End-to-End Neural Bridging Resolution

Hideo Kobayashi[1], **Yufang Hou**[2] and **Vincent Ng**[1]
[1] Human Language Technology Research Institute, University of Texas at Dallas, USA
[2] IBM Research Europe, Ireland
{hideo,vince}@hlt.utdallas.edu
yhou@ie.ibm.com

Abstract

The state of bridging resolution research is rather unsatisfactory: not only are state-of-the-art resolvers evaluated in unrealistic settings, but the neural models underlying these resolvers are weaker than those used for entity coreference resolution. In light of these problems, we evaluate bridging resolvers in an end-to-end setting, strengthen them with better encoders, and attempt to gain a better understanding of them via perturbation experiments and a manual analysis of their outputs.

1 Introduction

Bridging was used by Clark (1975) to refer to non-identity relations between anaphoric noun phrases (i.e., **bridging anaphors**) and their antecedents. In Example 1, "**ruling – party members**" is a bridging anaphor and its antecedent is "*Japan*".

(1) Yet another political scandal is racking *Japan*. But this time it's hurting opposition as well as **ruling - party members**.

Bridging resolution is the task of identifying bridging anaphors and linking them to their antecedents. Many tasks can benefit from bridging resolution, such as textual entailment (Mirkin et al., 2010) and question answering (Tseng et al., 2021).

Bridging resolution is arguably less studied but more challenging that entity coreference resolution, the task of determining which entity mentions refer to the same entity in the real world. Specifically, while linguistic constraints on coreference exist at the grammatical (e.g., gender and number agreement), syntactic (e.g., c-command), and semantic (e.g., semantic type agreement) levels that can be used to filter candidate antecedents, such constraints are largely absent for bridging resolution. For instance, a singular bridging anaphor (e.g., "the book") can refer to a plural antecedent (e.g., "books"), and bridging relations can be formed from mentions with different entity types (e.g., "the house" and "the window"). In fact, while many coreference relations can be identified via string matching facilities, it is not uncommon for bridging relations to be identified using background knowledge and/or sophisticated inference mechanisms. The complexity of bridging resolution is further complicated by the lack of a large corpus annotated with bridging relations: while the most extensively-used coreference-annotated corpus, OntoNotes, contain more than 2000 documents, two of the most commonly-used corpora for bridging resolution, ISNotes and BASHI, each contains only 50 documents taken from OntoNotes.

The current state of bridging resolution research is rather unsatisfactory. State-of-the-art bridging resolvers are typically evaluated in unrealistic settings: in the *bridging anaphora resolution* task, the goal is to identify the antecedent of a given bridging anaphor; and in the *full bridging resolution* task, the goal is to first identify the bridging anaphors given a set of gold mentions and then resolve each anaphor to its antecedent (Yu and Poesio, 2020). While such unrealistic settings have been considered unacceptable for evaluating entity coreference resolvers for more than a decade, they are still extensively used to evaluate bridging resolvers nowadays simply because the *end-to-end* setting, where a resolver needs to identify bridging relations given a raw document, is perceived to be overly challenging. Worse still, while bridging resolution is more challenging than coreference resolution, models for bridging resolution are less sophisticated than those for coreference resolution. For instance, while SpanBERT, a version of BERT specifically pre-trained to identify text spans (Joshi et al., 2020), has been used successfully as an encoder in span-based entity coreference models, the state-of-the-art neural bridging resolver developed by Yu and Poesio (2020) simply uses a bidirectional LSTM to encode the input document.

Our goal in this paper is to gain a better un-

derstanding of the state of the art in bridging resolution. First, we conduct a systematic evaluation of bridging resolvers in an *end-to-end* setting. In particular, we focus on evaluating three state-of-the-art approaches, including a rule-based approach by Rösiger et al. (2018), a neural approach by Yu and Poesio (2020), and a hybrid rule-based and learning-based approach by Kobayashi and Ng (2021), showing how each of them can be extended so that they can be applied in an end-to-end setting. Next, we strengthen Yu and Poesio's (2020) neural model by replacing the biLSTM encoder it uses with stronger encoders such as BERT (Devlin et al., 2019) and SpanBERT (Joshi et al., 2020). These experiments can help us determine whether the commonsense knowledge encoded in these pre-trained language models can be profitably exploited for bridging resolution, which could be important given that state-of-the-art bridging resolvers are trained on small annotated corpora. Further, to better understand the extent to which a bridging resolver relies on certain words/phrases in the input, we conduct perturbation experiments. Finally, to complement the quantitative analysis in our experiments, we conduct a qualitative analysis of the outputs produced by our best-performing resolver.

Our contributions in this paper are three-fold. First, our experiments reveal that end-to-end bridging resolution is not as challenging as typically perceived: for the most part, end-to-end bridging resolution lags behind its "gold mention" counterpart by less than 3 points in absolute F-score. These results provide suggestive evidence that time is ripe for abandoning unrealistic evaluations of bridging resolvers. Second, we establish baseline results for end-to-end bridging resolution against which future work can be compared on two commonly-used referential bridging corpora, ISNotes and BASHI. While our evaluations have largely focused on the end-to-end setting, for comparison purposes we also present evaluation results in the "gold mention" setting, in which our strongest models achieve state-of-the-art results on ISNotes and BASHI.

2 Related Work

Many previous computational studies on bridging have focused on one of the two sub-tasks of bridging resolution, namely *bridging anaphora recognition* and *bridging anaphora resolution* (see Kobayashi and Ng (2020) for a comprehensive overview of this area of research). Bridging anaphora recognition has been tackled as part of the information status (IS) classification problem (Rahman and Ng, 2011, 2012; Hou et al., 2013; Hou, 2020b). Recall that the goal of IS classification is to assign an IS to each discourse entity that indicates how these entities are referred to in a text (Prince, 1981; Nissim et al., 2004; Markert et al., 2012): an entity is *old* if it is coreferent with an entity that has been mentioned before, *new* if it is introduced into the discourse for the first time and is not known to the hearer before, and *mediated* if it has not been introduced in the discourse but can be inferred from previously mentioned entities. Bridging anaphors are a type of *mediated* entities that are discourse-new but hearer-old. Bridging anaphora resolution, on the other hand, focuses on selecting antecedents for bridging anaphors (Poesio et al., 2004; Pandit et al., 2020; Hou, 2020a). There are a few works tackling full bridging resolution (i.e., recognizing bridging anaphors and linking them to the antecedents), ranging from rule-based approaches (Hou et al., 2014; Rösiger et al., 2018), to machine learning-based approaches (Hou et al., 2018; Yu and Poesio, 2020) and hybrid methods (Kobayashi and Ng, 2021; Kobayashi et al., 2022). However, these resolvers all assume that gold mentions are given, which hinders the application of bridging resolution in downstream tasks.

