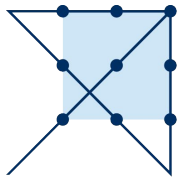


HITS
Heidelberg Institute for Theoretical Studies

TECHNISCHE UNIVERSITÄT DARMSTADT

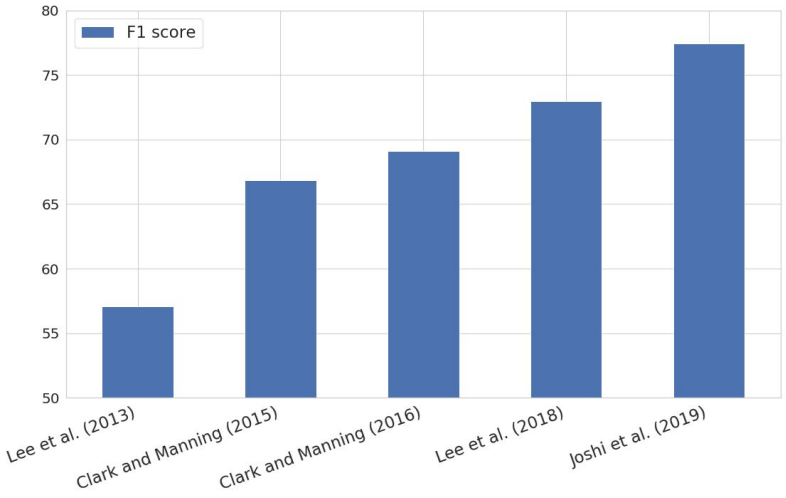
Evaluating Coreference Resolvers on Community-based Question Answering: From Rule-based to State of the Art

Haixia Chai[1] Nafise Sadat Moosavi[2,3] Iryna Gurevych[2] Michael Strube[1]

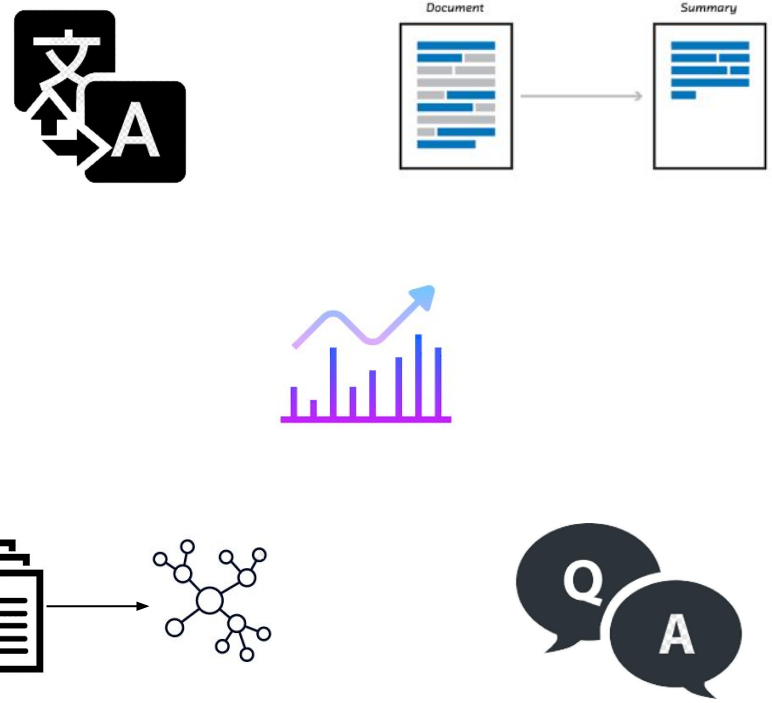
[1]Heidelberg Institute for Theoretical Studies (HITS)
[2]UKP Lab, Technische Universität Darmstadt
[3]The University of Sheffield

Oct 16th, 2022

Problem: Unclear Impact on Downstream Tasks



Coreference Resolution on CoNLL-2012



Downstream Tasks

Evaluation of Coreference Resolvers on CQA

CQA: Community-based Question Answering

Evaluation of Coreference Resolvers on CQA

CQA: Community-based Question Answering

Q:	Do I need a UK visa to enter UK from Ireland?
A1:	What is your nationality? According to the UK government service information website (URL), people from the countries who are mentioned in URL would still need to acquire a visa to enter the country.
A2:	Data sharing means only that they share data, so while the officers in Ireland are able to see details of your failed UK visa when they process your Irish visa, that doesn't mean you will be refused to get the visa to enter the country.

