Improving Bridging Reference Resolution using Continuous Essentiality from Crowdsourcing

Nobuhiro Ueda and **Sadao Kurohashi**
Graduate School of Informatics, Kyoto University
{ueda, kuro}@nlp.ist.i.kyoto-u.ac.jp

Abstract

Bridging reference resolution is the task of finding nouns that complement essential information of another noun. The essentiality varies depending on noun combination and context and has a continuous distribution. Despite the continuous nature of essentiality, existing datasets of bridging reference have only a few coarse labels to represent the essentiality (Poesio and Artstein, 2008; Hangyo et al., 2012). In this work, we propose a crowdsourcing-based annotation method that considers continuous essentiality. In the crowdsourcing task, we asked workers to select both all nouns with a bridging reference relation and a noun with the highest essentiality among them. Combining these annotations, we can obtain continuous essentiality. Experimental results demonstrated that the constructed dataset improves bridging reference resolution performance. The code is available at https://github.com/nobu-g/bridging-resolution.

1 Introduction

The meaning of natural language texts is supported by cohesion among various linguistic units such as words, sentences, and paragraphs (Halliday and Hasan, 2014). Analyzing cohesion is indispensable for capturing the semantic structure of natural language texts.

Among cohesion analysis tasks, predicate-argument structure (PAS) analysis and semantic role labeling (SRL) have been actively studied (Shibata and Kurohashi, 2018; Omori and Komachi, 2019; He et al., 2018). These tasks aim to find nouns that complement a predicate's essential meaning, such as *who* does/did *what* to *whom*.

On the other hand, bridging reference resolution is the task of finding nouns that complement a noun's essential meaning. It is a special case of an anaphora resolution in which the anaphor and its antecedent have non-identical yet associated relations (Kobayashi and Ng, 2020).

(1) I can see a house over there. **The roof** is covered with snow.

In the above example, *the roof* is semantically insufficient by itself, and *a house* plays an essential role in complementing the meaning of *the roof*. Here, *the roof* is called an anaphor, and *a house* is called an antecedent. The performance of bridging reference resolution is only 40-60%, while PAS analysis and SRL have reached 70-90% (Ueda et al., 2020; Konno et al., 2021; Umakoshi et al., 2021; Zhang et al., 2021).

One challenge of bridging reference resolution is the continuous distribution of the strength of the relation between nouns (We call the strength **essentiality**, hereafter). In the example (2), *he*, *world swimming championships*, and *100m breaststroke* all modify *record* semantically but have different essentiality.

(2) He won the world swimming championships with a world **record** in 100m breaststroke.

The most essential information for *record* is what kind of event the *record* was set in, i.e., *100m breaststroke*. Although other phrases, *he* and *world swimming championships*, also complement the meaning of *record*, their essentiality is lower than that of *100m breaststroke*. Therefore, essentiality varies depending on noun combination and context and has a continuous distribution.

Although predicates and their modifiers also have essentiality, their continuity is less than that of nouns. Predicates have syntactically required highly essential modifiers called arguments. For example, intransitive verbs always have their subject, and transitive verbs always have their subject and object. In contrast to arguments, less essential modifiers are called adjuncts. The argument/adjunct distinction is ambiguous, especially in prepositional phrases. Thus the essentiality distributes continuously, like nouns. However, many of the modifiers

Proceedings of the Fifth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2022) , pages 74–87
Gyeongju, Republic of Korea, October 16–17, 2022.

are syntactically linked to the predicate. On the other hand, few nouns have such syntactic links, making it more difficult to distinguish between essential and non-essential due to many implicit modifiers.

Despite their continuous nature, existing datasets of bridging references have only a few coarse labels, such as *essential*, *ambiguous*, and *optional* (Poesio and Artstein, 2008; Hangyo et al., 2012). This fact suggests that there is a gap between the phenomenon of bridging reference and the annotations in existing datasets. This gap leads to performance degradation in bridging reference resolution.

In this work, we utilize crowdsourcing to obtain annotations in which continuous essentiality is considered. Crowdsourcing makes it possible to obtain multiple annotations for each example at a low cost. We asked crowd workers to select all nouns that have a bridging reference relation with a given noun. We also asked them to select the most essential one from the selected nouns. We assigned eight workers per example. Considering the number of votes as essentiality between nouns, we collected annotations of essentiality on a 16-point scale.

We used this method to create a corpus (**Crowd** hereafter) consisting of about 3,900 documents. Each document in **Crowd** consists of three sentences, which add up to 11,700 sentences. We compared **Crowd** with an existing corpus annotated with coarse labels by experts (**Expert** hereafter). In the experiment, we trained bridging reference resolution models on **Crowd**, **Expert**, and the combination of them. The models trained on **Crowd** or the combination always outperformed models trained only on **Expert**, which demonstrated the effectiveness of using **Crowd** as a training dataset. Our general-purpose crowdsourcing interface is publicly available for further research.[1] Our constructed dataset and training code are also publicly available.[2]

2 Existing Corpora for Bridging Reference

This section compares our dataset with existing corpora for bridging reference resolution. First, we introduce corpora for English bridging reference resolution, which is most actively studied, and then we describe Japanese corpora, which we compare in this work, in detail.

2.1 English Corpora

Some of the most widely used corpora in English are ARRAU (Poesio and Artstein, 2008), ISNotes (Markert et al., 2012), BASHI (Rösiger, 2018), and SciCorp (Roesiger, 2016). Most of these corpora contain only a few thousand bridging anaphors. Even the largest ARRAU contains 5,512 bridging anaphors, which is insufficient to apply neural network-based methods. Some works proposed data augmentation methods to address the issue. Hou (2020) converted examples into QA format and augmented the examples with existing QA datasets, and Yu and Poesio (2020) performed multi-task learning with coreference resolution. However, even with these methods, the accuracy is around 40–60%.[3] On the other hand, our corpus consists of 3,933 documents, including 25,217 bridging anaphors[4], which is large enough to train a neural network model. In addition, while all the four corpora have coarse labels to distinguish bridging reference relations, our corpus has more continuous annotations.

