global Inference based on MLNs [Hou et al., 2013]	✓ Rule-based System [Hou et al., 2014]
✓ Cascade Collective Classification [Hou et al., 2013]	✓ Bridging Embeddings [Hou 2018a, 2018b]	✓ Learning-based Pipeline Model [Hou 2016]
✓ Incremental Classification Using Attention-based LSTMs [Hou 2016]	✓ Bridging Anaphora Resolution as Question Answering [Hou 2020]	✓ Constrained Multi-task Learning Model [Kobayashi et al., 2022a]
✓ Discourse Context-Aware BERT [Hou 2020]		✓ End-to-end Bridging Resolution [Kobayashi et al., 2022b]
✓ End-to-end Information Status Classification [Hou 2021]		✓ PairSpanBERT Model [Kobayashi et al., 2023]

https://github.com/IBM/bridging-resolution

Models for Bridging Resolution



https://github.com/IBM/bridging-resolution

Models for Bridging Resolution

Bridging Anaphora Recognition	Bridging Anaphora Resolution	Full Bridging Resolution
✓ Collective Classification [Markert et al., 2012]	✓ Global Inference based on MLNs [Hou et al., 2013]	✓ Rule-based System [Hou et al., 2014]
✓ Cascade Collective Classification [Hou et al., 2013]	✓ Bridging Embeddings [Hou 2018a, 2018b]	✓ Learning-based Pipeline Model [Hou 2016]
✓ Incremental Classification Using Attention-based LSTMs [Hou 2016]	✓ Bridging Anaphora Resolution as Question Answering [Hou 2020]	✓ Constrained Multi-task Learning Model [Kobayashi et al., 2022a]
✓ Discourse Context-Aware BERT [Hou 2020]		✓ End-to-end Bridging Resolution [Kobayashi et al., 2022b]
✓ End-to-end Information Status Classification [Hou 2021]		✓ PairSpanBERT Model [Kobayashi et al., 2023]

https://github.com/IBM/bridging-resolution

End-to-end Information Status Classification (EMNLP Findings 21)

- **Extract mentions and determine the information status for each mention in a raw text**

Input	Output
S1: The Bakersfield Supermarket went bankrupt last May.	S1: [The Bakersfield Supermarket]$_{new}$ went bankrupt [last May]$_{med/know}$.
S2: The business located in northern Manhattan closed when its owner was murdered.	S2: [The business located in [northern Manhattan]$_{med/know}$]$_{old}$ closed when [[its]$_{old}$ owner]$_{med/syn}$ was murdered.
S3: Friends expressed outrage at the murder.	S3: [Friends]$_{med/bridging}$ expressed [outrage]$_{new}$ at [the murder]$_{old}$.
S4: Unfortunately, such crimes are not unusual.	S4: Unfortunately, [such crimes]$_{med/comparative}$ are not unusual.

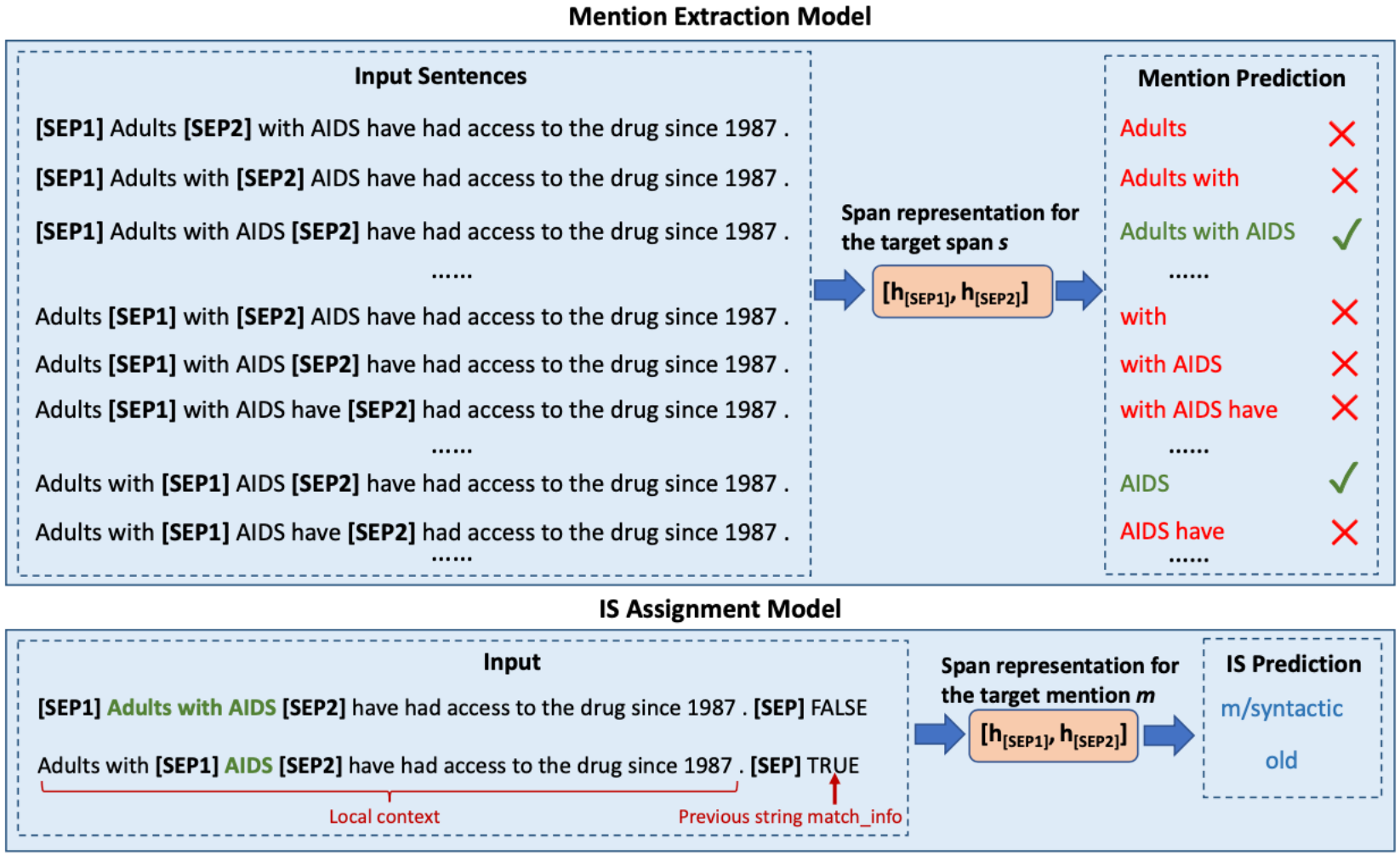
End-to-end Information Status Classification

- **Extract mentions and determine the information status for each mention in a raw text**

Input	Output
S1: The Bakersfield Supermarket went bankrupt last May.	S1: [The Bakersfield Supermarket]$_{new}$ went bankrupt [last May]$_{med/know}$.
S2: The business located in northern Manhattan closed when its owner was murdered.	S2: [The business located in [northern Manhattan]$_{med/know}$]$_{old}$ closed when [[its]$_{old}$ owner]$_{med/syn}$ was murdered.
S3: Friends expressed outrage at the murder.	S3: [Friends]$_{med/bridging}$ expressed [outrage]$_{new}$ at [the murder]$_{old}$.
S4: Unfortunately, such crimes are not unusual.	S4: Unfortunately, [such crimes]$_{med/comparative}$ are not unusual.

- **Compared to mention extraction on coreference resolution (CR)**
  - ✓ **The mention extraction component on CR normally focuses on identifying non-singleton mentions**
  - ✓ **We extract all singleton as well as non-singleton mentions and assign information status to them**
  - ✓ **We aim to identify referential bridging anaphors in an end-to-end setting**

System Architecture



Training: all spans up to L words (L=10)

Inference: **all spans**

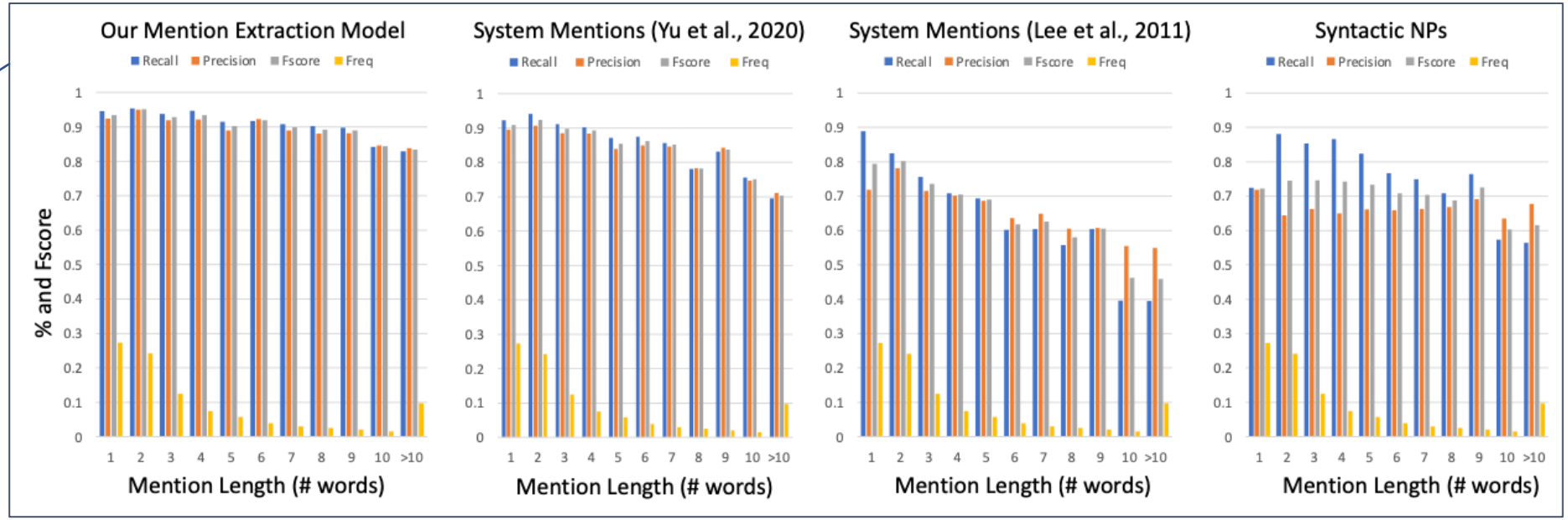
- ✓ Time complexity $O(n^2)$
- ✓ For a sentence with 100 words
  - 5 times slower than the pruning with L=10
  - 1.7 times slower than the pruning with L=30
- ✓ Can be speed up using a dictionary to filter out spans starting with prepositions, verbs, or punctuations