Issues:

- A1: the need for a visa from Ireland to UK
- A2: getting an Irish visa given that your UK visa has been rejected

Evaluation of Coreference Resolvers on CQA

CQA: Community-based Question Answering

Q:	Do I need a UK visa to enter UK from Ireland?
A1:	What is your nationality? According to the UK government service information website (URL), people from the countries who are mentioned in URL would still need to acquire a visa to enter the country.
A2:	Data sharing means only that they share data, so while the officers in Ireland are able to see details of your failed UK visa when they process your Irish visa, that doesn't mean you will be refused to get the visa to enter the country.

Issues:

- A1: the need for a visa from Ireland to UK
- A2: getting an Irish visa given that your UK visa has been rejected

Similar text sequence:

- A1: need to acquire a visa to enter the country
- A2: get the visa to enter the country

Evaluation of Coreference Resolvers on CQA

CQA: Community-based Question Answering

Q:	Do I need a UK visa to enter UK from Ireland?
A1:	What is your nationality? According to the **UK** government service information website (URL), people from the countries who are mentioned in URL would still need to acquire a visa to enter **the country**.
A2:	Data sharing means only that they share data, so while the officers in **Ireland** are able to see details of your failed UK visa when they process **your Irish visa**, that doesn't mean you will be refused to get **the visa** to enter **the country**.

Issues:

- A1: the need for a visa from Ireland to UK
- A2: getting an Irish visa given that your UK visa has been rejected

Similar text sequence:

- A1: need to acquire a visa to enter the country
- A2: get the visa to enter the country