Recently, Elazar et al. (2022) created a corpus annotated with a wide range of noun phrase relations, including bridging reference. They annotated all noun phrase pairs whose relation type can be expressed by an English preposition. Their corpus comprises 5.5k documents covering over 1 million noun phrase relations. However, they do not deal with the strength of the relations. In addition, their annotation method relies heavily on English prepositions and does not apply to languages that do not have prepositions, such as Japanese (Masuoka and Takubo, 1992).

2.2 Japanese Corpora

There are two large corpora with bridging reference annotations in Japanese, KWDLC (Hangyo et al., 2012) and Kyoto Corpus (Kurohashi and Nagao, 2003; Kawahara et al., 2002). KWDLC consists of 5,124 documents containing 16,038 sentences annotated with various linguistic information, including bridging reference relations. Each

[1] https://github.com/nobu-g/bridging-annotation

[2] https://github.com/nobu-g/bridging-resolution

[3] This is the result in the setting of gold anaphors are given. The score would be even lower when anaphor detection is also performed.

[4] This is calculated for anaphors that at least half of the workers considered to be bridging.

label	example
essential	*Amerika no* ***shuto*** **the capital** of the US
ambiguous	*watashi no* ***megane*** **glasses** of mine
optional	*50 sento no* ***ame*** A 50 cent **candy**

Table 1: Labels of bridging reference relations defined in KWDLC (Hangyo et al., 2012).

document in KWDLC consists of the leading three sentences of web pages. Kyoto Corpus is also a corpus with various linguistic annotation but originated from newspaper articles. Kyoto Corpus has the same types of annotations as KWDLC, and bridging reference relations are annotated to 1,909 documents containing 15,872 sentences. This work focuses on KWDLC because of the diversity of texts it contains.

Both KWDLC and Kyoto Corpus have three types of labels for bridging reference relations: *essential*, *ambiguous*, and *optional*. These labels distinguish the strength of bridging reference relations (i.e., essentiality). Table 1 shows some examples. In the top example, the anaphor "the capital" is semantically insufficient by itself, and the antecedent "the US" makes up the insufficiency, which means "the US" has an *essential* relation for "the capital." *Optional* indicates the anaphor is already semantically sufficient by itself, or even if it is semantically insufficient, the antecedent does not make up the insufficiency. In the bottom example, "candy" is already semantically sufficient and the price is supplementary information. These two examples are typical, and there are many examples where it is hard to distinguish between *essential* and *optional*, and they are labeled as *ambiguous*.

3 Data Construction with Crowdsourcing

In Japanese, a noun pair which has a bridging reference relation can typically be connected by a genitive case "*no*."[5] In other words, when an anaphor *noun B* has a bridging reference relation with an antecedent *noun A*, "*A no B*" is a semantically valid noun phrase.

Although it is difficult for non-experts to judge whether two nouns have a bridging reference relation, they can judge whether "*A no B*" is a valid noun phrase. We showed crowd workers a text in which one word (i.e., *noun B*) was underlined. We asked them to select all words (i.e., *noun A*) where "*A no B*" is semantically valid, based on the contexts.

The *noun A*s selected by the workers have continuous latent values of essentiality for the *noun B*. In order to obtain the continuous essentiality values, we adopt the following strategies: (1) we asked workers to select the most essential noun for the *noun B*; (2) for each sample, we assigned eight workers to obtain multiple annotations.

We constructed the new corpus based on KWDLC (Hangyo et al., 2012) (i.e., **Expert**) in order to evaluate the quality of crowd workers' annotations. As shown in Table 2, we collected crowd workers' annotations (i.e., **Crowd**) for a subset of **Expert**, which correspond to approximately 77% of **Expert**.

We plan to make the annotations publicly available in the future. Workers agreed that the annotations will be used for academic research purposes in a non-personally identifiable manner.

3.1 Filtering Nouns to Annotate

In crowdsourcing, reducing the burden on workers leads to improved data quality. A possible burden in this task is the number of candidate noun pairs. **Expert** has an approximately 250 noun pairs per document, while only a few of them have bridging reference relations. So we used the following conditions to reduce the number of candidates of *noun B* and *noun A*.

The conditions of selecting *noun B*

- *noun B* is not a nominal predicate
- *noun B* is the tail noun if *noun B* is a part of a noun phrase
- *noun B* is not a numeral

The condition of selecting *noun A*

- *noun A* appears in the same or preceding sentence as *noun B*

Applying the above conditions reduced the number of candidate noun pairs by about 56%. Meanwhile, only 28% of the noun pairs in **Expert** with the relations of *essential*, *ambiguous*, or *optional* were excluded.

The conditions require linguistic features for each noun. We used the Japanese morphological

[5] "*no*" roughly corresponds to "of" in English, but has a broader usage than "of."