In contrast, we focus on end-to-end bridging resolution. Note that some recent attempts have been made in this direction. For example, Hou's (2020a) approach to bridging anaphora resolution does not require gold mentions when constructing the list of antecedent candidates; nevertheless, it still needs gold bridging anaphora information. In addition, while Hou (2021) proposes an end-to-end neural approach to the related tasks of IS classification and bridging anaphora recognition, it has not been extended to bridging resolution. More recently, in the Bridging track of the CODI-CRAC shared task on Anaphora, Bridging, and Discourse Deixis in Dialogue in 2021 (Khosla et al., 2021) and 2022 (Yu et al., 2022), the participants built resolvers for performing end-to-end bridging resolution in dialogue in the "Predicted" phase (Kim et al., 2021; Kobayashi et al., 2021; Li et al., 2022).

3 State-of-the-Art Approaches

Existing approaches to bridging resolution can be broadly divided into rule-based approaches, learning-based approaches, and hybrid approaches.

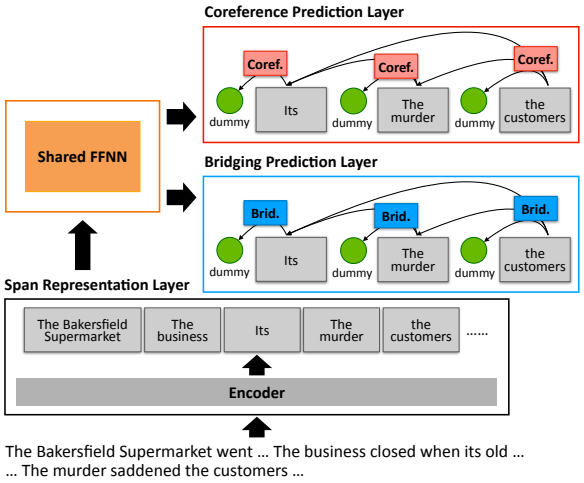

Figure 1: The MTL framework for bridging resolution.

In this section, we overview the state-of-the-art approach in each of these three categories, as we will extend them to the end-to-end setting in Section 4.

3.1 Yu and Poesio's (2020) Model

Yu and Poesio's (Y&P) approach is a state-of-the-art learning-based approach to bridging resolution. Their model is a span-based neural model that takes as input a document D represented as a sequence of word tokens and the associated set of gold mentions, and performs joint bridging resolution and coreference resolution, which we define below, in a multi-task learning (MTL) framework.

The bridging resolution task aims to find a bridging antecedent b_i for each span i in D. The set of possible values for b_i is $\mathcal{B}(i) = \{1, ..., i-1, \epsilon\}$, the preceding spans or a dummy antecedent (if the mention underlying i is not a bridging anaphor). Y&P define the following scoring function:

$$s_b(i,j) = \begin{cases} 0 & j = \epsilon \\ s_a(i,j) & j \neq \epsilon \end{cases} \quad (1)$$

where $s_a(i,j)$ is a pairwise score computed over i and a preceding span j suggesting their likelihood of having a bridging link. The antecedent of i is predicted to be $y_b^* = \arg\max_{y_b \in \mathcal{B}(i)} s_b(i, y_b)$.

The entity coreference task aims to find a coreference antecedent c_i for each span i based on a scoring function s_c that is defined analogously as the s_b function in the bridging resolution task.

Figure 1 illustrates the structure of MTL framework, which we describe in detail below.

Span Representation Layer To encode the tokens and the surrounding contexts of a gold mention, Y&P use a bidirectional LSTM (Hochreiter and Schmidhuber, 1997) that takes as input BERT and GloVe embeddings. They define $\mathbf{g}_i$, the representation of span i, as $[\mathbf{x}_{start(i)}; \mathbf{x}_{end(i)}; \mathbf{x}_{head(i)}; \phi_i]$, where $\mathbf{x}_{start(i)}$ and $\mathbf{x}_{end(i)}$ are the hidden vectors of the start and end tokens of i, $\mathbf{x}_{head(i)}$ is an attention-based head vector and ϕ_i is a span width feature embedding.

Bridging Prediction Layer To predict bridging links, Y&P first calculate the pairwise score between spans i and j as follows:

$$s_a(i,j) = \text{FFNN}_b([\mathbf{g}_i; \mathbf{g}_j; \mathbf{g}_i \circ \mathbf{g}_j; \psi_{ij}]) \quad (2)$$

where $\text{FFNN}_b(\cdot)$ represents a standard feedforward neural network, and $\circ$ denotes element-wise multiplication. This pairwise score includes $\mathbf{g}_i \circ \mathbf{g}_j$, which encodes the similarity of i and j, and ψ_{ij}, which denotes the distance between them.

Coreference Prediction Layer To predict coreference links, Y&P calculate the pairwise score that is defined analogously as in Equation 2 using another FFNN, FFNN_c. The model shares the first few hidden layers of FFNN_b and FFNN_c as well as the span representations.

The loss function is the weighted sum of the losses of the bridging task (L_b) and the coreference task (L_c). L_b and L_c are defined as the negative marginal log-likelihood of all correct bridging antecedents and coreference antecedents, respectively. The weights associated with the losses are tuned using grid search to maximize the average bridging resolution F-scores on development data.

3.2 Rösiger et al.'s (2018) Approach

Rösiger et al.'s approach, which builds upon the rules designed by Hou et al. (2014), is by far the best-performing rule-based approach to bridging resolution. These rules are shown in Appendix A. When evaluating on BASHI, all nine rules are applied, but when evaluating on ISNotes, only the first eight are used. The reason is that the last rule aims to resolve comparative anaphors, which are not annotated in ISNotes. Each rule is composed of an "anaphor" condition and an "antecedent" condition. When two mentions satisfy the two conditions of a rule, they will be extracted as a bridging pair.

