

# Improving Bridging Reference Resolution using Continuous Essentiality from Crowdsourcing

Nobuhiro Ueda and Sadao Kurohashi, Kyoto University, Japan

October 16th, CRAC2022



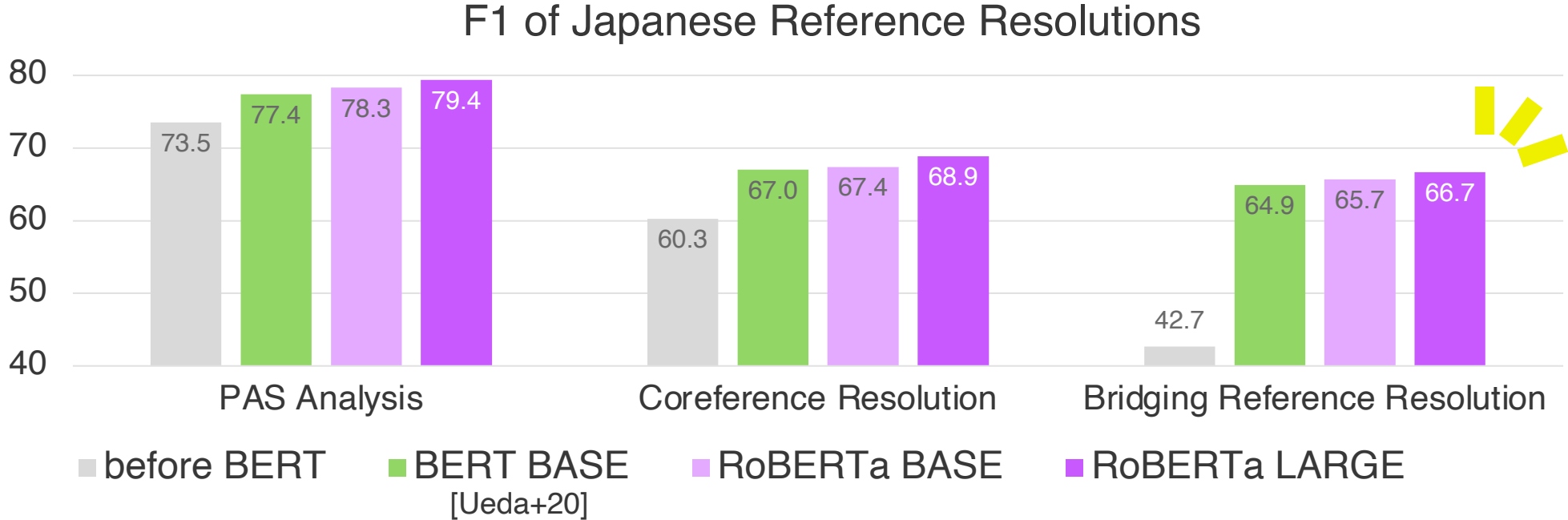
# Abstract

---

- Bridging reference resolution is a reference resolution task of finding non-identical antecedents
- **Challenge:** Continuous strength of bridging relations, which is not well-represented in existing datasets
- **Method:** We propose a crowdsourcing-based annotation method to obtain continuous labels
- **Result:** Adding our constructed dataset improved the resolution performance

# Background: Current status of Japanese Reference Resolution

- Train and evaluate models based on labels annotated by experts
- The performance has greatly improved using large pre-trained models such as BERT<sub>[devlin+19]</sub>




- However, the performance of bridging reference resolution is still low

# Introduction: Bridging Reference

- Reference between non-identical nouns
- Especially the case where an anaphor is semantically insufficient by itself, and its antecedent complements its meaning
- **Essentiality**: the importance of the complemented meaning for the anaphor

antecedent                      anaphor

I can see a house over there. The roof is covered with snow



bridging reference

# Existing Japanese Corpora for Bridging Resolution

- The size of

We focus on this dataset due to its diversity

	Documents	# of bridging anaphors
KWDLC [Hangyo+12] (Web domain)	5,124	16,038
Kyoto Corpus [K&N03] (News domain)	1,909	15,872

- Bridging-related labels defined in KWDLC

label	example
<i>essential</i>	the capital of the US
<i>ambiguous</i>	glasses of mine
<i>optional</i>	A 50-cent candy

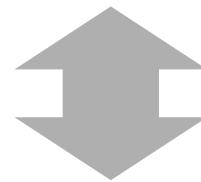


# A Challenge of Bridging Reference Resolution

There is a gap between the phenomenon of bridging reference and the annotations

The essentiality has a continuous distribution

He won the world swimming championships with a world record in 100m breaststroke.






