Experiments on ISNotes: Mention Extraction

- 10,980 mentions, 50 news texts
- 10-fold cross-validation on documents

	R	P	F
Baselines			
syntactic NPs	78.0	67.1	72.1
Lee et al. (2011)	74.2	70.8	72.5
Yu et al. (2020)	88.7	86.6	87.7
Our model			
our model (test $L = 10$)	83.5	**92.4**	87.7
our model	**92.8**	91.5	**92.1**

Our mention extraction model performs significantly better than the three baseline models for all length groups



Experiments on ISNotes: IS Classification on Gold Mentions

- 10,980 mentions, 50 news texts
- 10-fold cross-validation on documents

	Local F	*Collective* F	*Cascading Collective* F
old	84.4	85.2	84.4
med/worldKnowledge	72.4	71.4	71.9
med/syntactic	69.3	81.4	81.3
med/aggregate	62.5	74.8	74.3
med/function	68.5	69.8	70.8
med/comparative	82.9	83.1	83.1
med/bridging	33.5	34.1	**46.1**
new	76.6	79.9	79.5
acc	75.5	78.9	78.4

- Cascade Collective Classification [Hou et al., 2018]
- 34 lexical/semantic/discourse structure features
- The model is optimized to recognize bridging anaphors

	baseline *self-attention with* $BERT_{LARGE}$			*this work* *baseline with* $RoBERTa_{LARGE}$			*our model* $RoBERTa_{LARGE}$		
	R	P	F	R	P	F	R	P	F
old	88.4	90.0	89.2	89.0	92.0	**90.5**	88.8	91.8	90.3
m/worldKnow.	77.7	79.5	78.6	78.0	80.9	**79.4**	79.2	79.4	79.3
m/syntactic	83.7	81.1	82.4	85.7	81.8	83.7	84.7	83.5	**84.1**
m/aggregate	80.1	79.3	**79.7**	77.7	75.9	76.8	76.8	77.5	77.1
m/function	73.4	85.5	79.0	71.9	80.7	76.0	90.6	81.7	**85.9**
m/comparative	90.5	86.7	**88.6**	78.7	85.4	81.9	87.7	88.1	87.9
m/bridging	51.0	54.5	52.7	47.8	53.5	50.5	54.1	59.9	**56.9**
new	86.6	85.2	85.9	88.2	84.9	86.5	88.8	85.8	**87.3**
acc		83.7			84.3			**85.1**	

- End-to-end IS classification [Hou 2021]
- Mention boundary embeddings based on RoBERTa + 1 simple lexical feature
- No special treatment for bridging

Experiments on ISNotes: End-to-end IS Classification

- 10,980 mentions, 50 news texts
- 10-fold cross-validation on documents

optimized to identify non-singletons (e.g., old, m/worldKnow.) for CR

	NPs from predicted syntactic trees			*baselines* *system mentions* *Lee et al. (2011)*			*system mentions* *Yu et al. (2020)*			*this work* *our mention extraction model*		
	R	P	F	R	P	F	R	P	F	R	P	F
old	66.4	77.3	71.4	79.8	81.4	80.6	83.2	86.6	84.9	85.0	89.1	**87.0**
m/worldKnow.	49.8	58.9	54.0	65.0	61.3	63.1	72.4	69.3	70.8	74.9	70.8	**72.8**
m/syntactic	67.0	53.9	59.8	57.0	60.8	58.9	77.4	70.2	73.6	80.4	76.3	**78.3**
m/aggregate	60.7	61.5	61.1	28.0	31.9	29.8	61.1	63.2	62.2	69.7	72.1	**70.8**
m/function	76.6	63.6	69.5	70.3	50.6	58.8	73.4	73.4	73.4	81.2	81.2	**81.2**
m/comparative	73.5	56.9	64.1	52.2	55.2	53.7	74.3	72.3	73.3	82.2	79.7	**80.9**
m/bridging	50.2	42.1	45.8	42.2	44.5	43.3	51.6	53.1	52.3	51.3	53.5	**52.4**
new	70.8	49.1	58.0	58.2	49.8	53.7	74.1	70.6	72.3	79.0	75.4	**77.1**
overall	65.8	56.7	60.9	63.4	60.5	61.9	75.5	73.8	74.6	78.9	77.8	**78.3**

our proposed system outperforms the three baselines for all IS categories

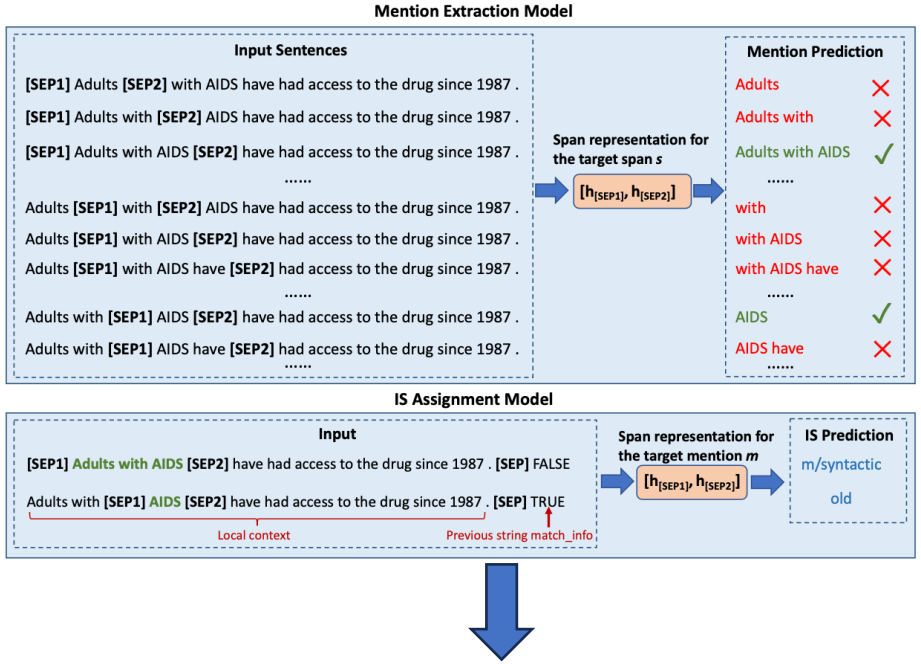
Experiments on BASHI and SciCorp: Bridging Anaphora Recognition

	R	P	F
BASHI			
Yu and Poesio (2020)	34.2	34.4	34.3
our system	59.7	25.5	**35.8**
SciCorp			
Yu and Poesio (2020)	35.7	45.0	39.8
our system	47.6	36.0	**41.0**

- ✓ Yu and Poesio (2020): based on the models trained on the in-domain data using 10-fold CV
- ✓ Our system: trained on ISNotes

- Some predicted bridging anaphors in genetics and computational linguistic scientific papers from SciCorp
  - the underlying siRNA
  - the target mRNA
  - the previous optimization
  - the objective function
  - the most predictive features
- It seems that our IS assignment model can capture some of the bridging referential patterns and generalize them into different domains

Conclusions



- We propose a simple and effective model for fine-grained IS classification in the end-to-end setting.
- Our system achieves strong results for both mention extraction and IS classification compared to other baselines on ISNotes.
- We demonstrate that our system trained on ISNotes can be applied to identify bridging anaphors in different domains.

Models for Bridging Resolution

Bridging Anaphora Recognition	Bridging Anaphora Resolution	Full Bridging Resolution
✓ Collective Classification [Markert et al., 2012]	✓ Global Inference based on MLNs [Hou et al., 2013]	✓ Rule-based System [Hou et al., 2014]
✓ Cascade Collective Classification [Hou et al., 2013]	✓ Bridging Embeddings [Hou 2018a, 2018b]	✓ Learning-based Pipeline Model [Hou 2016]
✓ Incremental Classification Using Attention-based LSTMs [Hou 2016]	✓ Bridging Anaphora Resolution as Question Answering [Hou 2020]	✓ Constrained Multi-task Learning Model [Kobayashi et al., 2022]
✓ Discourse Context-Aware BERT [Hou 2020]		✓ End-to-end Bridging Resolution [Kobayashi et al., 2022]
✓ End-to-end Information Status Classification [Hou 2021]		✓ PairSpanBERT Model [Kobayashi et al., 2023]

https://github.com/IBM/bridging-resolution

Bridging Anaphora Resolution as Question Answering (ACL20)

s1: In the hard - hit ***Marina*** neighbourhood, life after the earthquake is often all too real, but sometimes surreal.

s2: Some scenes: -- Saturday morning, **a resident** was given 15 minutes to scurry into a sagging building and reclaim what she could of her life's possessions.

…

s23: In post-earthquake parlance, her building is a "red".

s24: After being inspected, ***buildings with substantial damage*** were color-coded.

s25: Green allowed **residents** to re-enter; yellow allowed **limited access**; red allowed **residents one last entry** to gather everything they could within 15 minutes.