3.3 Kobayashi and Ng's (2021) Approach

Motivated by the observation that Rösiger et al.'s rule-based approach and Y&P's MTL approach are complementary rather than competing, Kobayashi

and Ng (K&N) propose a hybrid approach to bridging resolution that combines these two approaches in a pipeline fashion. Given a document, they first use the rules to extract the bridging pairs and then use Y&P's neural model to resolve all and only those mentions that are not resolved by the rules.

From a modeling perspective, however, K&N's approach is not particularly elegant, as there are two models (i.e., the rules and the neural models are still separate). Consequently, we propose a variant of the hybrid approach where we *integrate* the rules into the MTL model. Recall that each rule posits a bridging link when the anaphor and antecedent conditions are both satisfied. To incorporate these predictions into the MTL model, we first define a rule score function $r(i, j)$ whose value is the precision of the rule that posits a bridging link between spans i and j. This rule score function is incorporated into Equation 1 as follows:

$$s_{b'}(i,j) = \begin{cases} 0 & j = \epsilon \\ s_b(i,j) + \alpha r(i,j) & j \neq \epsilon \end{cases} \quad (3)$$

where α is a positive constant that controls the impact of the rule information on s'_b. The smaller α is, the less impact rule information has on s'_b. $s_{b'}$ is then used as the bridging score function when ranking the candidate antecedents of span i. Note that (1) if no rule posits i and j as bridging, $r(i, j)$ is 0; (2) rule precision is computed on the training set; and (3) α is tuned on the development set.

We will henceforth refer to the original K&N approach as H_1 and our proposed variant as H_2.

4 End-to-End Models

Next, we show how to create end-to-end versions of the three approaches described in Section 3.

4.1 Yu and Poesio's (2020) Model

We present two approaches to create end-to-end versions of Y&P's model.

Joint approach. The first approach learns mention boundaries jointly with bridging and coreference resolution. Specifically, following Joshi et al. (2019), for each document, we enumerate all possible intra-sentence spans of up to length L_m. We compute a score s_m for each span i that indicates i's likelihood of being a mention.

$$s_m(i) = \text{FFNN}_m(\mathbf{g}_i) \quad (4)$$

where FFNN_m is a feedforward neural network used to calculate mention scores. Using these scores, the model prunes candidate spans and retains only the top N spans for further processing in order to maintain computational tractability. These scores are then incorporated into the bridging score function in Equation 1 as additional terms:

$$s_b(i,j) = \begin{cases} 0 & j = \epsilon \\ s_m(i) + s_m(j) + s_a(i,j) & j \neq \epsilon \end{cases} \quad (5)$$

We incorporate these scores into the coreference score function $s_c(i, j)$ in a similar manner.

Pipeline approach. In this approach, we do not make any changes to Y&P's model. Rather, during testing, we first apply a mention extractor to extract mentions and then employ Y&P's model from Section 3 to resolve mentions.

Next, we describe our mention extractor. For ISNotes, we use Hou's (2021) mention extraction model, which has achieved state-of-the-art results on ISNotes and outperformed Yu et al.'s (2020) neural mention extractor. For BASHI, we use all the noun phrases extracted from the automatic parse trees that are obtained using Stanford CoreNLP (Manning et al., 2014).[1] The reason is that in BASHI gold mentions are not annotated, and bridging links are annotated over the noun phrases extracted from gold parse trees.[2]

Using BERT and SpanBERT as encoders. We strengthen Y&P's model by replacing its biLSTM encoder with Transformer-based encoders, including BERT and SpanBERT, the latter of which has been successfully applied to entity coreference resolution. To do so, we follow Joshi et al. (2019) and replace the LSTM-based encoder in Y&P's model, which takes frozen BERT and Glove embeddings as input, with BERT or SpanBERT, which corresponds to the "Encoder" component in Figure 1. We adopt the independent version of Joshi et al. (2019), where an input document is split into non-overlapping segments of up to length L_s.

4.2 Rösiger et al.'s (2018) Approach

While the rules were designed to operate on gold mentions, they can be applied to mentions extracted

[1] These mention extractors achieve F-scores of 92.1 (when extracting gold mentions in ISNotes) and 92.0 (when extracting gold noun phrases in BASHI).

[2] In preliminary experiments, we applied Hou's (2021) mention extractor to extract mentions in BASHI, but the results were poorer than those obtained using noun phrases extracted from automatic parse trees.

Corpora	Docs	Tokens	Mentions	Anaphors
ISNotes	50	40,292	10,980	663
BASHI	50	57,709	18,561	452

Table 1: Statistics on ISNotes and BASHI.

using one of the mention extractors described in Section 4.1 (i.e., Hou's (2021) mention extractor for ISNotes and the noun phrases extracted from system parse trees for BASHI) with just one caveat. Specifically, Rösiger et al. use gold annotations (i.e., gold POS tags, gold parse trees, and gold entity types) when computing the information needed by the rules. To make the rules applicable in an end-to-end setting, we use Stanford CoreNLP to provide automatic constituency and dependency parse trees and spaCy (Honnibal and Montani, 2017) to provide automatic POS tags and entity types.

4.3 Kobayashi and Ng's (2021) Approach

Now that we have end-to-end versions of Rösiger et al.'s rule-based approach and Y&P's MTL approach, we can simply use them to create an end-to-end version of K&N's hybrid approach.

5 Evaluation

5.1 Experimental Setup

Corpora. Since we focus on *anaphoric referential bridging resolution*, which corresponds to "referential bridging" in Rösiger et al. (2018) where bridging anaphors are truly anaphoric and bridging relations are context-dependent[3], we employ two widely used English referential bridging corpora: ISNotes (Markert et al., 2012) and BASHI (The Bridging Anaphors Hand-annotated Inventory) (Rösiger, 2018), both of which are composed of different sets of 50 WSJ articles in OntoNotes with anaphoric referential bridging annotations. Table 1 shows statistics on these corpora. We perform 5-fold cross validation (70% for model training, 10% for development, and 20% for testing).

Evaluation setting. We evaluate bridging resolvers in the end-to-end setting, meaning that they extract bridging relations given a raw document.

Evaluation metrics. Bridging results are reported in terms of precision (P), recall (R), and F-score (F) for recognition and resolution. For completeness we also report the results of entity coreference. Entity coreference results are expressed in terms of the CoNLL score (Pradhan et al., 2014), which is the unweighted average of the F-scores provided by three coreference evaluation metrics, MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), and $CEAF_e$ (Luo, 2005).