The existing corpora have only a few discrete labels

He won the world swimming championships with a world record in 100m breaststroke.



# Proposed Method: Utilizing Crowdsourcing

- We utilize crowdsourcing to obtain multiple labels for each example, and we can obtain more fine-grained annotations

	He	world...	world	record	100m...	...
						
						
						
						
						
						
						
score	6	3	11		8	5



Select all nouns such that “**record** of *sth*” is semantically valid



Select the most essential noun from the selected ones

× 38,840 questions

# Crowdsourcing Interface

- Original (Japanese)

問題7

【該当なし】

【書き手】

【読み手】

【その他（人）】

【その他（物）】

selected when none of noun As are related to noun B

生き物ではないので、関連グッズは非常に

- English translation

Question7

【NULL】

【Writer】

【Reader】

【Other (Person)】

【Other (Object)】

In this **corner**, we will introduce **goods** related to the **Japanese giant salamander**.  
related goods. If I find something interesting, I will introduce it in this corner.

As it is not a familiar creature, there are very few

noun A

noun A  
(exophora)

noun B

the most essential  
noun A



# Dataset Construction and Results

- We re-annotated a portion of KWDLC (**Expert** hereafter) and constructed a dataset called **Crowd**
- Krippendorff's alpha: 0.28
- We define **essentiality score** for noun  $A$

$$\begin{cases} n(A) \times 2 & \text{if noun } A \text{ is [NULL],} \\ n(A) + N(A) & \text{otherwise,} \end{cases}$$

$n(A)$  = (# of workers who selected noun  $A$ )

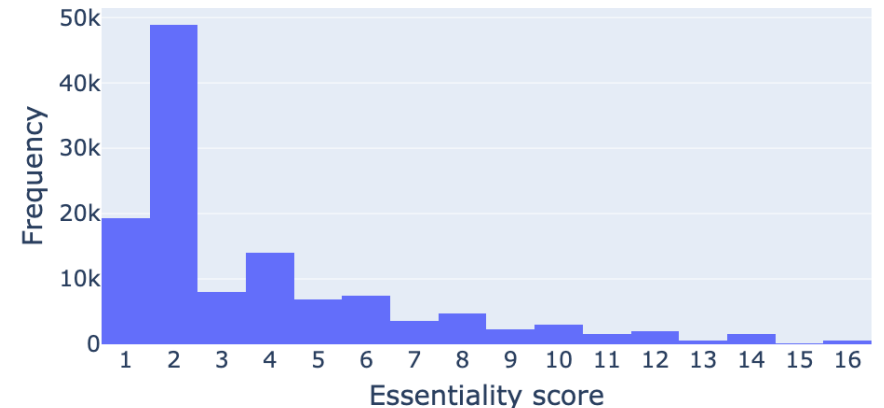
$N(A)$  = (# of workers who selected noun  $A$  as the most essential)

Corpus statistics

	# of docs	# of bridging anaphors
Expert	5,124	13,496
Crowd	3,933	*25,217

\*This is calculated for anaphors that at least half of the workers considered to be bridging

The distribution of essentiality score






# Constructed Dataset: Comparison to Expert

- We assume noun As with *essential* or *ambiguous* relation in Expert as ground truth and evaluate Crowd

	Precision	Recall	F1	Accuracy
Endophora	29.9	71.6	42.2	58.4
Exophora	6.7	48.9	11.8	52.9

- The low precision of exophora is reasonable as in most cases, an entity is owned by someone or is part of something

	color	essentiality
		1
		⋮
		16

【NULL】 (2)    【Writer】    【Reader】    【Other (Person)】    【Other (Object)】 (14)

The usage time includes the time for preparation and cleanup . Please return chairs and desks to their original state when you have used them.