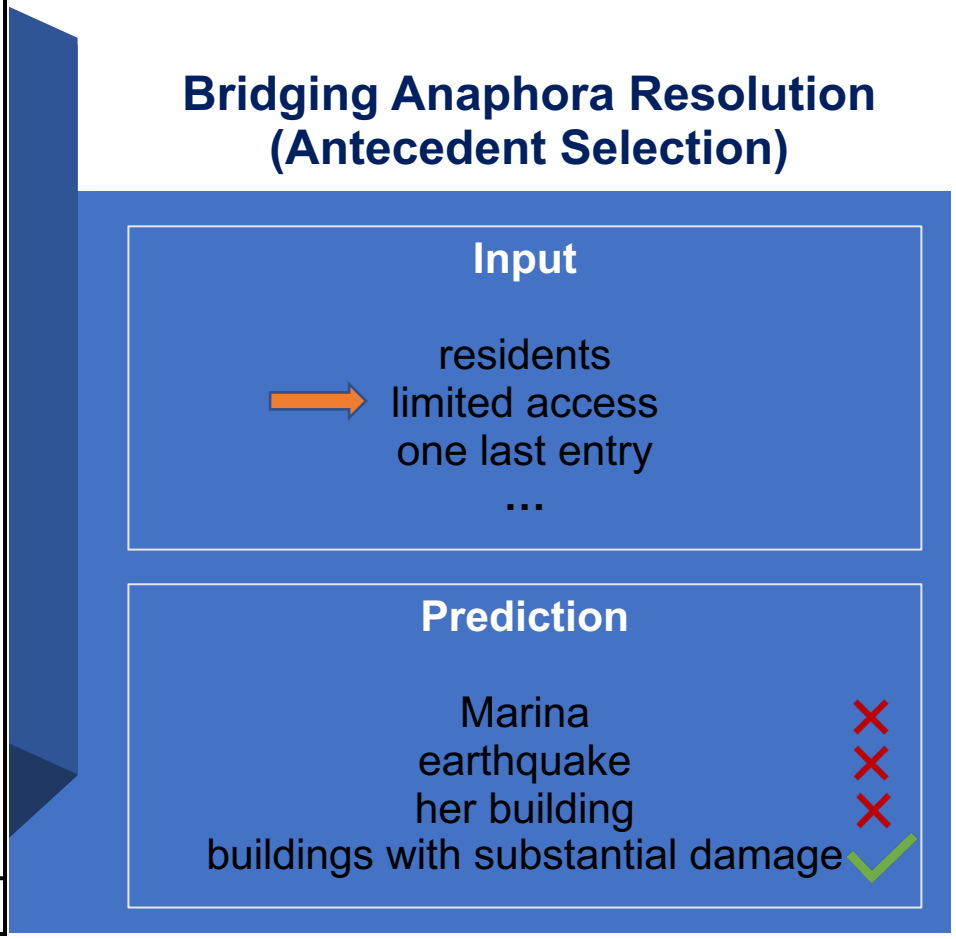
…

s34: ***One building*** was upgraded to red status while people were taking things out, and **a resident who wasn't allowed to go back inside** called up **the stairs** to his girl friend, telling her keep sending things down to **the lobby**.

…

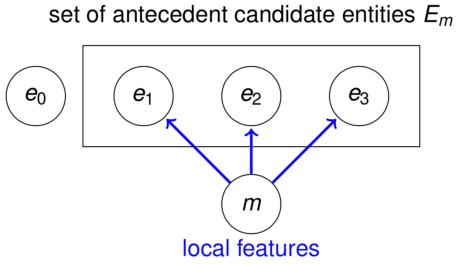
s36: Enforcement of restricted - entry rules was sporadic, **residents** said.

Bridging: - - - - - - ▸

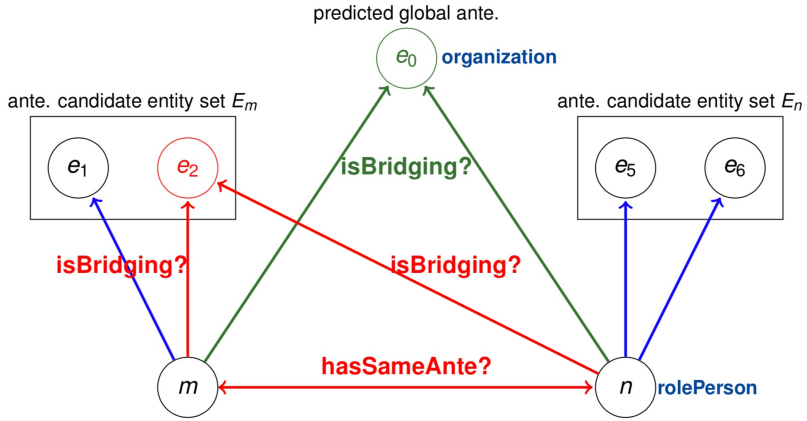


Bridging Anaphora Resolution: From NPs to the Surrounding Contexts

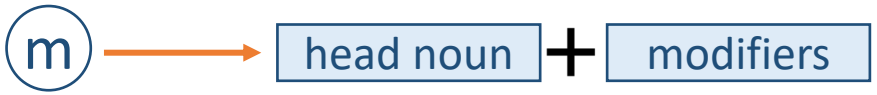
❑ Pairwise mention-entity model [Poesio et al. (2004) , Lassalle and Denis (2011)]



❑ Global model based on MLNs [Hou et al. (2013)]



❑ A deterministic algorithm based on *embeddings_bridging* [Hou (2018)]



Problems

- ✓ Previous work mainly considers the semantic of NPs and often fails to resolve context-dependent bridging anaphors.
- ✓ Previous studies assume that the gold mention or entity information is given.

BARQA: A QA System for Bridging Anaphora Resolution

- Built on top of the vanilla BERT QA framework
- Question generation
  - ✓ The preposition "**of**" in the syntactic structure "np1 **of** np2" encodes different associative relations between NPs that cover a variety of bridging relations
    - the price **of** the stock → an attribute of an object
    - the chairman **of** IBM → a professional function in an organization
  - ✓ Transform a bridging anaphor into a question

[Premodifiers][NP head]~~[postmodifiers]~~ → [Premodifiers][NP head] of what?

Anaphor: a painstakingly document report~~, based on hundreds of interview with randomly selected refugees~~

Question: a painstakingly document report of what?

BARQA: A QA System for Bridging Anaphora Resolution

- Answer generation
  - ✓ A new evaluation strategy for assessing the task where no gold mention/entity information is given
  - ✓ For an antecedent NP n, we consider its meaningful sub-parts that keep the main semantics of n as valid antecedents
    - the head noun of n → the total potential **claims** from the disaster
    - remove all postmodifiers from n → **the total potential claims** from the disaster
    - remove all postmodifiers and the determiner from n → the **total potential claims** from the disaster
- Inference
  - ✓ For each bridging anaphor a, we only predict text spans which appear before a from its context

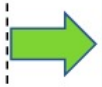
BARQA: A QA System for Bridging Anaphora Resolution

Input Text

...

In post-earthquake parlance, her building is a ``red''. After being inspected, buildings with substantial damage were color-coded. Green allowed **residents** to re-enter; yellow allowed **limited access**; red allowed **residents one last entry** to gather everything they could within 15 minutes.

...

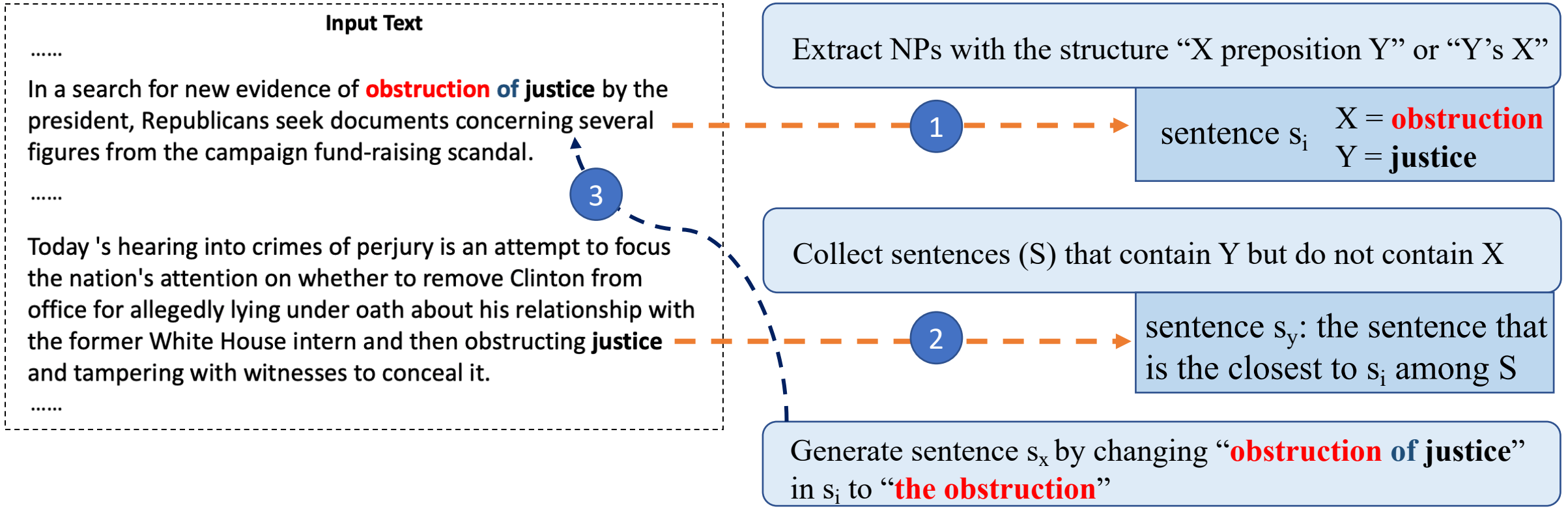


BARQA

Question	Context	Answers	Predicted Spans
residents of what?	In post-earthquake parlance, her building is a ``red''. After being inspected, buildings with substantial damage were color-coded. Green allowed residents to re-enter; yellow allowed limited access; red allowed residents one last entry to gather everything they could within 15 minutes.	(1) buildings with substantial damage (2) buildings	(1) buildings with substantial damage (2) buildings (3) her building (4) damage (5) Green (6) damage were color-coded ...
limited access of what?	In post-earthquake parlance, her building is a ``red''. After being inspected, buildings with substantial damage were color-coded. Green allowed residents to re-enter; yellow allowed limited access; red allowed residents one last entry to gather everything they could within 15 minutes.	(1) buildings with substantial damage (2) buildings	(1) buildings with substantial damage (2) buildings (3) her building (4) substantial damage (5) Green allowed residents ...
...	...	...	