[3]We excluded ARRAU (Uryupina et al., 2020) from our evaluation because most bridging links in ARRAU are non-anaphoric referential bridging pairs (e.g., *Europe-Spain*), which Rösiger et al. (2018) refer to as *lexical bridging*.

Model	Bridging	
	Recognition	Resolution
	ISNotes	
Rösiger et al. (2018)	25.6	17.5
Our re-implementation	28.1	18.1
	BASHI	
Rösiger et al. (2018)	27.2	14
Our re-implementation	28.5	14.1

Table 2: Comparison of Rösiger et al's (2018) resolver and our re-implementation on ISNotes and BASHI.

Implementation details. For the MTL model, we extend a publicly-available implementation of Y&P's resolver[4] so that it can operate in an end-to-end setting. The BERT and SpanBERT encoders we use are BERT_{LARGE} and SpanBERT_{LARGE}, and BERT_{LARGE} is used to obtain BERT embeddings with segment length 512. We set all parameters to the ones reported in Yu and Poesio (2020) except (1) the task learning rate, which is searched out of $\{1 \times 10^{-4}, 2 \times 10^{-4}, 3 \times 10^{-4}, 4 \times 10^{-4}\}$ and is decayed linearly; and (2) the learning rates for BERT and SpanBERT, which are searched out of $\{1 \times 10^{-5}, 2 \times 10^{-5}, 3 \times 10^{-5}\}$ and are decayed linearly. Each document is split into segments of length 384. We generate all spans of length up to 15 and prune these candidate spans by retaining the top 30%. For training, we use document-sized mini-batches and train models for up to 1600 epochs for both ISNotes and BASHI. α, the weight parameter associated with the rule score, is searched out of {0.1, 0.5, 1.0, 10, 100, 200, 300}.

For the rule resolver, while conceptually we can extend Rösiger et al.'s resolver so that it can operate in an end-to-end setting, the way they structured their code has made it non-trivial to do so. Consequently, we (1) re-implement their resolver, which operates on gold mentions; (2) extend it so that it operates on automatically extracted mentions, as described in Section 4.2; and (3) report rule-based results using this duplicated resolver. As we can see in Table 2, our re-implementation outperforms the original resolver in terms of recognition and resolution F-scores when evaluated on both ISNotes and BASHI in the gold-mention setting.

[4]https://github.com/juntaoy/dali-bridging

	Model	Bridging						Coref.	Bridging						Coref.
		Recognition			Resolution			Res.	Recognition			Resolution			Res.
		P	R	F	P	R	F	CoNLL	P	R	F	P	R	F	CoNLL
		ISNotes							**BASHI**						
	MTL														
1	Rules	49.4	17.4	25.7	31.8	11.2	16.5	-	33.1	22.5	26.8	15.2	10.3	12.3	-
2	LSTM^J	49.6	19.1	29.3	26.0	9.7	14	59.1	40.5	9.3	15.1	19.8	4.5	7.3	58.1
3	LSTM^P	50.0	24.0	32.4	25.4	12.2	16.5	61.1	37.3	16.0	22.4	16.3	7.0	9.8	48.6
4	H_1(LSTM^P)	45.8	32.4	**37.9**	25.9	18.3	21.5	61.1	33.6	32.3	**33.0**	16.4	15.8	**16.1**	48.6
5	H_2(LSTM^P)	53.3	26.5	35.4	32.6	16.2	**21.6**	60.6	41.2	22.8	29.3	22.6	12.5	**16.1**	47.1
6	BERT^J	72.1	7.8	14.1	45.7	5.0	9.0	61.9	68.7	3.1	5.9	53.6	2.4	4.6	60.3
7	BERT^P	33.3	32.0	32.5	19.0	18.2	18.5	56.9	37.1	23.4	28.7	15.0	9.4	11.6	43.8
8	H_1(BERT^P)	33.8	39.5	**36.4**	20.1	23.5	21.7	56.9	32.4	36.5	**34.3**	15.1	17.0	16.0	43.8
9	H_2(BERT^P)	33.7	32.7	33.2	22.4	21.8	**22.1**	55.1	45.8	23.7	31.2	23.9	12.4	**16.3**	48.2
10	SBERT^J	79.0	15.6	26.1	52.7	10.4	17.4	68.8	64.5	9.0	15.8	40.1	5.6	9.8	66.5
11	SBERT^P	34.4	30.9	32.6	22.3	20.1	21.1	59.5	34.7	29.4	31.8	15.3	12.9	14.0	47.5
12	H_1(SBERT^P)	35.1	38.8	**36.8**	22.2	24.6	23.4	59.5	31.3	41.6	**35.7**	14.8	19.6	16.9	47.5
13	H_2(SBERT^P)	39.7	31.6	35.1	27.0	21.5	**23.9**	59.2	36.0	27.5	31.2	19.7	15.0	**17.0**	45.4
	STL														
14	LSTM^J	57.4	11.6	19.3	20.4	4.2	6.9	-	59.5	8.6	14.9	15.9	2.3	4.0	-
15	LSTM^P	54.5	15.4	24.0	22.7	6.3	9.9	-	41.7	11.3	17.7	14.6	4.0	6.2	-
16	BERT^J	66.5	4.9	9.1	47.6	3.5	6.5	-	86.7	2.3	4.5	63.3	1.6	3.2	-
17	BERT^P	24.5	36.5	29.2	12.5	18.6	14.9	-	30.8	19.0	23.5	8.9	5.5	6.8	-
18	SBERT^J	67.9	16.7	26.6	38.2	9.1	14.6	-	61.7	5.9	10.8	36.0	3.4	6.2	-
19	SBERT^P	35.2	32.9	33.9	17.9	16.7	17.3	-	26.5	26.8	26.6	11.0	11.1	11.0	-

Table 3: Results of different MTL and STL resolvers in the end-to-end setting. Each result is the average of two runs. The highest recognition and resolution F-scores are bolded for each encoder in MTL.

Finally, for the hybrid approach, we employ our extension of the publicly-available implementation of Y&P's resolver and our re-implementation of Rösiger et al.'s resolver, as described above.