# Other Collected Examples

- Many workers select correct nouns

color	essentiality
	1
⋮	⋮
	16

【NULL】 (2)    【Writer】    【Reader】    【Other (Person)】    【Other (Object)】 (2)

Athyrium niponicum (5) sometimes grows in flower beds (2) in urban areas . There are many species of this family (6) , and many hybrids , making them difficult to distinguish .

- The essentiality is represented as the number of votes

【NULL】    【Writer】    【Reader】 (2)    【Other (Person)】 (2)    【Other (Object)】

We want to own real estate (2) in the future and live off the rental (10) income . We would like to semi-retire from our current company.

Many people would like to do so.

# Evaluation with Bridging Reference Resolution

We compare the existing dataset (Expert) and constructed dataset (Crowd) in terms of the score on bridging reference resolution

- Training set

- Crowd (2,712 docs)
- Expert (3,912 docs)
- Crowd + Expert (6,633 docs)

- Evaluation set

- Crowd (700 docs)
- Expert (700 docs)

- Resolution model

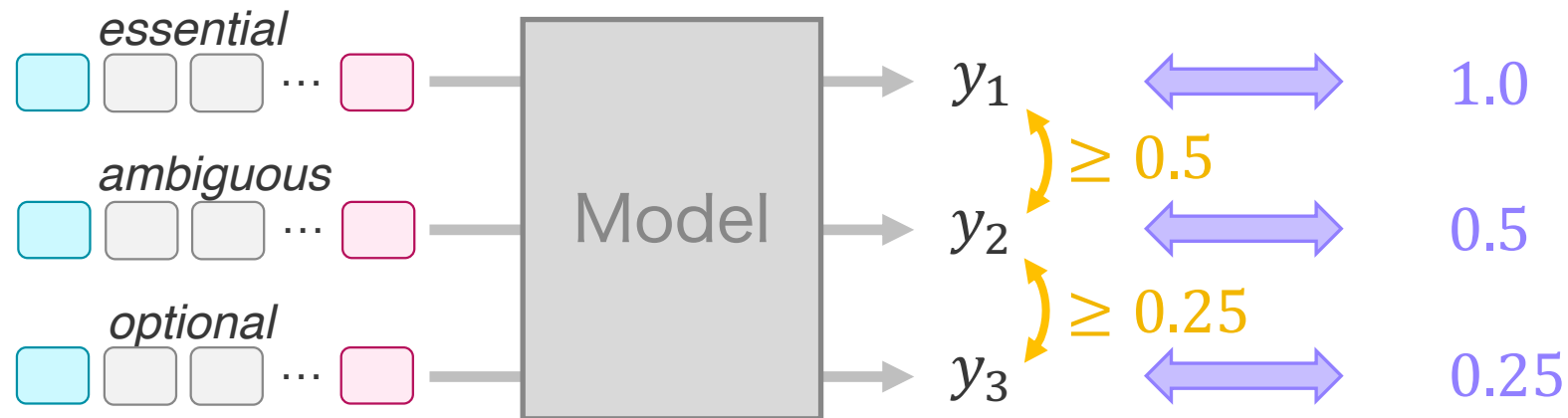
- learns to predict (normalized) essentiality score for each noun pair



# Training Objective

- For comparison, we convert the relations in Crowd and Expert into a value between 0 and 1
  - Crowd: normalize essentiality score
  - Expert: define a mapping table for conversion
- We use mean squared error (MSE) loss or margin ranking (MR) loss

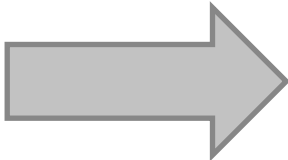
label	value
<i>essential</i>	1.0
<i>ambiguous</i>	0.5
<i>optional</i>	0.25