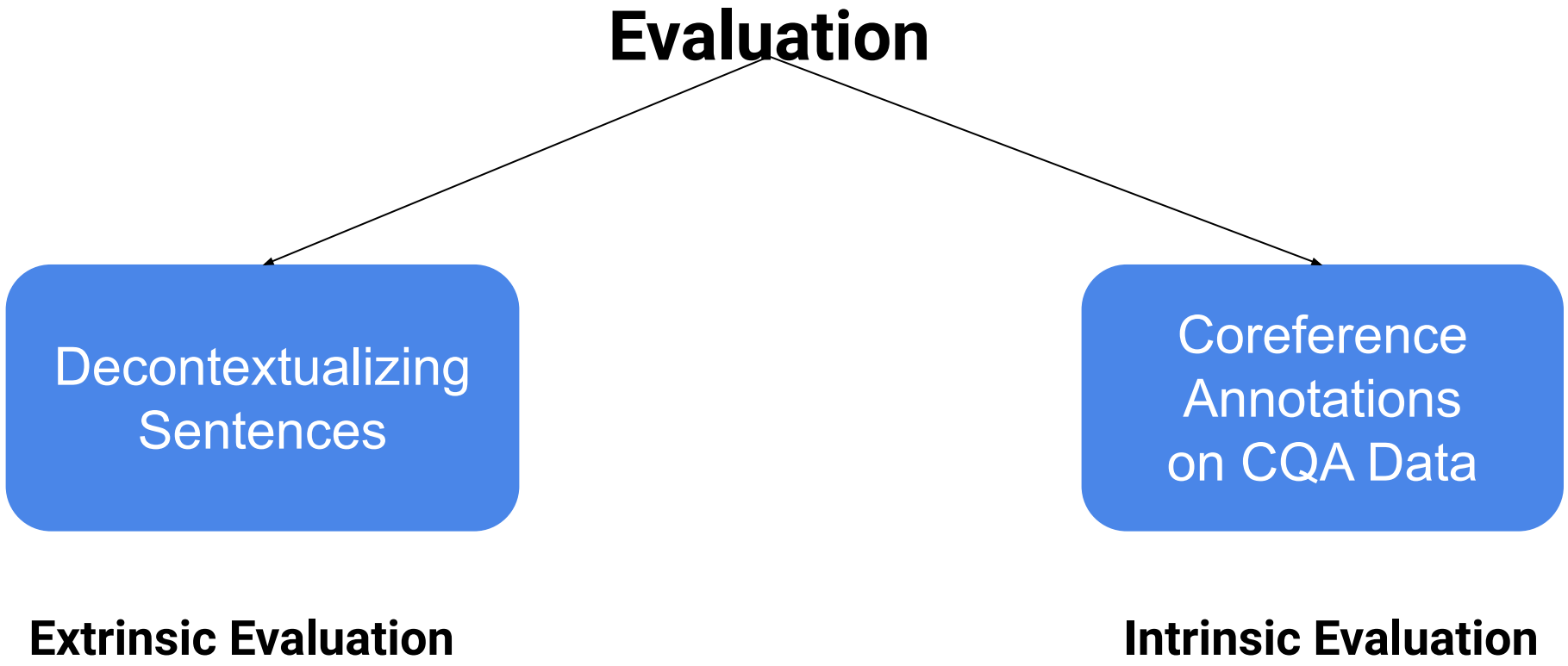
⇨

Given the coreference relations:

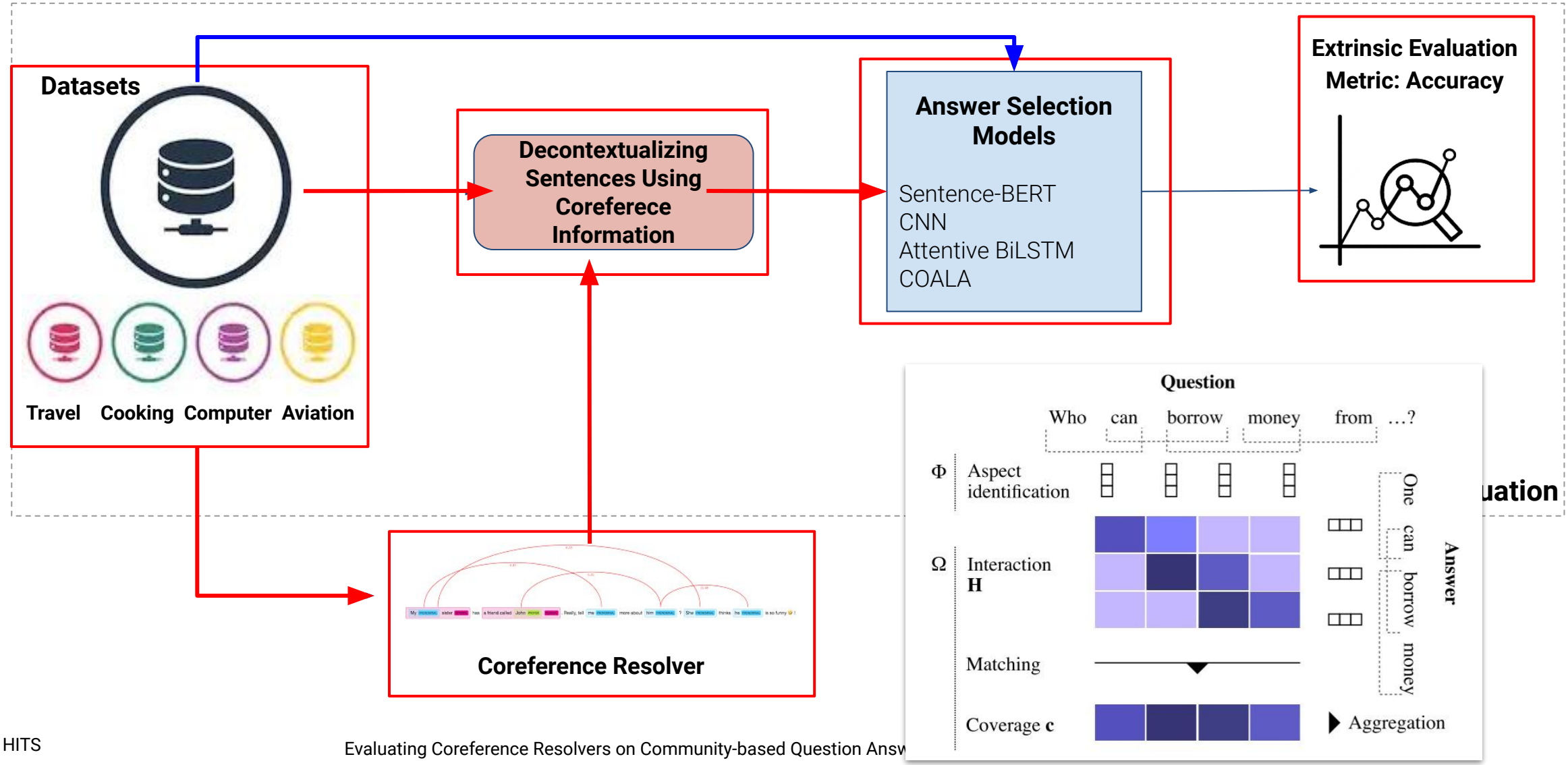
- A1: need to acquire a visa to enter UK
- A2: get your Irish visa to enter Ireland