5.2 Results and Discussion

Baselines. Strictly speaking, there are no baselines, as no one has reported results on ISNotes and BASHI in the end-to-end setting. Nevertheless, we will use the three approaches described in Section 4, namely our end-to-end versions of the state-of-the-art approaches to bridging resolution in the gold mention setting, as our baselines.

Results of bridging recognition and resolution in the end-to-end setting for both ISNotes and BASHI are shown in Table 3. For bookkeeping purposes, we also report results on entity coreference resolution. Rules (row 1) corresponds to our duplication of Rösiger et al.'s rule-based approach (Section 4.2). LSTM^J (row 2) and LSTM^P (row 3) are the joint and pipeline versions of Y&P's MTL approach using LSTM as the encoder (Section 4.1). H_1(LSTM^P) (row 4) and H_2(LSTM^P) (row 5) are the original version and our proposed variant of K&N's hybrid approach, respectively.

Several points deserve mention. First, LSTM^J underperforms LSTM^P. This is somewhat unexpected since pipeline models are prone to error propagation and have been shown to underperform their joint counterparts in many NLP tasks. A closer examination of the output reveals the reason: since the mention extraction F-scores on both datasets are above 90%, error propagation is by no means serious. In contrast, LSTM^J has particularly poor anaphor recognition recall, which translates to poor resolution recall and F-score. Second, LSTM^P does not perform better than Rules. While LSTM^P achieves considerably higher recognition recall than Rules, it does not perform better than Rules w.r.t. resolution. For BASHI, LSTM^P underperforms Rules w.r.t. both recognition and resolution. H_1(LSTM^P) outperforms both Rules and LSTM^P. This is perhaps not surprising: this hybrid variant has achieved the best results on ISNotes and BASHI in the gold mention setting and represents the prior state of the art. Our results suggest that the success of this hybrid variant can be extended to the end-to-end setting. Like in the gold setting, recognition and resolution F-scores are both better than other baselines in the end-to-end setting. Finally, consider H_2(LSTM^P). While H_2(LSTM^P) achieves lower recognition F-scores than H_1(LSTM^P), the resolution F-scores achieved by the two models are comparable, with H_2(LSTM^P) having higher resolution precision and lower resolution recall than H_1(LSTM^P). This suggests that H_2 could be more valuable than H_1 for downstream applications, as these applications

	Model	Bridging						Coref.	Bridging						Coref.
		Recognition			Resolution			Res.	Recognition			Resolution			Res.
		P	R	F	P	R	F	CoNLL	P	R	F	P	R	F	CoNLL
		ISNotes							**BASHI**						
	MTL														
1	Rules	52.7	19.2	28.1	34.0	12.4	18.1	-	35.8	23.6	28.5	17.8	11.7	14.1	-
2	LSTM	53.0	25.6	34.5	24.2	11.7	17.2	64.0	38.2	16.5	23.0	17.3	7.5	10.4	57.4
3	H_1(LSTM)	48.1	34.6	**40.3**	27.2	19.6	22.8	64.0	34.7	33.7	**34.2**	18.1	17.5	**17.8**	57.4
4	H_2(LSTM)	56.5	28.8	38.1	34.0	17.3	**22.9**	63.8	45.4	22.8	30.4	26.6	13.4	**17.8**	57.4
5	BERT	34.9	33.2	33.9	20.3	19.3	19.7	60.2	37.6	23.4	28.8	15.2	9.4	11.6	52.1
6	H_1(BERT)	35.6	42.2	**38.6**	21.1	25.0	22.8	60.2	33.6	37.8	**35.6**	16.4	18.5	17.4	52.1
7	H_2(BERT)	36.3	35.7	36.0	23.7	23.3	**23.5**	58.3	46.3	24.8	32.3	25.8	13.8	**18.0**	50.7
8	SBERT	37.1	33.1	35.0	24.5	21.9	23.1	62.9	35.0	29.7	32.1	16.1	13.7	14.8	54.9
9	H_1(SBERT)	37.6	42.4	**39.8**	23.6	26.6	25.0	59.5	32.2	43.0	**36.8**	16.3	21.7	18.6	54.9
10	H_2(SBERT)	43.8	34.6	38.6	30.4	24.1	**26.8**	62.6	37.6	28.8	32.6	21.6	16.6	**18.7**	55.1
	STL														
11	LSTM	57.4	16.4	25.4	25.1	11.0	15.2	-	42.8	11.5	18.1	15.5	4.2	6.5	-
12	BERT	26.0	38.2	30.8	13.1	19.2	15.5	-	28.9	20.5	24.0	9.1	6.5	7.6	-
13	SBERT	37.7	34.4	35.9	19.9	18.1	18.9	-	27.8	26.8	27.3	12.2	11.7	12.0	-

Table 4: Results of different MTL and STL resolvers in the gold mention setting. Each result is the average of two runs. The highest recognition and resolution F-scores are bolded for each encoder in MTL.

typically cannot benefit from bridging information if many links are erroneous.[5]

LSTM vs. BERT/SpanBERT. Results for BERT and SpanBERT (SBERT) are shown in rows 6–9 and rows 10–13 respectively. Comparing the BERT results with the corresponding LSTM results, we see that the two achieve comparable F-scores w.r.t. resolution except for three cases (BERT^P outperforms LSTM^P on both datasets, and H_2(BERT^P) outperforms H_2(LSTM^P) on ISNotes). In terms of recognition, the results are mixed: on BASHI the BERT models outperform their LSTM counterparts, while the reverse is true on ISNotes. Next, consider the SBERT results. Each SBERT model considerably outperforms the corresponding BERT and LSTM models. Generally, the higher resolution F-scores achieved by SBERT can be attributed to its higher recall on BASHI and its higher precision on ISNotes. These results show the usefulness of SBERT for bridging resolution.

MTL vs. STL. Y&P show that multi-task learning for entity coreference and bridging outperforms single-task learning (i.e., learning bridging resolution without coreference). The question is: would MTL still outperform STL in the end-to-end setting? To answer this question, we obtain STL results by removing the coreference prediction layer in Y&P's model and retraining it. Results of this experiment using different encoders in the Y&P model are shown in rows 11–13. As can be seen, regardless of which encoder is used, the resolution F-scores achieved by STL are lower than those achieved by MTL for both datasets.