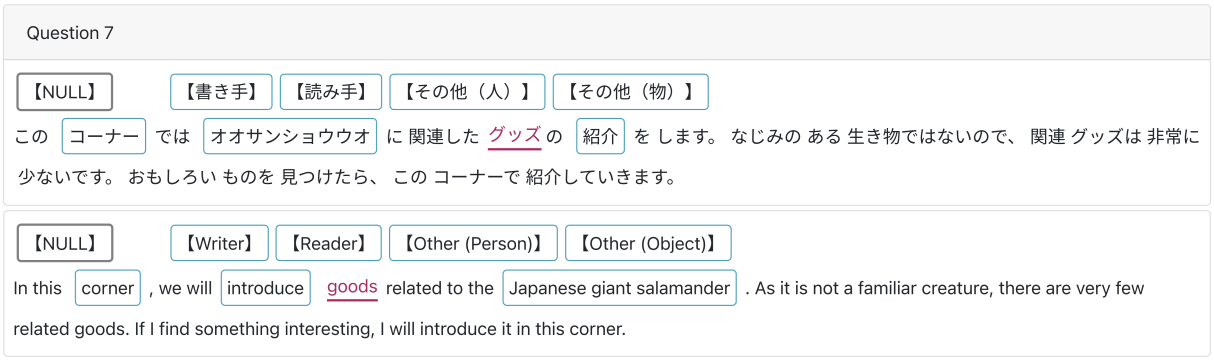

Figure 1: A sample question in our crowdsourcing interface. The upper one is from the original interface, and the lower one is the English translation. Workers select *noun A*s from framed words so that the *noun A*s have a relation of "*A no B*" ("*B of A*" in English) for the *noun B* (red underlined word). Workers can select *noun A*s easily by clicking the framed words.

corpus	train	dev	test
Expert	3,912	512	700
Crowd	2,721	512	700

Table 2: The number of documents contained in each corpus. **Expert** provides an official split and we split **Crowd** following **Expert**.

analyzer Juman++ (Morita et al., 2015; Tolmachev et al., 2018) and the Japanese syntactic analyzer KNP (Kurohashi and Nagao, 1994) to obtain the features. Juman++ performs morphological segmentation and assigns linguistic features such as parts of speech to morphemes. Based on the features from Juman++, KNP identifies noun phrases and assigns linguistic features to them.

3.2 Special Targets

Following the setting in **Expert** corpus (Hangyo et al., 2012), we introduce five special targets. Workers can select the special targets in addition to the nouns in a text. The first is the `[NULL]`, which is selected when none of the nouns in a text is related to the *noun B*. Introducing the `[NULL]` target enables us to require workers to answer all questions, which is expected to prevent workers from skipping questions.

The others are used for collecting annotations for exophora. Exophora is a reference to entities that do not appear in the text. In Japanese, exophora occurs 13% of all the bridging references. As exophora has no definite textual antecedents, we introduce the following four typical types of exophora.

- `[Writer]`:
  The one who wrote the text
- `[Reader]`:
  Someone who would read the text
- `[Other:Person]`:
  Someone except for the above
- `[Other:Object]`:
  Some entity external to text

Hereafter, we refer to the reference targets, including these special targets, as *noun A*s.

3.3 Crowdsourcing Interface

An annotation interface also plays an essential role in reducing the burden on crowd workers. However, existing crowdsourcing platforms of Japanese[6] do not provide an interface flexible enough to conduct this task. Therefore, we developed our own interface and directed workers to the interface from the existing platform.

Figure 1 shows one question sample of our interface. In this sample, the underlined red word corresponds to *noun B*, and the words with blue frames correspond to *noun A* candidates. For the given *noun B*, workers click to select *noun A*s from the framed words. By clicking on one of the selected words twice, the workers can select it as the most essential noun. If they select `[NULL]`, they can select none of the other words, and there is no need to select the most essential noun.

In addition to the question part, our interface consists of task instructions and practice questions. Workers first read the task instructions, solve the practice questions, and then start annotation. Appendix A.2 shows the interface of the task instructions and the practice questions.

[6] https://crowdsourcing.yahoo.co.jp/ https://crowdworks.jp/

	Multi				Single			
	Multi-Prec.	Multi-Rec.	Multi-F1	mAP	Single-Prec.	Single-Rec.	Single-F1	Acc.
Endophora	38.4	40.1	39.2	30.0	29.9	71.6	42.2	58.4
Exophora	12.2	20.7	15.4	7.3	6.7	48.9	11.8	52.9

Table 3: The evaluation result of **Crowd**, considering **Expert** as the gold. Endophora is a reference to words that appears in the text. Exophora is a reference to entities that do not appear in the text.

3.4 Cost of the Data Construction

We used Yahoo! Crowdsourcing[7] as our crowdsourcing platform. It charges 17.7 yen per task, including the commission fee. Overall, the cost of the data construction was approximately 580,000 yen for 31,200 tasks. In contrast, the cost of constructing **Expert** was over 6,000,000 yen. Although the cost is not directly comparable because **Expert** has other types of linguistic annotations besides bridging reference relations, the cost for **Crowd** would be less than half of that for **Expert**.

4 Corpus Evaluation

In order to verify the quality of the constructed corpus (i.e., **Crowd**), we compared it with the corpus with expert annotations (i.e., **Expert**). For the quantitative evaluation, we define **essentiality score** for a *noun A* in **Crowd** as follows.

$$\begin{cases} n(A) \times 2 & \text{if } \textit{noun A} \text{ is } \texttt{[NULL]}, \\ n(A) + N(A) & \text{otherwise,} \end{cases} \quad (1)$$

where $n(A)$ denotes the number of workers who selected *noun A*, and $N(A)$ denotes the number of workers who selected *noun A* as the most essential noun. In this work, since eight workers annotated each noun pair, essentiality score takes values from 0 to 16. Since workers cannot select `[NULL]` as the most essential noun, we double $n(\texttt{[NULL]})$ for normalization.

4.1 Evaluation Metrics

We want to evaluate whether essentiality score reflects the essentiality for the *noun B*. To evaluate **Crowd** in this criteria, we assumed **Expert** as the ground truth and calculated Multi-F1, Single-F1, mean average precision (mAP), and accuracy.