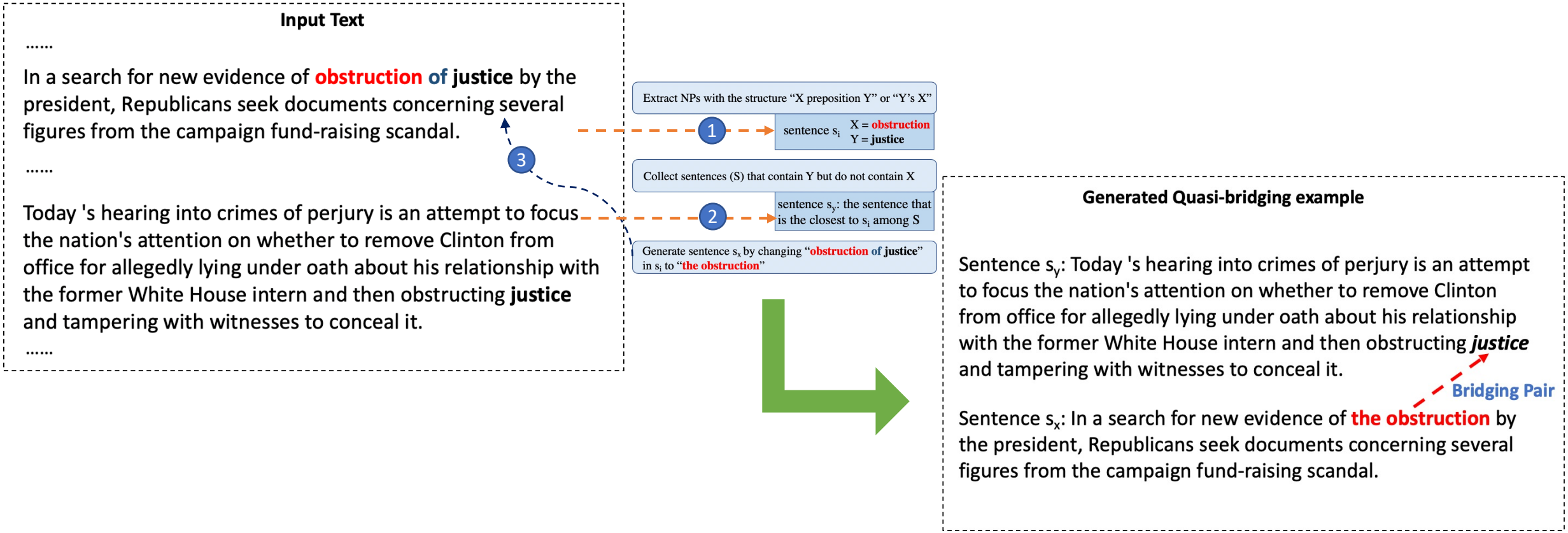
Generate "Quasi-bridging" Training Data

- There are no large-scale corpora available for referential bridging
- A novel method to generate "quasi-bridging" annotations



Generate “Quasi-bridging” Training Data

- A novel method to generate “quasi-bridging” annotations



Generate “Quasi-bridging” Training Data

- A novel method to generate “quasi-bridging” annotations
  - ✓ We apply the method to the NYT19 section of the Gigaword corpus
  - ✓ A large amount of “quasi-bridging” training data (~2.8 million bridging pairs)
  - ✓ Manually annotate 100 sentence pairs randomly sampled from all pairs

Quality Score	# Sent Pairs
2	25%
1	37%
0	38%

Experiments

Results on ISNotes Compared to Previous Approaches

System	Use Gold Mentions	Accuracy
Models from Hou et al. (2013b)		
pairwise model III	yes	36.35
MLN model II	yes	41.32
Models from Hou (2018a)		
embeddings_bridging (NP head + modifiers)	yes	39.52
MLN model II + embeddings_bridging (NP head + modifiers)	yes	46.46
This work		
BARQA with gold mentions/semantics, strict accuracy	yes	**50.08**
BARQA without mention information, strict accuracy	no	36.05
BARQA without mention information, lenient accuracy	no	47.21

Controlled experiment condition

A more realistic scenario in practice

Experiments

Datasets

Corpus	Genre	Bridging Type	# of Anaphors	# QA Pairs
ISNotes	WSJ news articles	referential bridging	663	1,115
BASHI	WSJ news articles	referential bridging	344	486
SQuAD 1.1 (train)	Wikipedia paragraphs	-	-	87,599
QuasiBridging	NYT news articles	quasi bridging	2,870,274	2,870,274

Results Using Different Training Strategies

large-scale out-of-domain training data

large-scale in-domain noisy training data

small in-domain training data

Best strategy

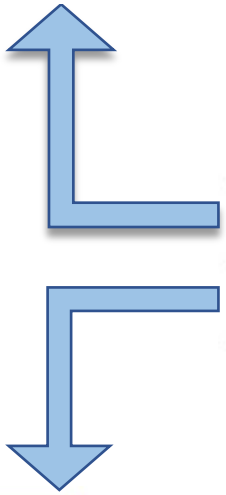
BARQA	Lenient Accuracy on ISNotes	Lenient Accuracy on BASHI
Large-scale (out-of-domain/noisy) training data		
SQuAD 1.1	28.81	29.94
QuasiBridging	25.94	17.44
Small in-domain training data		
BASHI	38.16	-
ISNotes	-	35.76
Pre-training + In-domain fine-tuning		
SQuAD 1.1 + BASHI	42.08	-
QuasiBridging + BASHI	**47.21**⋆	-
SQuAD 1.1 + ISNotes	-	35.76
QuasiBridging + ISNotes	-	**37.79**

Error Analysis

Bridging/World-knowledge

s1: While the discussions between Delmed and National Medical Care have been discontinued, Delmed will continue to supply ***dialysis products*** through National Medical after their exclusive agreement ends in March 1990, Delmed said.

s2: In addition, Delmed is exploring **distribution arrangements** with Fresenius USA, Delmed said.



	# pairs	*BARQA*	*MLN II + emb*
Know.	256	71.88	**88.28**
Context	407	**36.36**	19.90

Bridging/Context-dependent

In post-earthquake parlance, her building is a "red". After being inspected, ***buildings with substantial damage*** were color-coded. Green allowed **residents** to re-enter; yellow allowed **limited access**; red allowed **residents one last entry** to gather everything they could within 15 minutes.

Conclusions

Context-dependent bridging anaphors

In post-earthquake parlance, her building is a "red". After being inspected, ***buildings with substantial damage*** were color-coded. Green allowed **residents** to re-enter; yellow allowed **limited access**; red allowed **residents one last entry** to gather everything they could within 15 minutes.

BARQA [Hou 2020]

- We formalize bridging anaphora resolution as a context-dependent question answering problem
  - A QA system (BARQA)
- We explore a novel method to generate a large amount of quasi-bridging training dataset
  - pre-training with large scale noisy in-domain datasets + fine-tuning with small in-domain datasets
- We propose a new evaluation strategy to assess the task in a more realistic scenario in which no any gold mention/entity information is given

Models for Bridging Resolution

Bridging Anaphora Recognition	Bridging Anaphora Resolution	Full Bridging Resolution
✓ Collective Classification [Markert et al., 2012]	✓ Global Inference based on MLNs [Hou et al., 2013]	✓ Rule-based System [Hou et al., 2014]
✓ Cascade Collective Classification [Hou et al., 2013]	✓ Bridging Embeddings [Hou 2018a, 2018b]	✓ Learning-based Pipeline Model [Hou 2016]
✓ Incremental Classification Using Attention-based LSTMs [Hou 2016]	✓ Bridging Anaphora Resolution as Question Answering [Hou 2020]	✓ Constrained Multi-task Learning Model [Kobayashi et al., 2022]
✓ Discourse Context-Aware BERT [Hou 2020]		✓ End-to-end Bridging Resolution [Kobayashi et al., 2022]
✓ End-to-end Information Status Classification [Hou 2021]		✓ PairSpanBERT Model [Kobayashi et al., 2023]

https://github.com/IBM/bridging-resolution

PairSpanBERT: An Enhanced Language Model for Bridging Resolution (ACL23)



Hideo Kobayashi Yufang Hou Vincent Ng

- A pre-trained model specialized for bridging resolution
- Aims to *learn the contexts in which two NPs are implicitly linked to each other*
- Uses SpanBERT as a starting point for pre-training
  - Adds a pre-training step to SpanBERT with a novel objective

Step1: labelled data creation

Step2: Masking schemes

Step3: Pre-training tasks

Step1: Labelled Data Creation

Create data where the two NPs are *likely* to have a bridging relation.

1. Collect **noun** pairs that are likely involved in a bridging relation in a *context-independent* manner

   - Heuristically via **syntactic structures of noun phrases** (X preposition Y)
     - *NP: the winner **of** an election* → <the winner, an election>
     - 9.7 million noun pairs from the parsed Gigaword corpus

   - Distance supervision with **ConceptNet** using selected ConceptNet relations
     - PartOf, RelatedTo, HasA
     - 1.8 million noun pairs from ConceptNet

Step1: Labelled Data Creation

2. Use these pairs to automatically label 4 million documents from Gigaword

NP Pairs <X, Y>

<the winner, an election>
<candidate, an election>
<race, an election>
…

Meek and Crist essentially are now in **an election** within an election, with **the winner** to become the viable alternative to Republican Marco Rubio.

"I want to say a word about the third **candidate** in this **race**, Gov. Crist," said Gore, about halfway through his 12 minutes of remarks.

NP Pairs Source	# NP pairs	# Gigaword docs	# pseudo bridging links
Syntactic structure	9.7 M	4 M	1.7 B
ConceptNet	1.8 M	4 M	65 M

Step2: Masking Schemes

1. Span masking from SpanBERT: randomly select spans to be masked

2. Anchor masking: randomly select antecedents in the pseudo bridging links to be masked

Meek and Crist essentially are now in **an election** within an election, with **the winner** to become the viable alternative to Republican Marco Rubio.

Meek and Crist essentially are now in **[MASK] [MASK]** within an election, with **the winner** to become the viable alternative to Republican Marco Rubio.