End-to-end vs. gold settings. Will the trends we have observed so far generalize to the gold mention setting? To answer this question, we repeat the experiments in Table 3 on gold mentions. There is a caveat involved in evaluating on gold mentions, however. In ISNotes and BASHI, some bridging anaphors have clausal antecedents that correspond to *events*. While clausal antecedents are annotated, they are *not* annotated as gold mentions, and previous studies differ in terms of how they should be handled. Specifically, some previous work (e.g., Hou et al. (2014), Hou et al. (2018)) chose not to include these clausal antecedents in the list of candidate antecedents and others (e.g., Rösiger et al. (2018), Yu and Poesio (2020)) did. Obviously, the setting in which gold clausal antecedents are not included in training/evaluation is harsher because it implies that anaphors with clausal antecedents will always be resolved incorrectly. We believe that including gold clausal antecedents during evaluation does not represent a realistic setting, and will therefore report results using the "harsh" setting when evaluating on gold mentions.

Results of the gold mention setting are shown in Table 4.[6] Recall that the distinction between joint

[5]We use the pipeline version rather than the joint version of Y&P's MTL model in the hybrid variants because of our desire to create stronger baselines.

[6]The baseline results in Table 4 are lower than those reported in the original papers because (1) we report results using the "harsh" setting; (2) Rösiger et al. (2018) and Kobayashi and Ng (2021) postprocess the system output with gold coreference information, and (3) Yu and Poesio (2020) and Kobayashi and Ng (2021) use additional labeled data for model training.

Perturbation Type	Example
Seen adj/adv	strategically → slightly
Nonexistent adj/adv	skeptical → lacitpeks
Seen verbs	start → reply
Nonexistent verbs	possess → ssessop
Seen nouns	honesty → wall
Nonexistent nouns	example → elpmaxe
Seen words	particularly → firmly
Nonexistent words	begun → nugeb

Table 5: Perturbation examples.

and pipeline approaches is no longer applicable in the gold mention setting. As can be seen, the conclusions we drew based on the end-to-end results are also applicable to the gold mention results.

One of the questions we aim to answer is: how much worse would the end-to-end results be compared to the corresponding gold mention results? We see that the end-to-end LSTM-based (LSTM^P, H_1(LSTM^P), H_2(LSTM^P)) and BERT-based (BERT^P, H_1(BERT^P), H_2(BERT^P)) resolvers underperform their counterparts in the gold setting by up to 2.8% F-score in recognition and up to 1.7% F-score in resolution. This performance gap widens with SBERT (SBERT^P, H_1(SBERT^P), H_1(SBERT^P)), having a difference of up to 3.5% F-score in recognition and up to 2.9% F-score in resolution. Overall, while the gold results are better than the corresponding end-to-end results, the difference between them is less than 1.8% for LSTM and BERT and less than 3.0% for SBERT. These results are encouraging considering that the end-to-end evaluation setting is very challenging.

H_1 vs. H_2. While the performance difference between H_1 and H_2 tends to be small w.r.t. resolution in the end-to-end setting, there are cases in the gold mention setting in which this difference in resolution F-score is comparatively larger. Specifically, when BERT is used, H_2 outperforms H_1 by 0.6–0.7% points in F-score on the two corpora, and when SBERT is used, H_2 outperforms H_1 by 1.8% points in F-score on ISNotes.

5.3 Sensitivity to Perturbed Inputs

Next, we conduct experiments that involve perturbing the input. For each experiment, we replace a certain type of words with other words in all training documents, retrain our best-performing model, H_2(SBERT^P), on these perturbed training documents, and evaluate it on the (unperturbed) test set. The goal is to gain insights into the behavior of the best model by assessing how sensitive its performance is when training inputs are perturbed.

	Pertutbation Type	ISNotes		BASHI	
		Rec.	Res.	Rec.	Res.
1	No perturbation	**35.1**	**23.9**	**31.2**	**17.0**
2	Seen adj/adv	33.0	22.5	29.0	12.7
3	Nonexistent adj/adv	34.9	23.4	28.8	12.9
4	Seen verbs	35.0	23.2	28.9	13.3
5	Nonexistent verbs	35.0	23.5	30.4	13.4
6	Seen nouns	32.9	22.1	30.1	12.5
7	Nonexistent nouns	31.8	22.2	29.6	12.7
8	Seen words	33.2	20.6	30.4	12.2
9	Nonexistent words	32.3	21.3	25.4	12.7

Table 6: Perturbation results of the best model.

If performance drops a lot when a certain type of word is replaced, then it means that that type of words is important in the learning process. Note that we consider only mention-external perturbations, meaning that we only replace words that are not part of a bridging anaphor or its antecedent(s).

Specifically, we replace words from the following categories: adjectives and adverbs only, verbs only, nouns only, and all categories combined. For each category, we consider two replacement methods. One is to replace each word with another word of the same POS tag that is taken from the training documents but which has never appeared within a mention in the training set (**Seen**). This replacement is deterministic: all occurrences of a given word will be replaced with the same word. The other method involves replacing each word with a nonexistent word (**Nonexistent**), which we create by reversing the order of the characters of the word to be replaced. This latter method tests the impact of nonexistent words has on the model.

Results of these experiments are reported in Table 6 in terms of recognition and resolution F-scores. To facilitate comparison, we show in row 1 the results of the resolver when the input is not perturbed. Several points deserve mention. First, all results obtained via perturbations are lower than the "No perturbation" results in row 1. This implies that each kind of perturbation we considered affects the model learning process and negatively impacts bridging recognition and resolution performances. Second, Seen words appear to confuse the model more than Nonexistent words. This is perhaps not surprising: in the Nonexistent setting the model will not be confused by those Seen replacements that could cause a sentence to become unsensible. Finally, we see from rows 4 and 5 that verbs have the least impact on resolution F-scores, suggesting that adj/adv and nouns play more important roles than verbs in learning span-based models for

bridging resolution.

5.4 Analysis of Results

Error analysis of the best end-to-end model. To gain additional insights into our best end-to-end model (H_2(SBERT^P)), we conduct an error analysis of this resolver. First, the system is still struggling to detect the majority of the bridging anaphors and find their antecedents, having recall scores of 31.6% and 27.5% for bridging anaphora recognition on ISNotes and BASHI, respectively. Only a very small portion of the recall errors are from mention prediction errors: 3% and 1.3% of the gold bridging anaphors are misclassified as non-mentions in ISNotes and BASHI, respectively. The system makes more recall errors at predicting definite bridging anaphors (i.e., NPs modified by the definite article "the") than other bridging anaphors. For instance, on ISNotes, the recall scores of identifying definite bridging anaphors and other bridging anaphors are 20% and 25%, respectively.