Multi-F1 is an F measure to evaluate how well all the *noun A*s with bridging reference relations are selected. Multi-F1 measures the ability of a model to find *noun A*s with less essential relations as well as the most essential relation. We defined a threshold and selected *noun A*s that many workers selected. And then, we calculated precision and recall for the selected *noun A*s, regarding *noun A*s annotated as *essential* or *ambiguous* in **Expert** as positive. Multi-F1 is the harmonic mean of the precision and recall. We varied the threshold from 0 to 16 and picked the one with the highest Multi-F1 value. The threshold obtained was 7.

Single-F1 is an F measure to evaluate how well the most essential *noun A* is selected. In Single-F1, we consider one *noun A* for each *noun B*. For *noun A*s in **Crowd**, we select the one with the highest essentiality score. For *noun A*s in **Expert**, we select the one annotated as *essential*.[8] If none of the nouns are annotated as *essential*, we select the one annotated as *ambiguous*. In Single-F1, we ignored `[NULL]`, that is, we did not count `[NULL]` as a true positive or false positive.

Mean average precision (mAP) is the mean of average precision (AP) over all *noun B*s. AP is the average of precision at each recall value varying the threshold. The precision and the recall are defined in the same way as Multi-F1, but without a need to set a threshold. Accuracy is calculated, including `[NULL]`.

4.2 Evaluation Results

First, we evaluate **Crowd** in comparison to **Expert**. Table 3 shows the scores of each evaluation metric. In general, recall tends to be higher than precision, demonstrating that our method enabled us to collect a broader range of examples than the experts' annotation.

However, the precision, especially the Single-Precision of exophora, was considerably low. This result is partly due to the nature of `[Other:Person]` and `[Other:Object]`. Since, in most cases, an entity is owned by someone or is part of something, we can say that the entity has a bridging relation with `[Other:Person]`

[7] https://crowdsourcing.yahoo.co.jp/

[8] When multiple nouns are annotated as *essential*, we prioritize the one that most crowd workers selected.

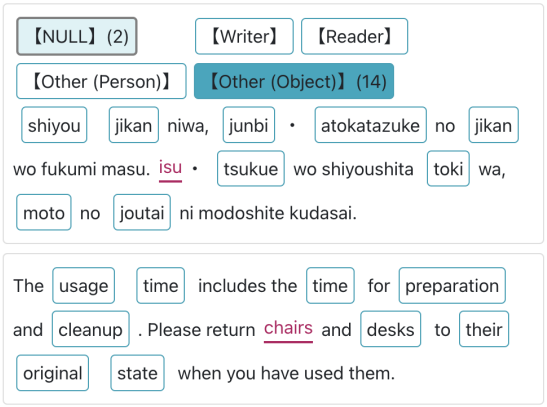

Figure 2: An example of collected annotations in Japanese (upper) and its English translation (lower). The numbers in parentheses and the color intensity indicates essentiality score.

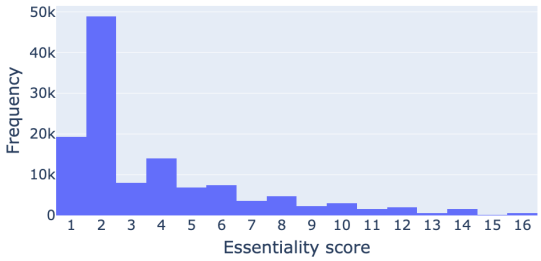

Figure 3: The distribution of essentiality score in **Crowd**.

or `[Other:Object]`. Although experts did not annotate such general modifiers because they are too obvious, many crowd workers did. In the example in Figure 2, many workers selected `[Other:Object]` because "chairs" are considered to be chairs of some facility, while experts annotated nothing.

Next, we evaluate **Crowd** itself. For each noun pair, we can formulate this task as a three-class classification: *select*, *select as the most essential noun*, and *do not select*. This formulation enables us to calculate Krippendorff's alpha (Krippendorff, 2018) to measure the inter-worker agreement. We found it to be about 0.28. For a more intuitive agreement measure, 57% of all workers selected *noun A* with the highest number of votes, and 52% selected such *noun A* as the most essential noun. Although this value is relatively low for an inter-annotator agreement, this is a minor problem because our purpose is to obtain diverse annotations for this inherently subjective task.

We can see the diversity of annotations in Figure 3. It shows the distribution of essentiality score

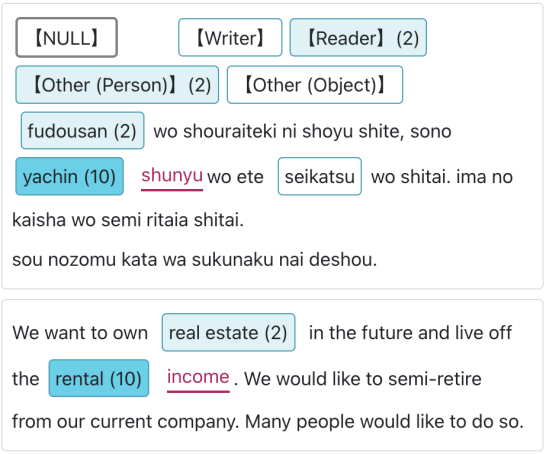

Figure 4: An example of collected annotations. The format is the same as Figure 2.

for *noun A*s except for `[NULL]`. *Noun A*s whose essentiality score is 0 are excluded because they are too frequent (449k). The figure shows high frequency of *noun A*s with low essentiality score. This means that the continuous nature of essentiality is reflected as the diversity of essentiality score in **Crowd**. We can also see the characteristic that even essentiality score is more frequent than odd one. This is because many workers selected only one *noun A*. When a worker selects only one *noun A*, it is necessarily the most essential noun, and the essentiality score increases by 2.