Evaluation of Coreference Resolvers on CQA

CQA: Community-based Question Answering



Extrinsic Evaluation

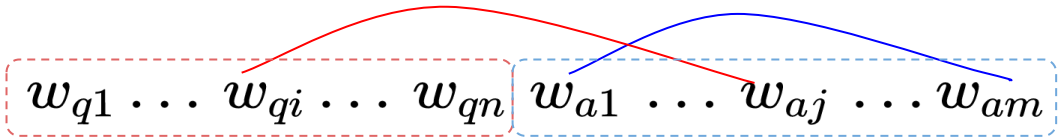


Extrinsic Evaluation

Where to apply coreference resolver?

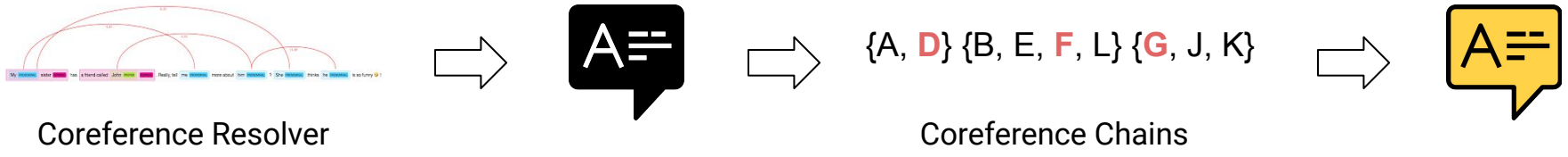
1. Q
2. A ✔
3. Q + A

Reasons:

1. Q	Questions are usually too short and do not contain coreference relations.
3. Q + A	 Example: Q: Do I need a ***UK visa*** to enter UK from Ireland? A: You can get ***it*** by going to the closest grocery store.

Extrinsic Evaluation

How to incorporate coreference annotation?



Rules (Lee et al., 2013):

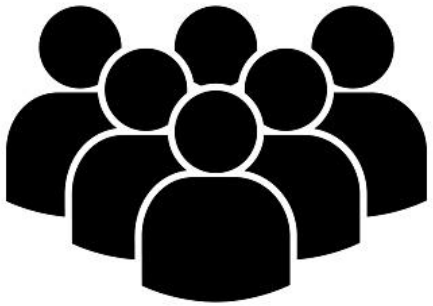
- **Mention Type**: proper names > common nouns > pronouns

  e.g., ‘the UK visa’ vs ‘it’