Step3: Pre-training Tasks

SpanBERT

1. MLM objective: predict a masked token using the encoding of its hidden state

$$L_{MLM}(\text{election}) = -\log P(\text{election} \mid t_9)$$

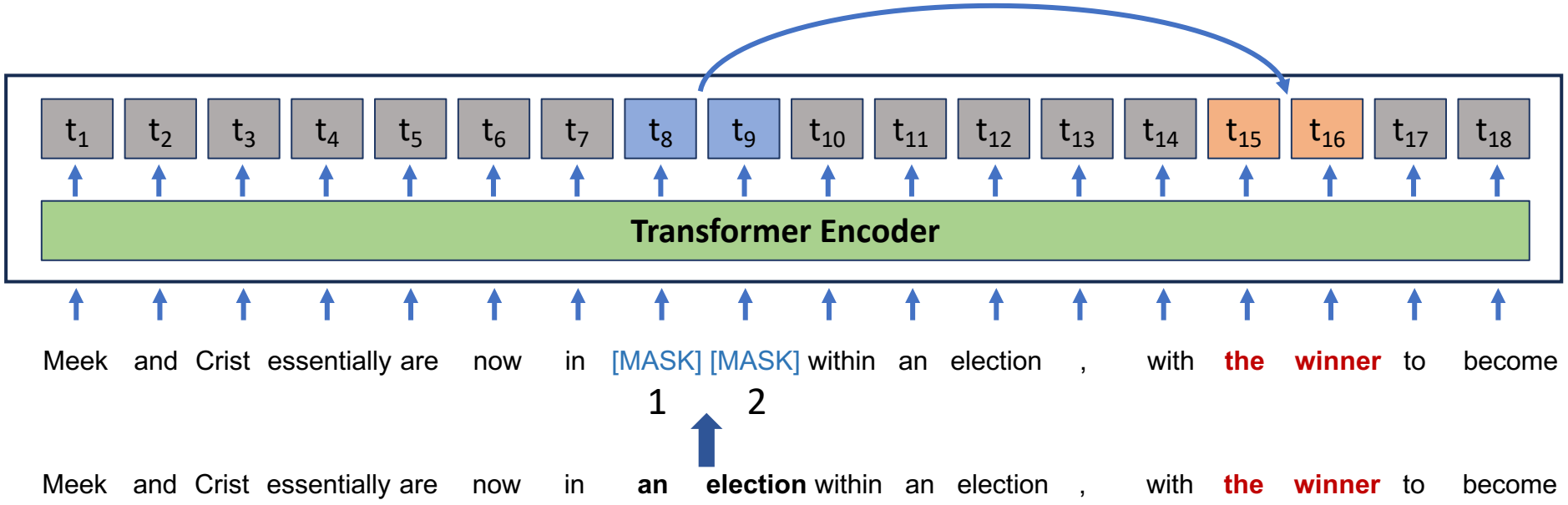
2. Span boundary objective: predict a masked token using encodings of external boundary tokens as well as the position embedding

$$L_{SBO}(\text{election}) = -\log P(\text{election} \mid t_7, t_{10}, p_2)$$

PairSpanBERT

3. Associative noun objective: apply to tokens masked by anchor masking, enable to model to learn the context in which two nouns are likely involved in a bridging relation

$$L_{ANO}(\text{an election}) = -\log P(\text{the winner} \mid [\text{MASK}][\text{MASK}])$$
$$= -\log P(t_{15} \mid t_8) \cdot P(t_{16} \mid t_9)$$



End-to-end Bridging Resolution System (Kobayashi et al., 2022)

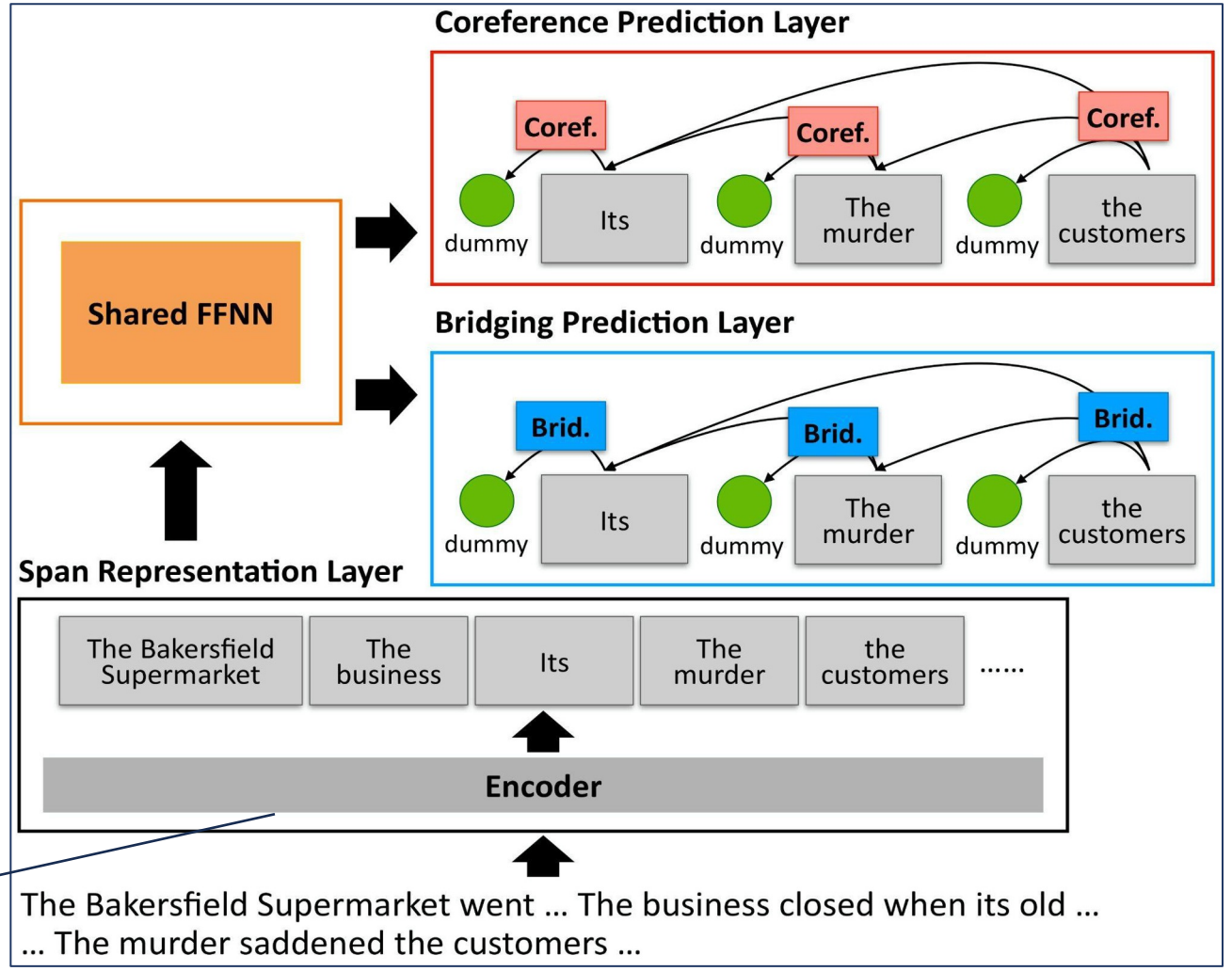
- Multi-task learning approach
- Take gold/predicted mentions as input
- Learn bridging/coreference score functions jointly
- A *Hybrid* approach to augment the MTL model with results from a rule-based resolver

$$s_{b'}(i,j) = \begin{cases} 0 & j = \epsilon \\ s_b(i,j) + \alpha r(i,j) & j \neq \epsilon \end{cases}$$

MTL bridging score function

Rule score function

SpanBERT -> PairSpanBERT



Experimental Results – Bridging Resolution

- Datasets

Anaphoric referential bridging

Anaphoric and Non-anaphoric referential bridging

Corpora	Docs	Tokens	Mentions	Anaphors
ISNotes	50	40,292	11,272	663
BASHI	50	57,709	18,561	459
ARRAU RST	413	228,901	72,013	3,777

- Our new resolver based on PairSpanBERT achieves the best results on three datasets for full bridging resolution

(a) ISNotes

Model	Recognition			Resolution		
	P	R	F	P	R	F
End-to-End Setting						
Rules(R)	49.4	17.4	25.7	31.8	11.2	16.5
Rules(H)	9.2	21.1	12.8	3.4	7.8	4.7
SBERT	34.4	30.9	32.6	22.3	20.1	21.1
SBERT(R)	39.7	31.6	35.1	27.0	21.5	23.9
SBERT(R,H)	34.6	37.1	35.8	22.8	24.4	23.6
PSBERT	36.3	36.8	36.6	22.3	22.6	22.5
PSBERT(R)	40.2	39.5	**39.9**	26.4	25.9	**26.2**

(b) BASHI

Model	Recognition			Resolution		
	P	R	F	P	R	F
End-to-End Setting						
Rules(R)	33.1	22.5	26.8	15.2	10.3	12.3
Rules(H)	3.5	15.1	5.7	1.0	4.3	1.6
SBERT	34.7	29.4	31.8	15.3	12.9	14.0
SBERT(R)	36.0	27.5	31.2	19.7	15.0	17.0
SBERT(R,H)	34.3	29.6	31.8	17.8	15.4	16.5
PSBERT	41.5	29.1	**34.2**	17.7	12.7	14.8
PSBERT(R)	43.0	25.6	32.1	25.4	14.3	**18.3**

(c) ARRAU AST

Model	Recognition			Resolution		
	P	R	F	P	R	F
End-to-End Setting						
Rules(R)	12.4	15.5	13.7	6.8	8.5	7.6
Rules(H)	6.6	14.5	9.0	1.6	3.6	2.2
SBERT	29.7	24.9	27.1	19.0	15.9	17.3
SBERT(R)	25.9	22.7	24.2	15.1	13.4	14.2
SBERT(R,H)	21.6	24.4	22.9	11.5	13.0	12.2
PSBERT	31.1	26.5	**28.6**	21.2	16.9	**18.8**
PSBERT(R)	28.1	23.2	25.4	16.7	14.1	15.3

Result Analysis

- Continue pre-train SpanBERT on the new training dataset with the original objectives?