# Evaluation Metrics (when evaluating on Crowd)

## Gold label (Crowd)

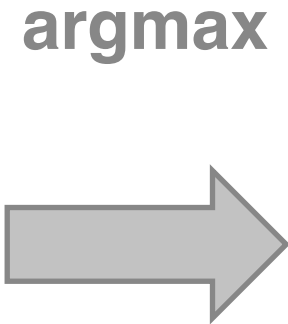
	He	world	100m	record	NULL
Noun B	He	-	-	-	16
	world	0	-	-	16
	100m	4	0	-	12
	record	5	7	8	2



He	NULL
world	NULL
100m	NULL
record	100m

## System prediction

	He	world	100m	record	NULL
Noun B	He	-	-	-	16.2
	world	0.3	-	-	15.8
	100m	9.1	3.5	-	6.2
	record	5.5	7.2	8.4	2.3



He	NULL
world	NULL
100m	He
record	100m

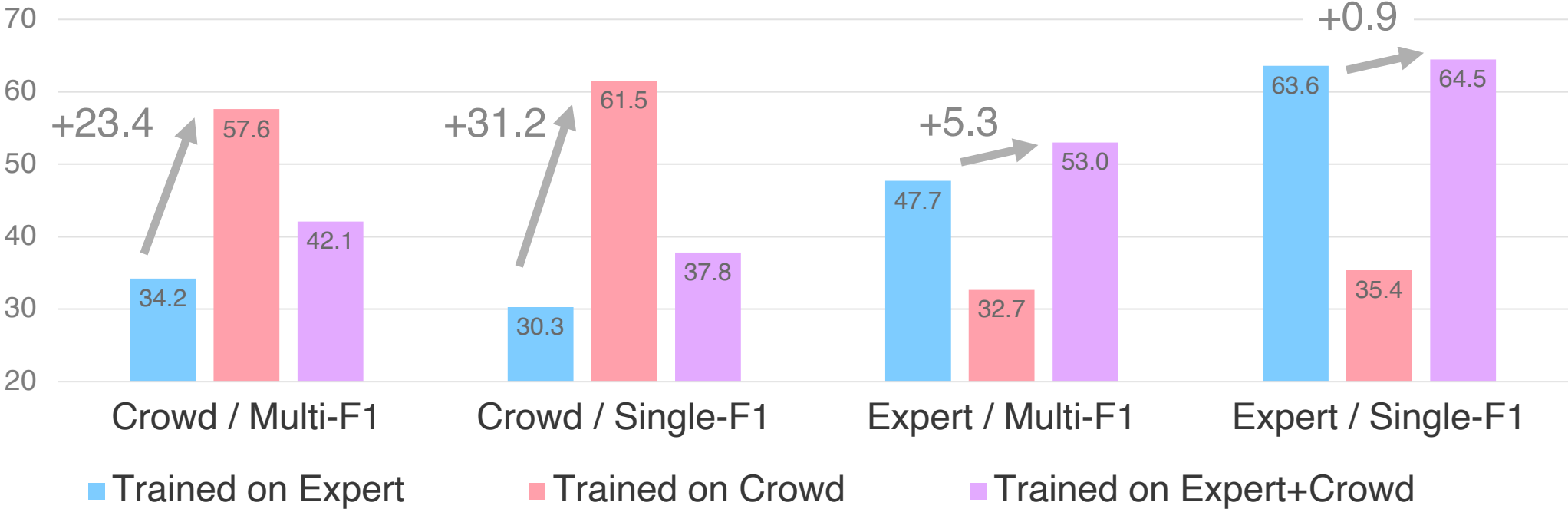
**Multi-F1** (threshold=7/16)

**Single-F1**

# Experimental Results

In all evaluation settings, adding Crowd improves F1 value

Multi-F1 and Single-F1 evaluated on Crowd and Expert



# Conclusion and future works

---

- To obtain continuous annotations for bridging reference resolution, we proposed to utilize crowdsourcing
- Experiments showed that collected data helps solve bridging resolution

## Future works

- Collect more examples for further improvement of bridging resolution
- Consider an effective way to combine Crowd and Expert

## Acknowledgments

- This work was supported by RIKEN