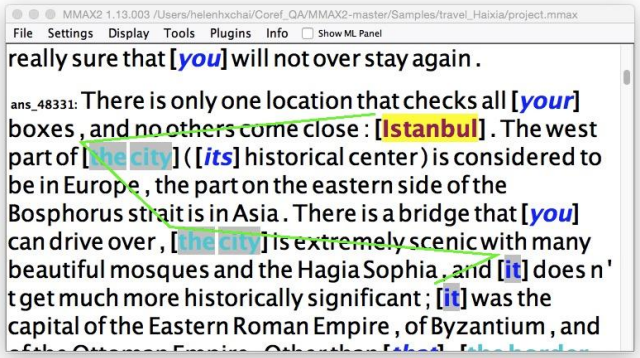
- **Number of words**: e.g., ‘the UK visa’ vs ‘the visa’

Intrinsic Evaluation on CQA Data

- Human Annotation for Coreference



Annotators



Tool: MMAX2

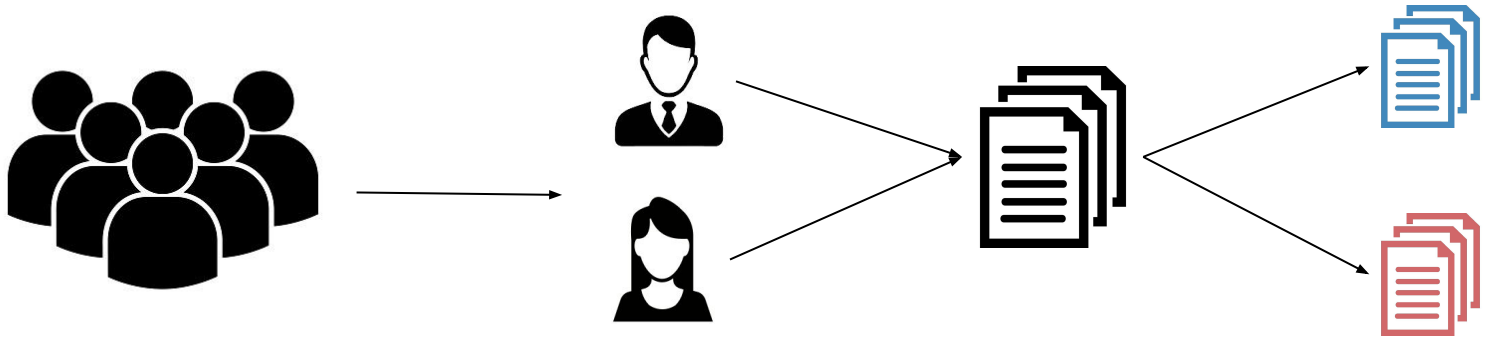
(Müller and Strube, 2006)