Figure 4 shows another collected example. The selected *noun A*s, *Reader*, *Other (Person)*, *real estate*, and *rental*, are all related to the *noun B*, *income*. In addition, the *noun A* with the highest essentiality score is *rental*, which is considered to be the most essential information for *income*. This example shows that in our corpus, the essentiality of nouns is represented as the number of crowd workers' votes.

5 Evaluation with Bridging Reference Resolution

In this section, we examine the effectiveness of **Crowd** for improving bridging reference resolution performance through evaluation experiments. In the experiments, we use three kinds of corpora, **Expert**, **Crowd**, and the combination of **Expert** and **Crowd**, as the training data. We compare the performance of the models trained on each corpus.

5.1 Task Definition

In Japanese, the formulation of bridging reference resolution is different from the one in English.

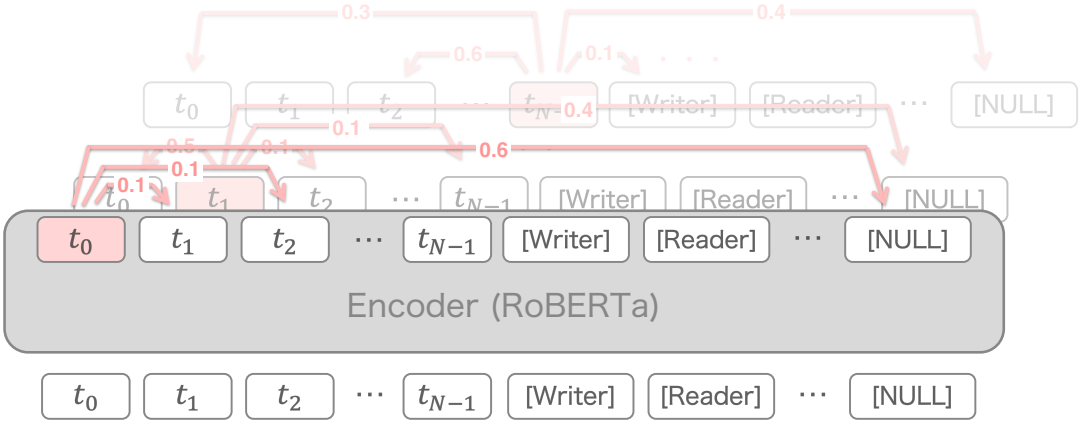

Figure 5: An overview of our model. For an input sequence of length N, the model outputs $N \times N$ values. Note that five special targets, `[Writer]`, `[Reader]`, `[Other:Person]`, `[Other:Object]`, and `[NULL]`, are appended to the input text.

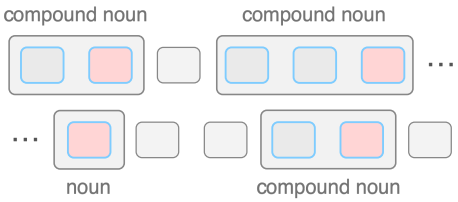

Figure 6: A simplified illustration of how we select *noun A*s and *noun B*s. Each cell denotes a word. The cells colored with light red and bordered with light blue are selected as *noun B*s and *noun A*s, respectively.

label	value
essential	1.0
ambiguous	0.5
optional	0.25

Table 4: The label conversion table in **Expert**.

In English, the task is formulated as a span prediction problem, similar to coreference resolution. In Japanese, on the other hand, it is formulated as a word prediction problem. Therefore, we can solve the task by performing binary classification for each noun pair in a text. Figure 6 shows an illustration of how we select target nouns.

Bridging reference resolution consists of two sub tasks: bridging anaphor recognition and antecedent selection. Many studies refer to bridging reference resolution (or bridging anaphora resolution) as a task of antecedent selection, that is, the gold anaphors are given (Kobayashi and Ng, 2020; Hou, 2020, 2018). In this work, we tackle full bridging resolution, in which we perform bridging anaphor recognition as well as antecedent selection.[9] Instead of performing bridging anaphor recognition, we use `[NULL]` as a special antecedent and perform antecedent selection for all *noun B*s. Reference to `[NULL]` means that the *noun B* is not an anaphor, similar to the data construction stage.

Furthermore, we also consider bridging exophora resolution. In a similar manner to full bridging resolution described above, we use additional special targets, `[Writer]`, `[Reader]`, `[Other:Person]`, `[Other:Object]`, and `[NULL]`.

5.2 Label Conversion

For the comparison between **Crowd** and **Expert**, we need to treat both corpora in a common framework. For this purpose, we convert the relation between each noun in both corpora into a value between 0 and 1, called **normalized essentiality score**. For **Crowd**, we just normalize essentiality score by dividing it by its maximum value, 16. For **Expert**, since the relation is defined as a label rather than a value, we define the label to value mapping heuristically as shown in Table 4.

5.3 Resolution Method

We train a model that outputs normalized essentiality scores for each token pair as shown in Figure 5. $s(t_a, t_b)$, the normalized essentiality score

[9] Because we limit anaphors and antecedents by the rule described in section 3.1, our task is a little easier than full bridging resolution.

Training Corpus	Evaluated on Crowd Multi-F1 (Prec. / Rec.)	Single-F1 (Prec. / Rec.)	mAP	Spearman
Expert	34.2 ± 7.0 (30.4 / 39.2)	30.3 ± 1.3 (63.2 / 19.9)	24.9 ± 5.1	36.5 ± 1.6
Crowd	**57.6 ± 1.3** (58.7 / 56.5)	**61.5 ± 1.0** (62.3 / 60.8)	**60.8 ± 1.0**	**53.3 ± 0.4**
Crowd+Expert (MR)	38.7 ± 4.3 (36.8 / 41.5)	42.8 ± 0.7 (74.7 / 30.0)	34.8 ± 5.2	46.9 ± 4.4
Crowd+Expert (MSE)	42.1 ± 3.9 (40.7 / 44.1)	37.8 ± 0.9 (69.5 / 26.0)	41.7 ± 1.7	49.2 ± 0.6

Table 5: Results of bridging reference resolution evaluated on **Crowd** corpora (%). The scores are the mean and 95% confidence interval over three training runs with different random seeds. MR and MSE represent the model is trained using MR loss and MSE loss, respectively.