New Training Data

NP Pairs Source	# NP pairs	# Gigaword docs	# pseudo bridging links
Syntactic structure	9.7 M	4 M	1.7 B
ConceptNet	1.8 M	4 M	65 M

New pre-training task?

SpanBERT

1. MLM objective: predict a masked token using the encoding of its hidden state

$$L_{MLM}(\text{election}) = -\log P(\text{election} \mid t_9)$$

2. Span boundary objective: predict a masked token using encodings of external boundary tokens as well as the position embedding

$$L_{SBO}(\text{election}) = -\log P(\text{election} \mid t_7, t_{10}, p_2)$$

PairSpanBERT

3. Associative noun objective: apply to tokens masked by anchor masking, enable to model to learn the context in which two nouns are likely involved in a bridging relation

$$L_{ANO}(\text{an election}) = -\log P(\text{the winner} \mid [\text{MASK}][\text{MASK}])$$
$$= -\log P(t_{15} \mid t_8) \cdot P(t_{16} \mid t_9)$$

- **PairSpanBERT's superior performance can be attributed to the addition of ANO rather than the additional pre-training steps**

Model	ISNotes	BASHI	ARRAU
SpanBERT (R)	23.9	17.0	14.8
ContinueSpanBERT (R)	23.6	16.7	14.9
PairSpanBERT (R)	**26.2**	**18.3**	**15.3**

End-to-end Bridging Resolution (F-score)

Result Analysis

- Error Analysis

Error Type	ISNotes
Bridging links	
Non-mentions -> Bridging anaphors	8.1%
Wrong bridging anaphors	73.1%
Correct bridging anaphor, wrong antecedent	18.8%
Bridging anaphor recognition	
New mentions -> Bridging anaphors	43%
Old mentions -> Bridging anaphors	25%

Additional Experimental Results on Relation Extraction

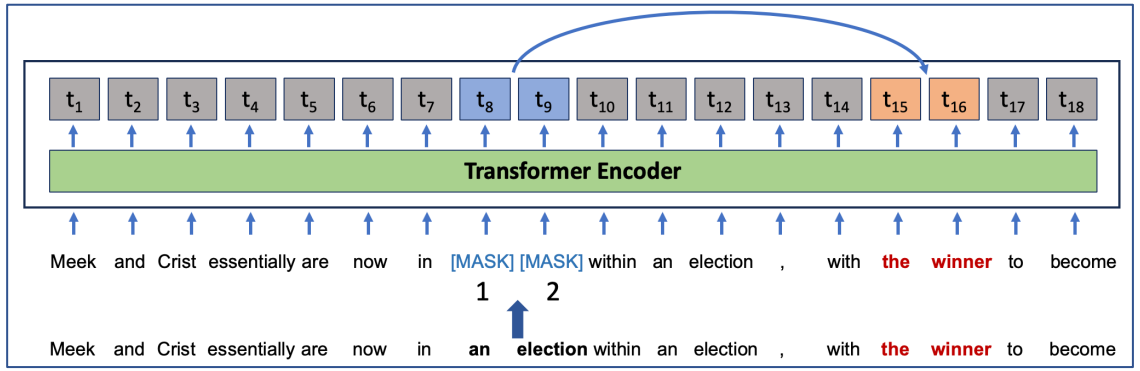
- TACRED dataset: 42 relation types

Data Split	# Ex.	Years
Train	75,050	2009–2012
Dev	25,764	2013
Test	18,660	2014

- Some initial results on relation extraction

Approach	Test F1
Zhou & Chen (2022) w SpanBERT	72.46
Zhou & Chen (2022) w PairSpanBERT	**73.53**

Conclusions



PairSpanBERT [Kobayashi et al., 2023]

- PairSpanBERT is a newly pre-trained model that can effectively capture bridging relations
- Our new resolver based on PairSpanBERT outperforms the previous SoTA models for bridging resolution
- Model can be download from: https://huggingface.co/utd/pairspanbert

Outline

- Information Status: A Model to Understand Discourse Entities
- Models for Bridging Resolution
- Probing LLMs for Bridging Inference
- Some Thoughts on Future Work

Probing for Bridging Inference in Transformer Language Models (NAACL21)

Onkar Pandit

Yufang Hou

- Bridging signals in each attention head of vanilla BERT
- **Fill-in-the-gap probing: Of-Cloze test**
  - Formulate bridging anaphora resolution as a masked token prediction task

S1: **[**The Bakersfield Supermarket**]** went bankrupt **[**last May**]**.

S2: **[**The business located in **[**northern Manhattan**]]** closed when **[[**its**]** owner**]** was murdered.

S3: **[Friends]**_bridging expressed outrage at the murder.

↓

S3': **Friends of [MASK]** expressed outrage at the murder.

↑

BERT/RoBERTa

Probing for Bridging Inference in Transformer Language Models (NAACL21)

Accuracy of selecting antecedents with different candidate scopes

Ante. Candidate Scope	# Anaphors	BERT-Base	BERT-Large	RoBERTa-Base	RoBERTa-Large
Salient/nearby mentions	531	31.64	33.71	34.08	**34.65**
All previous mentions	622	26.36	28.78	27.49	**29.90**

Probing for Bridging Inference in Transformer Language Models (NAACL21)

Accuracy of selecting antecedents with different candidate scopes

Ante. Candidate Scope	# Anaphors	BERT-Base	BERT-Large	RoBERTa-Base	RoBERTa-Large
Salient/nearby mentions	531	31.64	33.71	34.08	**34.65**
All previous mentions	622	26.36	28.78	27.49	**29.90**

Friends of [MASK]

Friends of [MASK] expressed outrage at the murder.

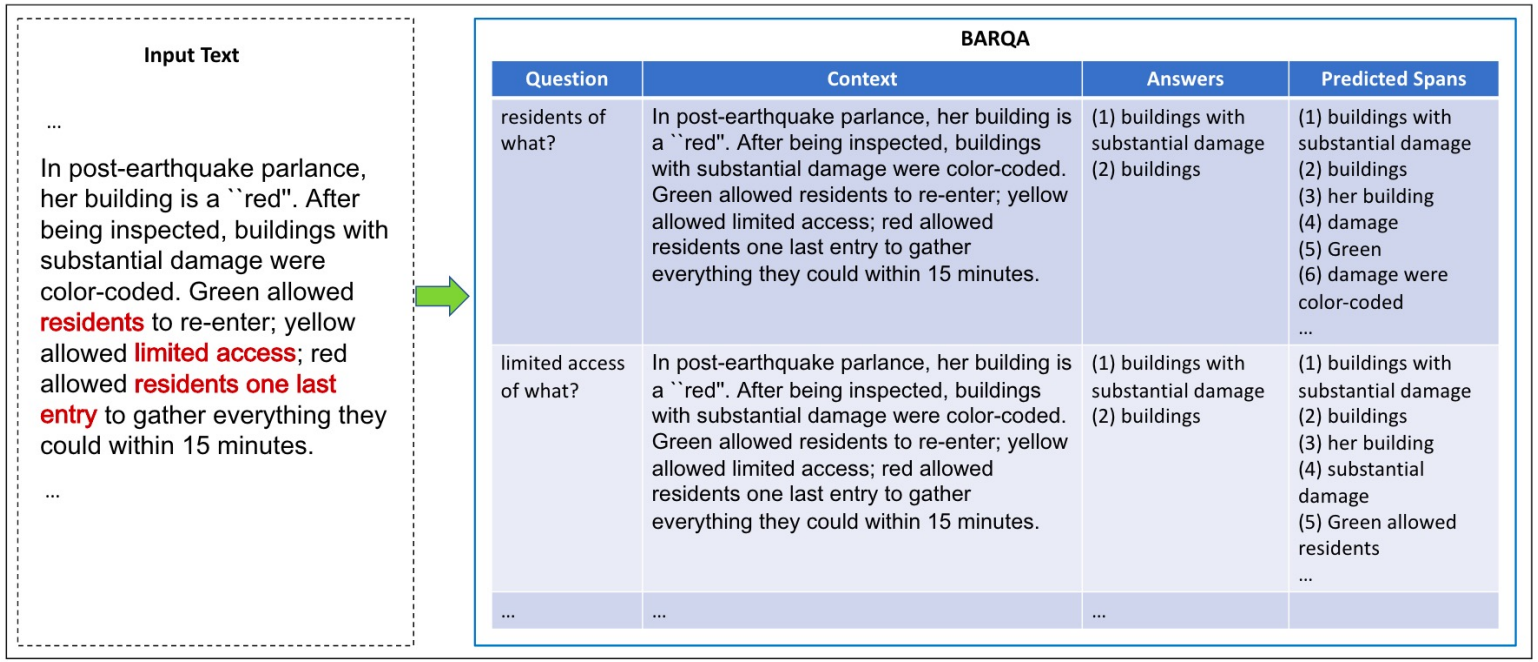
S1, S2 + Friends of [MASK] expressed outrage at the murder.