Annotation Guidelines

Annotations

Intrinsic Evaluation on CQA Data

- Inter-Annotator Agreement



$$Agreement = \frac{1}{N} \sum_{i=1}^{N} \alpha_i$$

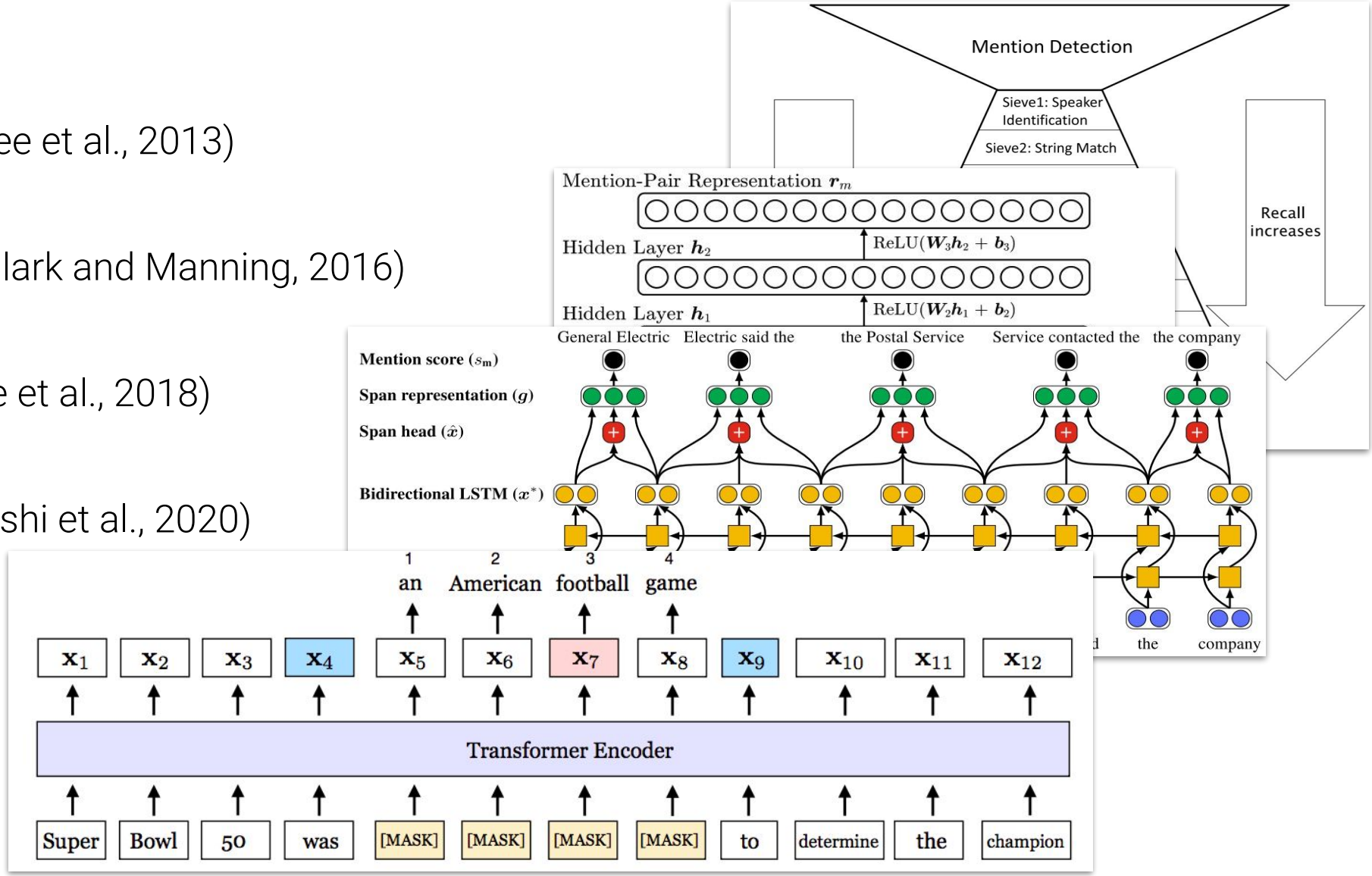
Results:

	MASI	Jaccard	Dice
Krippendorff's alpha	0.71	0.78	0.82

Interpretation of alpha value

Examined Coreference Resolvers

- rule-based (Lee et al., 2013)
- deep-coref (Clark and Manning, 2016)
- e2e-coref (Lee et al., 2018)
- bert-coref (Joshi et al., 2020)



Results: Extrinsic Evaluation (1/2)

- Evaluating CQA using coreference annotations in the test data

Coreference	Answer Selection	Travel	Cooking	Computer	Aviation
rule-based	Sentence-BERT	-1.57	-1.39	-0.96	-1.54
	CNN	**1.17**	**0.50**	**0.80**	0.00
	Att.-BiLSTM	***1.17***	***0.63***	***0.88***	***0.92***
	COALA	**0.78**	**0.13**	**1.44**	**0.46**
deep-coref	Sentence-BERT	-0.65	-0.63	-0.48	-1.54
	CNN	***0.52***	***0.75***	***0.40***	0.00
	Att.-BiLSTM	-0.13	**0.63**	**0.16**	**0.92**
	COALA	-0.40	**0.38**	**0.96**	**0.46**
e2e-coref	Sentence-BERT	**0.26**	-1.14	-0.48	-1.23
	CNN	***1.04***	***0.75***	***0.24***	***0.16***
	Att.-BiLSTM	-0.26	**0.50**	-0.24	-0.61
	COALA	**0.52**	-0.12	**0.24**	0.00
bert-coref	Sentence-BERT	-0.13	-1.01	0.00	-1.38
	CNN	**0.78**	-0.38	-0.40	**0.31**
	Att.-BiLSTM	**0.39**	**0.25**	**0.32**	**0.31**
	COALA	***0.39***	0.00	***0.56***	***0.61***

rule-based has a more positive impact and less negative impact on CQA compared to the state-of-the-art coreference resolver, bert-coref.