Training Corpus	Evaluated on Expert Multi-F1 (Prec. / Rec.)	Single-F1 (Prec. / Rec.)	mAP	Spearman
Expert	47.7 ± 1.1 (50.6 / 45.3)	63.6 ± 1.4 (66.7 / 60.8)	43.2 ± 2.3	43.7 ± 0.3
Crowd	32.7 ± 0.7 (33.0 / 32.3)	35.4 ± 1.2 (24.6 / 63.3)	27.3 ± 0.1	36.8 ± 0.5
Crowd+Expert (MR)	48.3 ± 2.5 (52.1 / 45.0)	62.8 ± 0.5 (59.8 / 66.1)	45.4 ± 3.1	42.4 ± 0.9
Crowd+Expert (MSE)	**53.0 ± 1.8** (57.5 / 49.3)	**64.5 ± 1.8** (63.6 / 65.5)	**52.2 ± 2.0**	**43.8 ± 0.3**

Table 6: Results of bridging reference resolution evaluated on **Expert** corpora (%). The representations are the same as in Table 5.

of token t_a for token t_b, is calculated as follows:

$$s(t_b, t_a) = \boldsymbol{v}^{\mathrm{T}} \tanh(W_1 \boldsymbol{t}_b + W_2 \boldsymbol{t}_a), \quad (2)$$

where W_1 and W_2 denote weight matrices. $\boldsymbol{v}$ is a weight vector. $\boldsymbol{t}_b$ and $\boldsymbol{t}_a$ denote hidden vectors of the encoder's final layer corresponding to the tokens t_b and t_a. The encoder was RoBERTa (Liu et al., 2019) model that has been pre-trained with Japanese web texts.[10]

In addition to a tokenized text, the encoder's input sequence contains special tokens at the end of the sequence, similar to Ueda et al. (2020). The special tokens are `[Reader]`, `[Writer]`, `[Other:Person]`, `[Other:Object]`, and `[NULL]`.

5.4 Training Objective

When training on **Crowd**, we employed mean squared error loss (MSE loss) as the loss function. As shown in the following equation, for each t_b and t_a, MSE loss optimizes the system output $s(t_b, t_a)$ to be close to the normalized essentiality score $e(t_b, t_a)$.

$$\mathcal{L}_{MSE} = \frac{1}{Z} \sum_{a,b} \Big(s(t_b, t_a) - e(t_b, t_a) \Big)^2, \quad (3)$$

where t_b and t_a are tokens in a input sequence. $s(t_b, t_a)$ and $e(t_b, t_a)$ are the system output and normalized essentiality score, respectively. Z is the normalization term.

When training on **Expert**, we employed margin ranking loss (MR loss) because we expected that using ranking-based loss mitigates the bias of arbitrarily defined values in the label conversion stage.[11] MR loss is calculated as follows:

$$\begin{aligned} \mathcal{L}_{MR} &= \frac{1}{Z} \sum_{a,b,c<a} \max(0, d_{abc}), \\ d_{abc} &= \mathrm{sign}(\Delta e_{abc}) \cdot (-\Delta s_{abc} + \Delta e_{abc}), \\ \Delta s_{abc} &= s(t_b, t_a) - s(t_b, t_c), \\ \Delta e_{abc} &= e(t_b, t_a) - e(t_b, t_c). \end{aligned}$$

MR loss optimizes the difference of the system outputs and normalized essentiality scores (Δs and Δe, respectively) rather than the values themselves. This way of optimization avoids forcing the model to output arbitrarily defined discrete values. See Appendix A.1 for more details on the implementation.

5.5 Experimental Results

Table 5,6 shows the results of the experiments when **Crowd**, **Expert**, and both are used for the training. For the evaluation metrics, in addition to Multi-F1, Single-F1, and mAP described in section 4, we used Spearman's rank correlation coefficient to measure the ranking-based agreement.

[10] https://huggingface.co/nlp-waseda/roberta-base-japanese

[11] Using MR loss showed better performances than using MSE loss in our preliminary experiments.

For Multi-F1, Single-F1, and mAP, we ignored [NULL]. Table 5 shows adding **Crowd** to the training data improved the performance by 8–15 points in all the evaluation metrics. Furthermore, when training with only **Crowd**, the performance was even higher. It makes sense that the model shows the higher performance when the evaluation set's data distribution matches the training set's distribution.

Moreover, we also confirmed the effectiveness of **Crowd** when evaluated on **Expert** (Table 6). Although training with only **Crowd** did not improve the performance, training with both **Crowd** and **Expert** improved the performance compared to only using **Expert**. Especially, the performance of Multi-F1 and mAP improved by 5.3 and 9.0 points, respectively.

The performance improvements in Multi-F1 are due to the high coverage of the relations annotated in **Crowd**. This high coverage is an advantage of crowdsourcing, which enables us to obtain diverse annotations by many people at a reasonable cost. The performance improvements in mAP and Spearman's rank correlation coefficient are due to annotations in **Crowd**, in which the continuous nature of essentiality are represented precisely.

6 Conclusion

In the existing datasets of bridging reference resolution, the strength of relations between nouns was unnaturally represented by coarse-grained labels despite the continuous distribution of the strength. We focused on this gap and proposed a crowdsourcing-based annotation method to construct a dataset with more continuous annotations. We have developed a general-purpose interface for the data collection with crowdsourcing. This interface can be applied to crowdsourcing not only for bridging reference resolution but also for any relational analysis tasks.