Context Scope	With "of"	Without "of"	Perturb
Only anaphor	17.20	5.62	-
Ana sent.	22.82	7.71	10.28
More context	26.36	12.21	11.41

Probing for Bridging Inference in BIG-bench

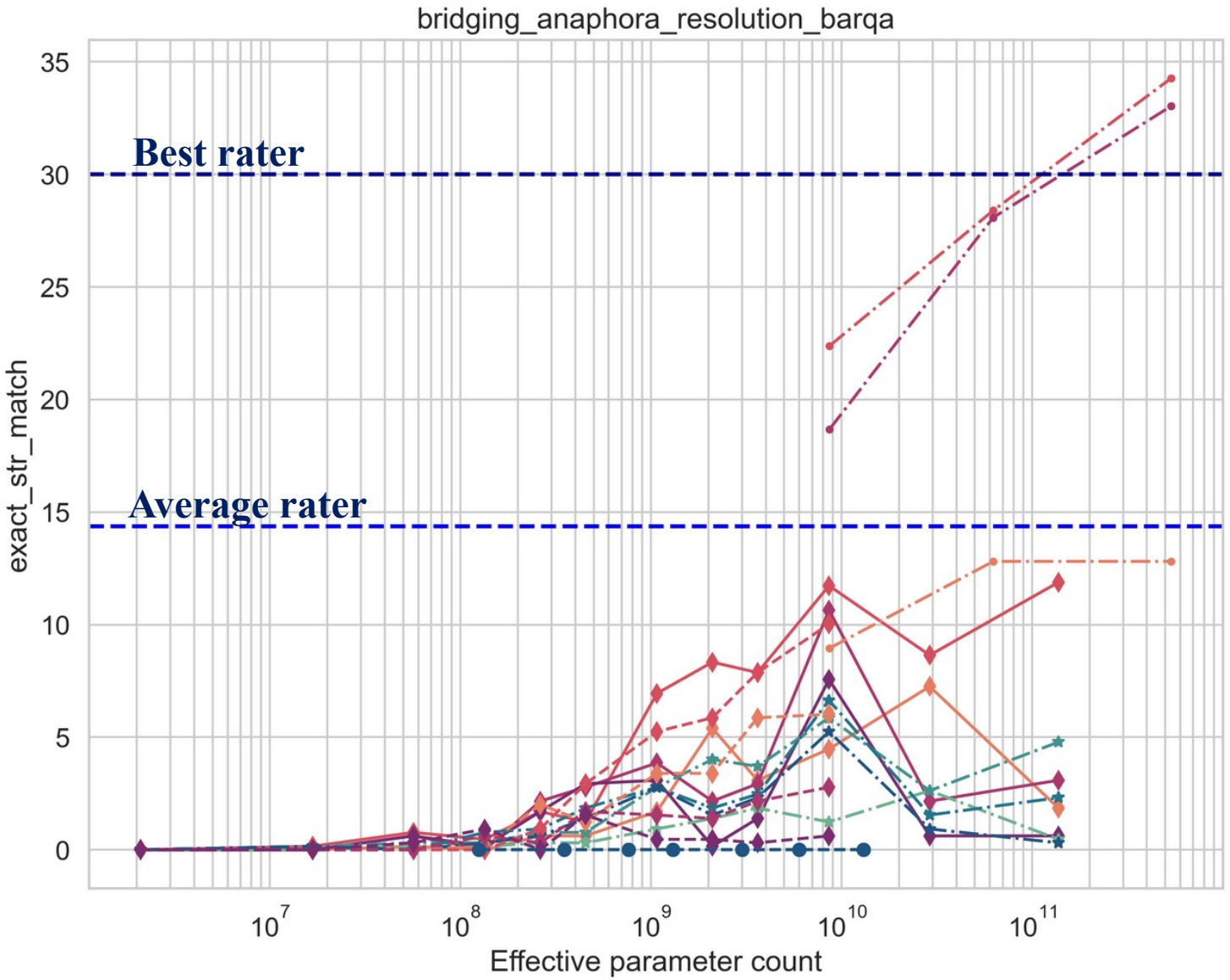
- Probing setup: antecedent selection as question answering
  - BARQA-ISNotes
    - ✓ 648 bridging anaphors and their lenient antecedent information from 50 WSJ news articles
    - ✓ Bridging anaphors in ISNotes are truly anaphoric and bridging relations are context dependent
    - ✓ **Context contains all previous sentences appearing before a bridging anaphor as well as the sentence that contains the anaphor**

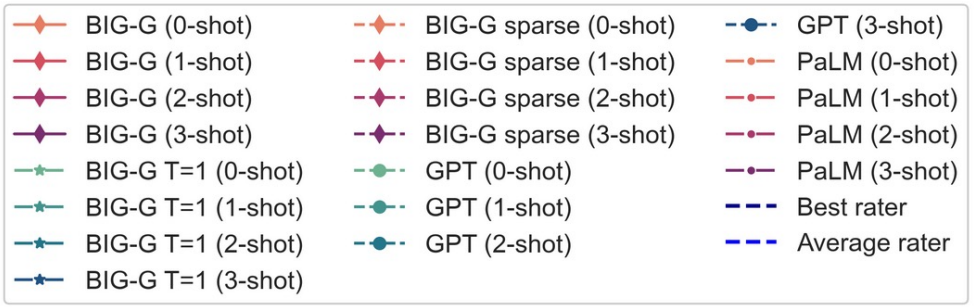
Context in BARQA only contain salient and nearby sentences



Question	Context	Answers	Predicted Spans
residents of what?	In post-earthquake parlance, her building is a ``red''. After being inspected, buildings with substantial damage were color-coded. Green allowed residents to re-enter; yellow allowed limited access; red allowed residents one last entry to gather everything they could within 15 minutes.	(1) buildings with substantial damage (2) buildings	(1) buildings with substantial damage (2) buildings (3) her building (4) damage (5) Green (6) damage were color-coded ...
limited access of what?	In post-earthquake parlance, her building is a ``red''. After being inspected, buildings with substantial damage were color-coded. Green allowed residents to re-enter; yellow allowed limited access; red allowed residents one last entry to gather everything they could within 15 minutes.	(1) buildings with substantial damage (2) buildings	(1) buildings with substantial damage (2) buildings (3) her building (4) substantial damage (5) Green allowed residents ...
...	...	...	

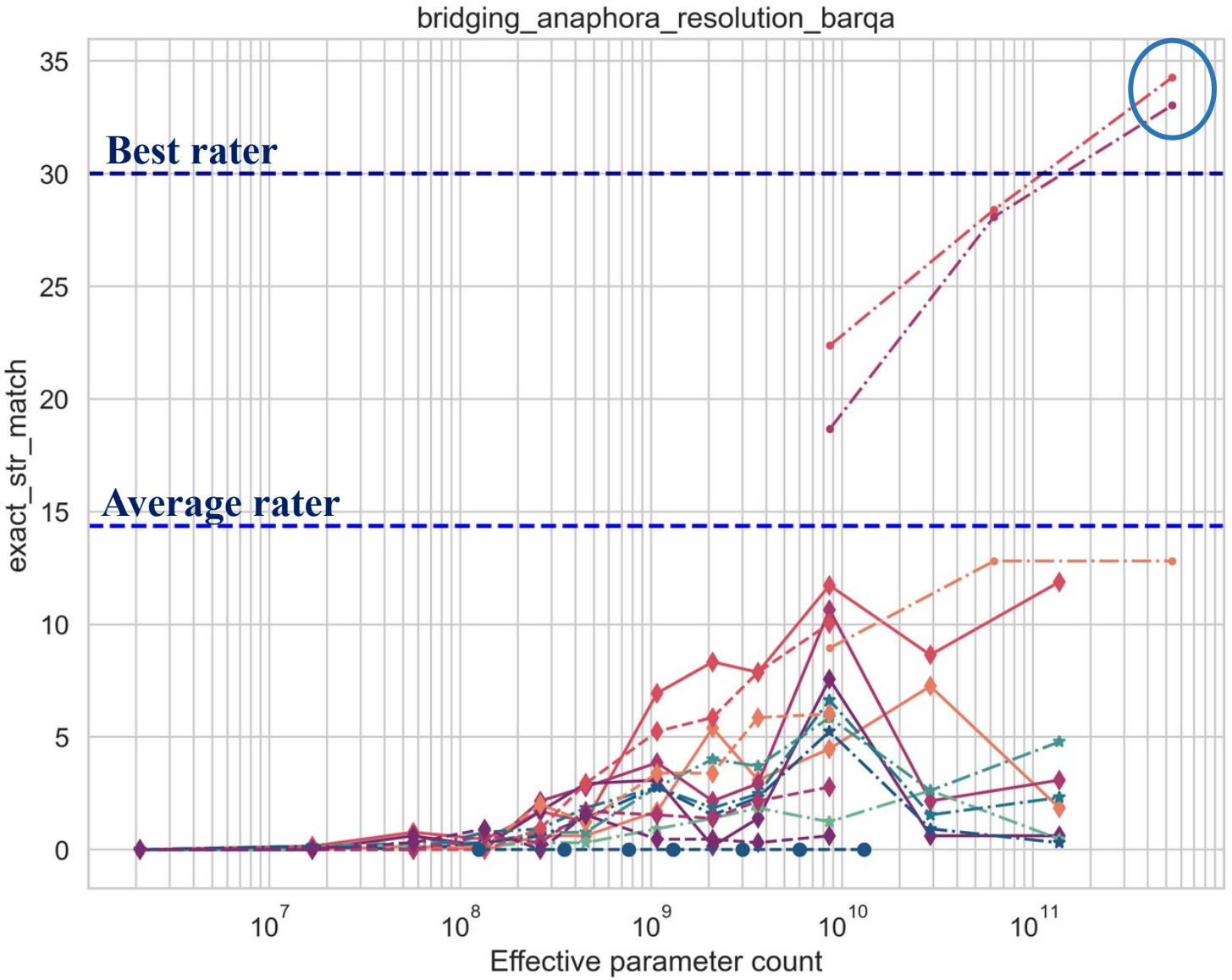
Probing for Bridging Inference in BIG-bench





- ✓ BIG-G: 12 dense decoder-only Transformer models from Google (2M – 137B)
- ✓ Big-G sparse: 10 Sparsely-activated expert models (51M – 48B)
- ✓ GPT: OpenAI GPT-3 model series (125M – 175B)
- ✓ PaLM: Pathways Language Models from Google (8B/62B/540B)
- ✓ Rater: crowd-source platform workers, strong lower bounds to a theoretical maximum of human performance