Results: Extrinsic Evaluation (1/2)

- Evaluating CQA using coreference annotations in the test data

Coreference	Answer Selection	Travel	Cooking	Computer	Aviation
rule-based	Sentence-BERT	-1.57	-1.39	-0.96	-1.54
	CNN	**1.17**	**0.50**	**0.80**	0.00
	Att.-BiLSTM	***1.17***	***0.63***	***0.88***	***0.92***
	COALA	**0.78**	**0.13**	**1.44**	**0.46**
deep-coref	Sentence-BERT	-0.65	-0.63	-0.48	-1.54
	CNN	***0.52***	***0.75***	***0.40***	0.00
	Att.-BiLSTM	-0.13	**0.63**	**0.16**	**0.92**
	COALA	-0.40	**0.38**	**0.96**	**0.46**
e2e-coref	Sentence-BERT	**0.26**	-1.14	-0.48	-1.23
	CNN	***1.04***	***0.75***	***0.24***	***0.16***
	Att.-BiLSTM	-0.26	**0.50**	-0.24	-0.61
	COALA	**0.52**	-0.12	**0.24**	0.00
bert-coref	Sentence-BERT	-0.13	-1.01	0.00	-1.38
	CNN	**0.78**	-0.38	-0.40	**0.31**
	Att.-BiLSTM	**0.39**	**0.25**	**0.32**	**0.31**
	COALA	***0.39***	0.00	***0.56***	***0.61***

Resolver	Mentions	Pronouns	Percentage
rule-based	99k	63k	64%
deep-coref	70K	51K	73%
e2e-coref	72K	56K	77%
bert-coref	81K	57K	70%

We hypothesize that resolving more nominal mentions and improving the precision of resolved pronouns will improve the effectiveness of coreference resolvers on downstream applications.

Results: Extrinsic Evaluation (1/2)

- Evaluating CQA using coreference annotations in the test data

Coreference	Answer Selection	Travel	Cooking	Computer	Aviation
rule-based	Sentence-BERT	-1.57	-1.39	-0.96	-1.54
	CNN	**1.17**	**0.50**	**0.80**	0.00
	Att.-BiLSTM	***1.17***	***0.63***	***0.88***	***0.92***
	COALA	**0.78**	**0.13**	**1.44**	**0.46**
deep-coref	Sentence-BERT	-0.65	-0.63	-0.48	-1.54
	CNN	***0.52***	***0.75***	***0.40***	0.00
	Att.-BiLSTM	-0.13	**0.63**	**0.16**	**0.92**
	COALA	-0.40	**0.38**	**0.96**	**0.46**
e2e-coref	Sentence-BERT	**0.26**	-1.14	-0.48	-1.23
	CNN	***1.04***	***0.75***	***0.24***	***0.16***
	Att.-BiLSTM	-0.26	**0.50**	-0.24	-0.61
	COALA	**0.52**	-0.12	**0.24**	0.00
bert-coref	Sentence-BERT	-0.13	-1.01	0.00	-1.38
	CNN	**0.78**	-0.38	-0.40	**0.31**
	Att.-BiLSTM	**0.39**	**0.25**	**0.32**	**0.31**
	COALA	***0.39***	0.00	***0.56***	***0.61***

Metric	rule-based	deep-coref	e2e-coref	bert-coref
MUC	64.7	74.2	80.4	**85.3**
B^3	52.7	63.0	70.8	**78.1**
$CEAF_e$	49.3	58.7	67.6	**75.3**
LEA	47.3	59.5	67.7	**75.9**

Intrinsic evaluation on CoNLL should be accompanied by extrinsic evaluation to approximate the utility of the coreference resolvers for the end tasks.