Next we analyze the precision errors on ISNotes because BASHI does not annotate mentions and their information status. We find that mention prediction errors (i.e., predicted bridging anaphors are not mentions) account for 8.7% of the precision errors for bridging anaphora recognition. In addition, 16.7% of the wrongly predicted bridging pairs contain correct bridging anaphors but wrong antecedents. The majority of the precision errors can be attributed to the fact that the system predicts *new* and *old* mentions as bridging anaphors, which account for 31% and 21% of the precision errors, respectively. This is in line with the previous studies on bridging recognition that suggest that systems often fail to distinguish bridging anaphors from generic *new* mentions with simple syntactic structures (Hou et al., 2018; Hou, 2021).

Comparison of different encoders and embeddings. We analyze the results from three end-to-end systems: H_2(LSTM^P), H_2(BERT^P), and H_2(SBERT^P). which correspond to rows 5, 9, and 13 in Table 3, respectively. As noted before, the LSTM encoder with BERT embeddings (i.e., H_2(LSTM^P)) is more conservative in link prediction, having higher precision but lower recall than the other two systems. In fact, on ISNotes, H_2(LSTM^P) only predicts half of the number of bridging pairs predicted by the other two systems. Interestingly, although both H_2(BERT^P) and H_2(SBERT^P) achieve higher recall scores on full bridging resolution compared to H_2(LSTM^P), they both make a relatively large portion of precision errors that involve linking a *mediated/syntactic* mention m to a previous mention that is often related to the premodification of m, such as {*Britain's voters – Britains's*} or {*Some Mobil executives – Mobil Corp.*}. On the contrary, this kind of error is rare in H_2(LSTM^P).

Finally, we analyze the recall scores based on the determiners of bridging anaphors. We divide bridging anaphors into three categories: (1) *the NPs* correspond to bridging anaphors that are modified by the definite article "the"; (2) *other determiner NPs* contain bridging anaphors that are modified by the indefinite articles "a/an" as well as other determiners (e.g., demonstratives or possessives); and (3) *bare NPs* are bridging anaphors that are not modified by any determiners, such as "subsidies" and "overseas operations". The majority of the correctly predicted bridging links from the above three models are *bare NPs*. H_2(BERT^P) is better at predicting bridging anaphors for all three categories compared to H_2(LSTM^P). The performance of H_2(SBERT^P) on *other determiner NPs* and *bare NPs* is on par with that of H_2(BERT^P), but the former achieves higher recall at identifying definite bridging anaphors than the latter (i.e., 20% vs. 14% on ISNotes).

6 Conclusion

We conducted a pioneering study on end-to-end neural bridging resolution in which we adapted three state-of-the-art bridging resolvers that were originally developed to operate on gold mentions, namely Rösiger et al.'s (2018) resolver, Yu and Poesio's (2020) resolver, and Kobayashi and Ng's (2021) resolver, to the end-to-end setting. To strengthen the resolvers, we replaced the LSTM encoders they use with BERT- and SpanBERT-based encoders. In an evaluation on ISNotes and BASHI, end-to-end bridging resolvers lagged behind their gold-mention counterparts by only 2-3% absolute F-score. These results suggested that time is ripe for researchers to focus on evaluating bridging resolvers in the end-to-end setting. In addition, our work suggested that H_2(SBERT^P), the hybrid score-based pipeline bridging resolver trained using SBERT (1) achieves better performance than other model variants; (2) is sensitive to all kinds of perturbations we considered; and (3) will likely be improved by improving mediated/syntactic errors.

Acknowledgments

We thank the three anonymous reviewers for their insightful comments on an earlier draft of the paper. This work was supported in part by NSF Grants IIS-1528037 and CCF-1848608. Any opinions, findings, conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views or official policies, either expressed or implied, of the NSF.

References

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *Proceedings of the Linguistic Coreference Workshop at the First International Conference on Language Resources and Evaluation (LREC'98)*, pages 563–566, Granada, Spain. European Language Resources Association (ELRA).

Herbert H. Clark. 1975. Bridging. In *Proceedings of the 1975 Workshop on Theoretical Issues in Natural Language Processing*, TINLAP '75, page 169–174, USA. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Yufang Hou. 2020a. 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.

Yufang Hou. 2020b. Fine-grained information status classification using discourse context-aware BERT. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6101–6112, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yufang Hou. 2021. End-to-end neural information status classification. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1377–1388, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yufang Hou, Katja Markert, and Michael Strube. 2013. Cascading collective classification for bridging anaphora recognition using a rich linguistic feature set. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 814–820, Seattle, Washington, USA. Association for Computational Linguistics.

Yufang Hou, Katja Markert, and Michael Strube. 2014. A rule-based system for unrestricted bridging resolution: Recognizing bridging anaphora and finding links to antecedents. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2082–2093, Doha, Qatar. Association for Computational Linguistics.

Yufang Hou, Katja Markert, and Michael Strube. 2018. Unrestricted bridging resolution. *Computational Linguistics*, 44(2):237–284.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. 2019. BERT for coreference resolution: Baselines and analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5803–5808, Hong Kong, China. Association for Computational Linguistics.

Sopan Khosla, Juntao Yu, Ramesh Manuvinakurike, Vincent Ng, Massimo Poesio, Michael Strube, and Carolyn Rosé. 2021. The CODI-CRAC 2021 shared task on anaphora, bridging, and discourse deixis in dialogue. In *Proceedings of the CODI-CRAC 2021 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue*, pages 1–15, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hongjin Kim, Damrin Kim, and Harksoo Kim. 2021. The pipeline model for resolution of anaphoric reference and resolution of entity reference. In *Proceedings of the CODI-CRAC 2021 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue*, pages 43–47, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hideo Kobayashi, Yufang Hou, and Vincent Ng. 2022. Constrained multi-task learning for bridging resolution. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 759–770, Dublin, Ireland. Association for Computational Linguistics.