By training with our newly constructed dataset, we improved the performance of bridging reference resolution on a Japanese standard benchmark dataset. Moreover, the performance improvement was over 16 points, evaluated on the constructed dataset.

References

Lukas Biewald. 2020. Experiment tracking with weights and biases. Software available from wandb.com.

Yanai Elazar, Victoria Basmov*, Yoav Goldberg, and Reut Tsarfaty. 2022. Text-based NP enrichment. *Transactions of the Association for Computational Linguistics*, 10:764–784.

Michael Alexander Kirkwood Halliday and Ruqaiya Hasan. 2014. *Cohesion in English*. 9. Routledge.

Masatsugu Hangyo, Daisuke Kawahara, and Sadao Kurohashi. 2012. Building a diverse document leads corpus annotated with semantic relations. In *Proceedings of PACLIC*, pages 535–544.

Shexia He, Zuchao Li, Hai Zhao, and Hongxiao Bai. 2018. Syntax for semantic role labeling, to be, or not to be. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2061–2071, Melbourne, Australia. Association for Computational Linguistics.

Yufang Hou. 2018. Enhanced word representations for bridging anaphora resolution. 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 1–7, New Orleans, Louisiana. Association for Computational Linguistics.

Yufang Hou. 2020. Bridging anaphora resolution as question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1428–1438, Online. Association for Computational Linguistics.

Daisuke Kawahara, Sadao Kurohashi, and Kôiti Hasida. 2002. Construction of a Japanese relevance-tagged corpus. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*. European Language Resources Association (ELRA).

Hideo Kobayashi and Vincent Ng. 2020. Bridging resolution: A survey of the state of the art. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3708–3721, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ryuto Konno, Shun Kiyono, Yuichiroh Matsubayashi, Hiroki Ouchi, and Kentaro Inui. 2021. Pseudo zero pronoun resolution improves zero anaphora resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3790–3806. Association for Computational Linguistics.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Sadao Kurohashi and Makoto Nagao. 1994. A syntactic analysis method of long Japanese sentences based on the detection of conjunctive structures. *Computational Linguistics*, 20(4).

Sadao Kurohashi and Makoto Nagao. 2003. Building a Japanese parsed corpus. In *Treebanks*, pages 249–260. Springer.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Katja Markert, Yufang Hou, and Michael Strube. 2012. Collective classification for fine-grained information status. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 795–804, Jeju Island, Korea. Association for Computational Linguistics.

Takashi Masuoka and Yukinori Takubo. 1992. *Kiso Nihongo bunpō: kaiteiban*. Kuroshio shuppan.

Hajime Morita, Daisuke Kawahara, and Sadao Kurohashi. 2015. Morphological analysis for unsegmented languages using recurrent neural network language model. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2292–2297, Lisbon, Portugal. Association for Computational Linguistics.

Hikaru Omori and Mamoru Komachi. 2019. Multi-task learning for Japanese predicate argument structure analysis. In *Proceedings of NAACL*, pages 3404–3414.

Massimo Poesio and Ron Artstein. 2008. Anaphoric annotation in the ARRAU corpus. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Ina Roesiger. 2016. SciCorp: A corpus of English scientific articles annotated for information status analysis. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1743–1749, Portorož, Slovenia. European Language Resources Association (ELRA).

Ina Rösiger. 2018. BASHI: A corpus of Wall Street Journal articles annotated with bridging links. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Tomohide Shibata and Sadao Kurohashi. 2018. Entity-centric joint modeling of Japanese coreference resolution and predicate argument structure analysis. In *Proceedings of ACL*, pages 579–589.

Arseny Tolmachev, Daisuke Kawahara, and Sadao Kurohashi. 2018. Juman++: A morphological analysis toolkit for scriptio continua. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 54–59, Brussels, Belgium.

Nobuhiro Ueda, Daisuke Kawahara, and Sadao Kurohashi. 2020. BERT-based cohesion analysis of Japanese texts. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1323–1333, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Masato Umakoshi, Yugo Murawaki, and Sadao Kurohashi. 2021. Japanese zero anaphora resolution can benefit from parallel texts through neural transfer learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1920–1934, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Juntao Yu and Massimo Poesio. 2020. Multi-task learning based neural bridging reference resolution. *ArXiv*, abs/2003.03666.

Yu Zhang, Qingrong Xia, Shilin Zhou, Yong Jiang, Zhenghua Li, Guohong Fu, and Min Zhang. 2021. Semantic role labeling as dependency parsing: Exploring latent tree structures inside arguments.

A Appendix

A.1 Implementation Details

Table 7–10 show the hyperparameters used in our experiments. We tuned training epochs, learning rate, and scheduler warmup steps based on the mAP score on the validation set. Learning rate was selected from {0.00001, 0.00005, 0.0001, 0.0002} in all the experiments. Training epochs was selected from {12, 16, 20, 24} when training on **Expert**, {16, 20, 24, 28} when training on **Crowd**, and {12, 16, 20} when training on the combination of **Expert** and **Crowd**. Scheduler warmup steps was selected from {160, 200, 240, 280} when training on **Expert**, {100, 140, 180, 220} when training on **Crowd**, and {200, 240, 280, 320} when training on **Expert** and **Crowd**. We used Weights and Biases (Biewald, 2020) for the hyperparameter tuning.

The computation was performed on NVIDIA TITAN X (Pascal) or NVIDIA GeForce RTX 2080 Ti GPUs. Each training run took about 0.5–2 hours on two GPUs.