Results: Extrinsic Evaluation (1/2)

- Evaluating CQA using coreference annotations in the test data

Coreference	Answer Selection	Travel	Cooking	Computer	Aviation
rule-based	Sentence-BERT	-1.57	-1.39	-0.96	-1.54
	CNN	**1.17**	**0.50**	**0.80**	0.00
	Att.-BiLSTM	***1.17***	***0.63***	***0.88***	***0.92***
	COALA	**0.78**	**0.13**	**1.44**	**0.46**
deep-coref	Sentence-BERT	-0.65	-0.63	-0.48	-1.54
	CNN	***0.52***	***0.75***	***0.40***	0.00
	Att.-BiLSTM	-0.13	**0.63**	**0.16**	**0.92**
	COALA	-0.40	**0.38**	**0.96**	**0.46**
e2e-coref	Sentence-BERT	**0.26**	-1.14	-0.48	-1.23
	CNN	***1.04***	***0.75***	***0.24***	***0.16***
	Att.-BiLSTM	-0.26	**0.50**	-0.24	-0.61
	COALA	**0.52**	-0.12	**0.24**	0.00
bert-coref	Sentence-BERT	-0.13	-1.01	0.00	-1.38
	CNN	**0.78**	-0.38	-0.40	**0.31**
	Att.-BiLSTM	**0.39**	**0.25**	**0.32**	**0.31**
	COALA	***0.39***	0.00	***0.56***	***0.61***

The impact of coreference resolvers varies for different CQA models. We suggest to consider the overall impact on multiple CQA models to investigate the effect of a coreference resolver on CQA.

Results: Extrinsic Evaluation (2/2)

- Evaluating CQA using coreference annotations in both training and test data

Resolver	CQA	Travel	Cooking
rule-based	CNN	-0.78	1.26
	Att.-BiLSTM	2.35	0.13
	COALA	0.91	0.63
bert-coref	CNN	2.22	0.63
	Att.-BiLSTM	2.09	-2.27
	COALA	0.13	-2.14

Analysis:

1. Incorporating coreference relations in both training and test datasets results in **higher improvements** compared to only incorporating them in the test data.

2. *bert-coref* performs better on the Travel domain, while rule-based shows most positive results on both domains.

Results: Intrinsic Evaluation

Metric	rule-based	deep-coref	e2e-coref	bert-coref
		Travel		
MUC	28.07	**55.36**	34.90	39.53
B^3	28.81	**50.66**	34.28	39.31
$CEAF_e$	33.56	**45.83**	38.95	44.62
LEA	23.19	**46.86**	30.19	35.29
ARCS	18.24	23.99	29.47	**36.80**
		Cooking		
MUC	31.58	**59.43**	37.82	43.07
B^3	30.99	**54.85**	36.17	40.70
$CEAF_e$	34.77	**52.42**	41.36	45.11
LEA	24.47	**50.01**	30.88	36.04
ARCS	15.49	24.37	26.27	**34.17**

ARCS (Tuggener, 2014): Evaluating coreference resolvers based on their potential impact on downstream applications.

Analysis:

1. All standard coreference evaluation metrics agree on the ranking of the examined resolvers on both domains.
2. ARCS ranks *bert-coref* higher than the rest of the systems on both domains.

➢ None of the rankings is consistent with our extrinsic evaluations.

➢ Existing evaluation metrics are linguistic-agnostic.

Summary and Discussion

- We perform a thorough investigation on the impact of coreference resolvers on a downstream task, community-based question answering.
  - ➔ Extrinsic evaluation: using coreference relations to decontextualize individual sentences.
  - ➔ Intrinsic evaluation: coreference annotated data
- Our method is efficient for the extrinsic evaluation covering the most coreference resolvers, downstream models and datasets.
- On the downside, the decontextualization results in unnatural sentences in some examples.
- Various evaluation methods could result in very different extrinsic evaluation results on different downstream models and datasets.

HITS

www.h-its.org

Thanks for your attention!

Haixia Chai
haixia.chai@h-its.org
PhD Student at NLP Group - HITS