Hideo Kobayashi, Shengjie Li, and Vincent Ng. 2021. Neural anaphora resolution in dialogue. In *Proceedings of the CODI-CRAC 2021 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue*, pages 16–31, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Hideo Kobayashi and Vincent Ng. 2021. Bridging resolution: Making sense of the state of the art. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1652–1659, Online. Association for Computational Linguistics.

Shengjie Li, Hideo Kobayashi, and Vincent Ng. 2022. Neural anaphora resolution in dialogue revisited. In *Proceedings of the CODI-CRAC 2022 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue*, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

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.

Shachar Mirkin, Ido Dagan, and Sebastian Padó. 2010. Assessing the role of discourse references in entailment inference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1209–1219, Uppsala, Sweden. Association for Computational Linguistics.

Malvina Nissim, Shipra Dingare, Jean Carletta, and Mark Steedman. 2004. An annotation scheme for information status in dialogue. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Onkar Pandit, Pascal Denis, and Liva Ralaivola. 2020. Integrating knowledge graph embeddings to improve mention representation for bridging anaphora resolution. In *Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 55–67, Barcelona, Spain (online). Association for Computational Linguistics.

Massimo Poesio, Rahul Mehta, Axel Maroudas, and Janet Hitzeman. 2004. Learning to resolve bridging references. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 143–150, Barcelona, Spain.

Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. 2014. Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 30–35, Baltimore, Maryland. Association for Computational Linguistics.

Ellen F. Prince. 1981. Toward a taxonomy of given-new information. In P. Cole, editor, *Syntax and semantics: Vol. 14. Radical Pragmatics*, pages 223–255. Academic Press, New York.

Altaf Rahman and Vincent Ng. 2011. Learning the information status of noun phrases in spoken dialogues. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1069–1080, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Altaf Rahman and Vincent Ng. 2012. Learning the fine-grained information status of discourse entities. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 798–807, Avignon, France. Association for Computational Linguistics.

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).

Ina Rösiger, Arndt Riester, and Jonas Kuhn. 2018. Bridging resolution: Task definition, corpus resources and rule-based experiments. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3516–3528, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Bo-Hsiang Tseng, Shruti Bhargava, Jiarui Lu, Joel Ruben Antony Moniz, Dhivya Piraviperumal, Lin Li, and Hong Yu. 2021. CREAD: Combined resolution of ellipses and anaphora in dialogues. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3390–3406, Online. Association for Computational Linguistics.

Olga Uryupina, Ron Artstein, Antonella Bristot, Federica Cavicchio, Francesca Delogu, Kepa Joseba Rodriguez, and Massimo Poesio. 2020. Annotating a broad range of anaphoric phenomena, in a variety of genres: the ARRAU corpus. *Natural Language Engineering*, 26(1):95–128.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*.

Juntao Yu, Bernd Bohnet, and Massimo Poesio. 2020. Neural mention detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1–10, Marseille, France. European Language Resources Association.

Juntao Yu, Sopan Khosla, Ramesh Manuvinakurike, Lori Levin, Vincent Ng, Massimo Poesio, Michael Strube, and Carolyn Rosé. 2022. The CODI-CRAC 2022 shared task on anaphora, bridging, and discourse deixis in dialogue. In *Proceedings of the CODI-CRAC 2022 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue*, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Juntao Yu and Massimo Poesio. 2020. Multitask learning-based neural bridging reference resolution. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3534–3546, Barcelona, Spain (Online). International Committee on Computational Linguistics.

A Rules

Rösiger et al. (2018) designed rules to resolve the bridging anaphors in ISNotes and BASHI. Table 7 shows these rules. For each rule, we describe the anaphor and antecedent conditions as well as its motivation and its resolution precision. The precision scores are calculated using our re-implementation or Rösiger et al.'s resolver. Note that the first nine rules are used for resolution in both ISNotes and BASHI while the last rule (comparative anaphora) is specifically designed for BASHI.

Rule	Description (anaphor)	Description (antecedent)	Motivation	Res. Precision (%)
Set: Percentage	Percentage NPs in subject position	Closest NP modifying another percentage NP via the preposition “of” (e.g. 22% of the firms)	Percentage expressions can indicate set bridging	I: 100.0 B: 0.0
Building part	Common NPs whose head is a building part without nominal pre-modifications	NP with the strongest semantic connectivity to anaphor	A typical case (building part) of meronym bridging	I: 62.5 B: 0.0
Set: Number or indefinite pronoun	Number expressions (e.g. two dogs) or indefinite pronouns (e.g. some ...)	Closest plural NP in subject position. If not found, closest plural NP in object position	Numbers or indefinite pronouns can indicate set bridging	I: 80.0 B: 33.3
Argument-taking NPs 1	NPs with high argument ratio and without nominal/adjective pre-modifications or indefinite determiners	1. take all nominal modifiers of NPs whose head is same as anaphor’s head. 2. closest NP that is a realization of these modification	Different instances of the same noun predicate likely maintain the same argument fillers indicated by nominal modifiers (extended claim from Laparra’13)	I: 40.0 B: 18.5
Relative person	Non-generic NPs whose head is a relative without no nominal/adjective pre-modifications	Closest non-relative person NP	Handles relative nouns, which tend to be bridging	I: 50.0 B: 42.9
GPE job title	Job titles with country pre-modifications (e.g., Italian mayor)	Most salient GPE (e.g., Italy)	Some job title NPs implicitly refer to the globally salient GPE	I: 45.0 B: 14.3
Professional role	Professional role NPs (e.g. professor)	Most salient organization name	A more general rule than "Relative person" and "GPE job title"	I: 62.0 B: 21.2
Argument-taking NPs 2	NPs in subject position with high argument ratio and without nominal/adjective pre-modifications	NP with the strongest semantic connectivity to the anaphor	An NP in subject position that is likely to take arguments tends to be bridging anaphor	I: 28.1 B: 0.0
Meronym relation	Unmodified definite NPs	NP classified as meronym with anaphor according to a relation classifier trained using WordNet	Handles meronym bridging	I: 11.8 B: 11.3
Comparative anaphora	NPs with comparative markers	Closest NP with same head and semantic category	Comparative anaphors are typically indicated by certain markers	B: 45.5

Table 7: Rules used by Rösiger et al. (2018) for resolving bridging anaphors in ISNotes and BASHI. ’I’ and ’B’ refer to ISNotes and BASHI, respectively.