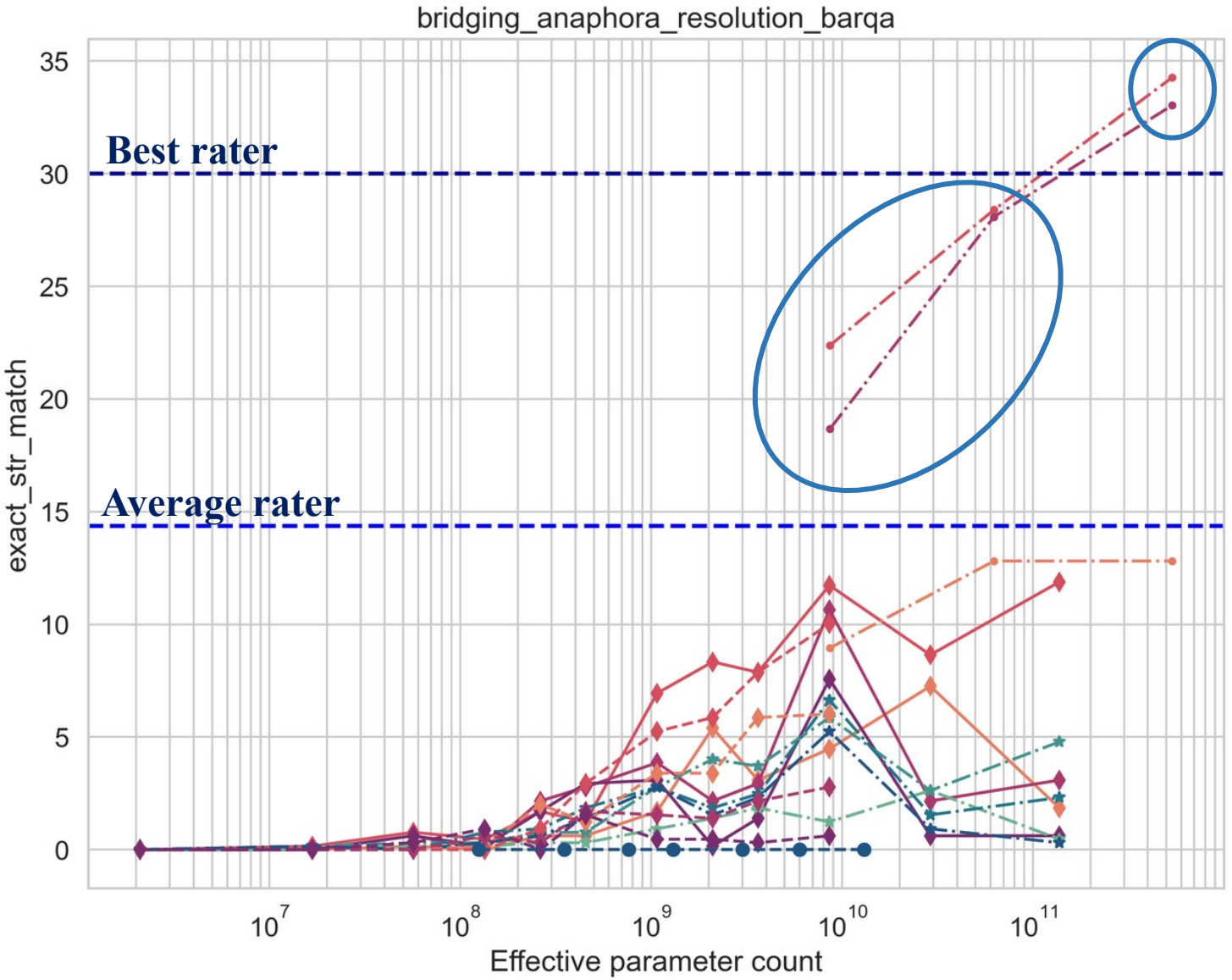
Probing for Bridging Inference in BIG-bench



Palm 540B 1-shot and 2-shot perform better than the best crowd-source worker

- Agreement for selecting bridging antecedents was around 80% for all expert annotator pairings

Probing for Bridging Inference in BIG-bench

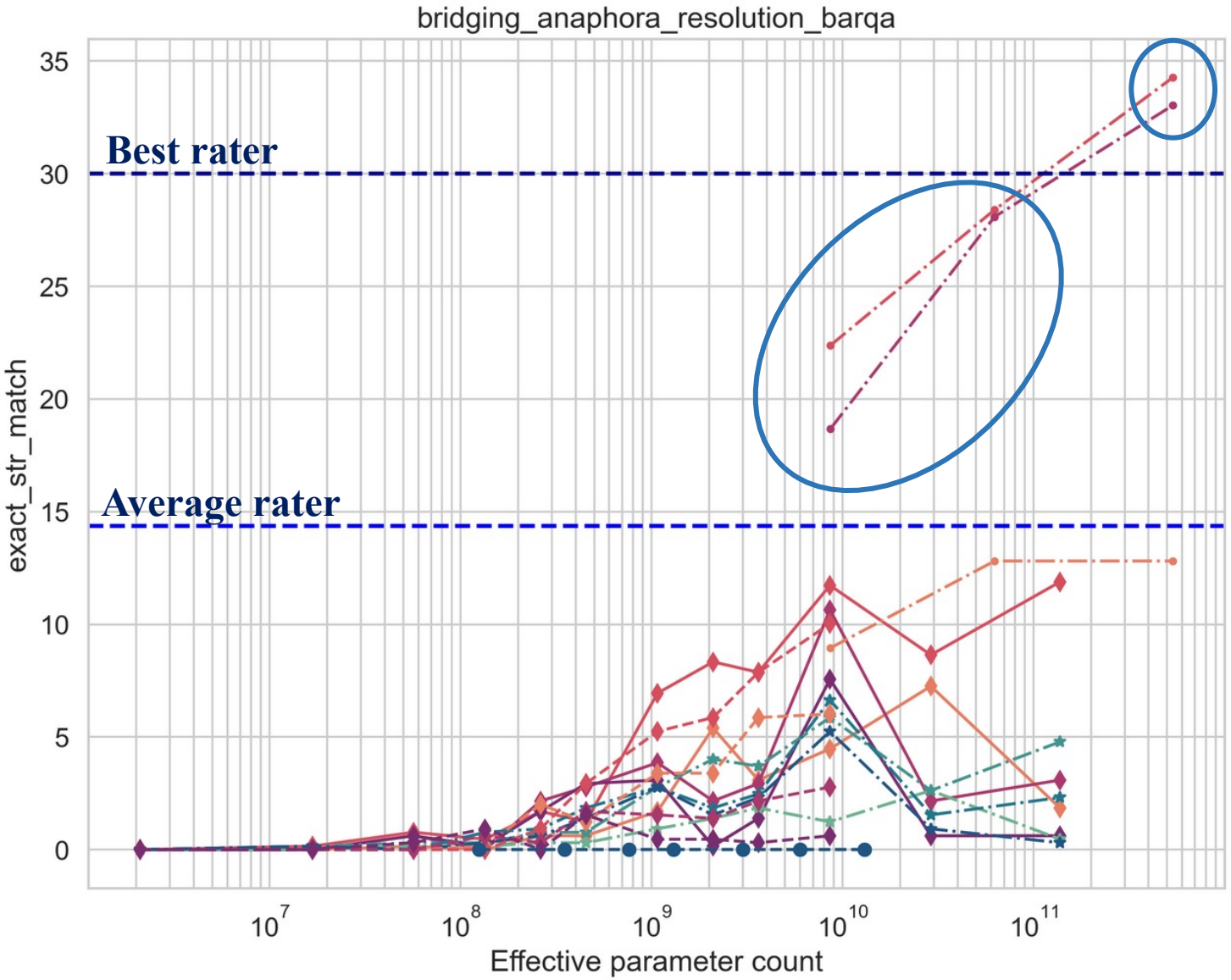


Palm 540B 1-shot and 2-shot perform better than the best crowd-source worker

- Agreement for selecting bridging antecedents was around 80% for all expert annotator pairings

Palm models (8B/62B) with 1-shot and 2-shot perform better than the other models of similar size and the average crowd-source workers

Probing for Bridging Inference in BIG-bench



Palm 540B 1-shot and 2-shot perform better than the best crowd-source worker

- Agreement for selecting bridging antecedents was around 80% for all expert annotator pairings

Palm models (8B/62B) with 1-shot and 2-shot perform better than the other models of similar size and the average crowd-source workers

No clear patterns w.r.t. the general trend of LLMs in BIG-bench

- Performance improves with model size
- BIG-G sparse models perform better than BIG-G models

Outline

- Information Status: A Model to Understand Discourse Entities
- Models for Bridging Resolution
- Probing LLMs for Bridging Inference
- Some Thoughts on Future Work

Bridging Definition

Anaphoric referential bridging (ISNotes/BASHI)

S1: The business located in northern Manhattan closed when **its owner** was murdered.

S2: **Friends** expressed outrage at the murder.

Non-anaphoric referential bridging (ARRAU)

S1: And, in some neighbourhoods, **rents** have merely hit a plateau.

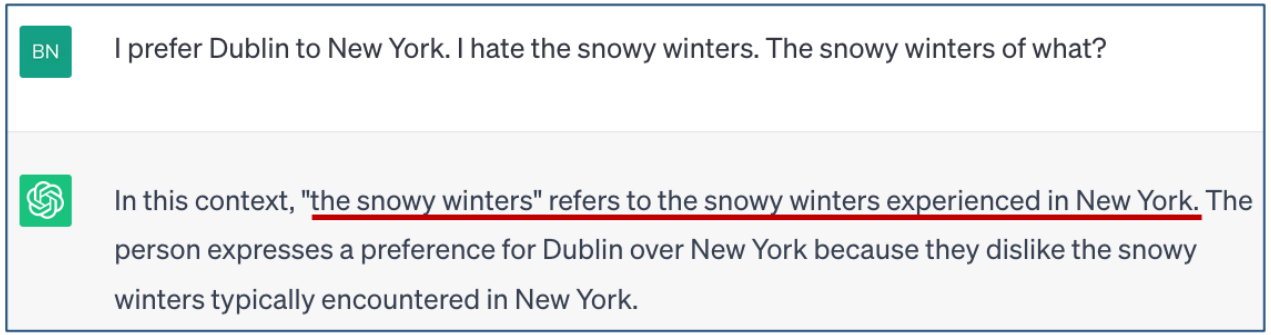
S2: But on average, **Manhattan retail rents** have dropped 10% to 15% in the past six months alone, experts say.

- Resolving different types of bridging requires different models.
- Different downstream applications might benefit from different bridging resolvers.

Bridging Relation

ISNotes bridging relation distribution

Relation Type	Bridging Pairs
Action	2.3%
Set/Membership	6.6%
Part-of/attribute-of	13.5%
Other	77.6%



- More fine-grained relation types
- Explainability: Explain the implicit inference that links a bridging anaphor to its antecedent by combining common sense knowledge and the discourse context

Michael Strube

Katja Markert

Vincent Ng

Hideo Kobayashi

Onkar Pandit

Thanks!

Bridging Examples