[12]This scheduler is implemented in Transformers (Wolf et al., 2019) and we used it.

Parameter name	Parameter value
Optimizer	AdamW
Training epochs	20
Learning rate	2×10^{-4}
Optimizer eps	1×10^{-8}
Weight decay	0.01
Dropout rate (RoBERTa layer)	0.1
Dropout rate (output layer)	0.0
LR scheduler	linear_schedule_with_warmup[12]
Scheduler warmup steps	160
Batch size	32

Table 7: Hyperparameters used for training on **Expert**.

Parameter name	Parameter value
Optimizer	AdamW
Training epochs	28
Learning rate	1×10^{-4}
Optimizer eps	1×10^{-8}
Weight decay	0.01
Dropout rate (RoBERTa layer)	0.1
Dropout rate (output layer)	0.0
LR scheduler	linear_schedule_with_warmup
Scheduler warmup steps	180
Batch size	32

Table 8: Hyperparameters used for training on **Crowd**.

Parameter name	Parameter value
Optimizer	AdamW
Training epochs	16
Learning rate	1×10^{-4}
Optimizer eps	1×10^{-8}
Weight decay	0.01
Dropout rate (RoBERTa layer)	0.1
Dropout rate (output layer)	0.0
LR scheduler	linear_schedule_with_warmup
Scheduler warmup steps	240
Batch size	32

Table 9: Hyperparameters used for training on the combination of **Expert** and **Crowd** with margin ranking loss.

Parameter name	Parameter value
Optimizer	AdamW
Training epochs	20
Learning rate	5×10^{-5}
Optimizer eps	1×10^{-8}
Weight decay	0.01
Dropout rate (RoBERTa layer)	0.1
Dropout rate (output layer)	0.0
LR scheduler	linear_schedule_with_warmup
Scheduler warmup steps	280
Batch size	32

Table 10: Hyperparameters used for training on the combination of **Expert** and **Crowd** with mean squared error loss.

A.2 Crowdsourcing Instructions and Practice Questions

Figure 7,8 show the task instructions, and Figure 9 shows the practice questions. After reading the task instructions, workers need to answer the practice questions which they can answer as many times as they want until they answer correctly.

タスク説明

▼ 概要

下線を引いた赤字の単語 △△ について、 文章中の枠線で囲まれた単語 ○○ の中から**「○○の△△」**という関係が成り立つ単語を**全て**選んでください。
例えば次の文では、「太郎の先生」「英語の先生」という関係が成り立つので、「太郎」「英語」を選択します。

昨日 、 太郎 は 英語 の 先生 に質問した。 ⇒ 昨日 、 太郎 は 英語 の 先生 に質問した。

ここで、単語 ○○ は単語 △△ から簡単に連想できる単語としてください。 具体的には、まず単語 △△ から**「○○の△△」と連想できる**単語を考えてください。

先生 ⇒ 教科(数学、英語...) 生徒(○○君、□□さん...) 場所(小学校、教習所...)

そして、選択肢の中に連想した単語（もしくは同じような単語）があればそれを選択します。 さらに、単語 △△ にとって**最も重要、あるいは必須**と考えられる単語も同時に選んでください。
上記の例では、**何の教科の先生なのか**が重要かつ必須的な情報なので、「英語」をもう1度クリックして選択します。 選ぶのが難しい場合は、一番最初に連想した単語を選んでも構いません。

昨日 、 太郎 は 英語 の 先生 に質問した。 ⇒ 昨日 、 太郎 は 英語 の 先生 に質問した。

単語 △△ からの連想の他の例を示します。

屋根 ⇒ 建物(民家、ガレージ...)

社員 ⇒ 会社(○○グループ、株式会社□□...)

記録 ⇒ 種目(マラソン、水泳...) 出来事(戦争、災害...) 保持者(高橋尚子、北島康介...)

問題は練習問題3問を含む全13問です。 全ての問題に回答し、「送信」ボタンを押すとタスク終了です。

▼ 注意点

- 「○○の△△」が**意味的に正しくなる**ような単語のみを選んでください。
  次の例では、下線部の「先生」は「太郎」が教わっている先生ではないため、「太郎の先生」は意味的に正しくありません。 したがって、この文では「太郎」は選択しません。

  太郎 は 英語 の 先生 になるのが夢だ。 ⇒ 太郎 は 英語 の 先生 になるのが夢だ。

- 連想した単語が**原文に存在しない**問題も多くあります。
  その単語が「私」など、原文の書き手であれば 【書き手】 を、 反対に「あなた」など、原文の読み手であれば 【読み手】 を選んでください。 どちらにも当てはまらない場合は、連想した単語が人か物かによって 【その他（人）】 または 【その他（物）】 を選んでください。

  【書き手】
  今日 は 先日 生まれた 息子 を紹介したいと思います。 ⇒ 【書き手】 今日 は 先日 生まれた 息子 を紹介したいと思います。

  【その他（物）】
  階段 を登って 街並み を 屋根 から見渡した。 ⇒ 【その他（物）】 階段 を登って 街並み を 屋根 から見渡した。

- 以下のように単語 ○○ が**連想しにくい**名詞も多くあります。 この場合は 【該当なし】 を選んでください。
  その他、選択が難しい場合も 【該当なし】 を選んでください。

  ピアニスト ⇒ ?

  政治家 ⇒ ?

  東京タワー ⇒ ?

  太郎 ⇒ ?

- 問題文はウェブサイトの文章を切り取ったものです。 文脈が不足している場合は、適宜話題を推測しつつお答えください。

Figure 7: A screen capture of the instruction page of our crowdsourcing task (1/2).

Figure 8: A screen capture of the instruction page of our crowdsourcing task (2/2).

Figure 9: A screen capture of the practice questions page of our crowdsourcing task.