Bridging Anaphora Resolution as Question Answering

Yufang Hou
IBM Research Europe
yhou@ie.ibm.com

Abstract

Most previous studies on bridging anaphora resolution (Poesio et al., 2004; Hou et al., 2013b; Hou, 2018a) use the pairwise model to tackle the problem and assume that the gold mention information is given. In this paper, we cast bridging anaphora resolution as question answering based on context. This allows us to find the antecedent for a given anaphor without knowing any gold mention information (except the anaphor itself). We present a question answering framework (*BARQA*) for this task, which leverages the power of transfer learning. Furthermore, we propose a novel method to generate a large amount of "quasi-bridging" training data. We show that our model pre-trained on this dataset and fine-tuned on a small amount of in-domain dataset achieves new state-of-the-art results for bridging anaphora resolution on two bridging corpora (ISNotes (Markert et al., 2012) and BASHI (Rösiger, 2018)).

1 Introduction

Anaphora accounts for text cohesion and is crucial for text understanding. An anaphor is a noun phrase (NP) that usually refers back to the same or a different entity (the antecedent) in text. *Anaphora resolution* is the task to determine the antecedent for a given anaphor. While *direct anaphora resolution* attracts a lot of attention in the NLP community recently, such as Winograd Schema Challenge (Rahman and Ng, 2012; Opitz and Frank, 2018; Kocijan et al., 2019), *indirect anaphora resolution* or *bridging anaphora resolution* is less well studied.

In this paper, we focus on *bridging anaphora resolution* where bridging anaphors and their antecedents are linked via various lexico-semantic, frame or encyclopedic relations. Following Hou et al. (2013b) and Rösiger et al. (2018), we mainly consider "referential bridging" in which bridging anaphors are truly anaphoric and bridging relations are context-dependent. In Example 1[1], both "her building" and "*buildings with substantial damage*" are plausible antecedent candidates for the bridging anaphor "**residents**" based on lexical semantics. In order to find the antecedent (*buildings with substantial damage*), we have to take the meaning of the broader discourse context into account.

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

Most previous studies on bridging anaphora resolution (Poesio et al., 2004; Lassalle and Denis, 2011; Hou et al., 2013b; Hou, 2018a) tackle the problem using the pairwise model and assume that the gold mention information is given. Most work (Poesio et al., 2004; Lassalle and Denis, 2011; Hou et al., 2013b) uses syntactic patterns to measure semantic relatedness between the head nouns of an anaphor and its antecedent. Hou (2018a) proposes a simple deterministic algorithm that also considers the semantics of modifications for head nouns. These approaches, however, do not take the broader context outside of noun phrases (i.e., anaphors and antecedent candidates) into account and often fail to resolve context-dependent bridging anaphors as demonstrated in Example 1.

Resolving bridging anaphors requires context-dependent text understanding. Recently, Gardner et al. (2019) argue that question answering (QA) is a natural format to model tasks that require question understanding. In this paper, we cast bridging anaphora resolution as question answering based

[1] All examples, if not specified otherwise, are from ISNotes (Markert et al., 2012). Bridging anaphors are typed in boldface, antecedents in italics throughout this paper.

on context. We develop a QA system (*BARQA*) for the task based on BERT (Devlin et al., 2019). Given a context as shown in Example 1, we first rephrase every anaphor as a question, such as "***residents** of what?*". By answering the question, the system then identifies the span of the antecedent from the context. Compared to the pairwise model, our QA system does not require the gold or system mention information as the antecedent candidates. In addition, this framework allows us to integrate context outside of NPs when choosing antecedents for bridging anaphors. For instance, "Green" and "damage were color-coded" are among the top predicted answers for the above question.

Different from coreference resolution, there are no large-scale corpora available for referential bridging resolution due to its complexity. In this paper we propose a new method to generate a large amount of quasi-bridging training data from the automatically parsed Gigaword corpus (Parker et al., 2011; Napoles et al., 2012). We demonstrate that our quasi-bridging training data is a better pre-training choice for bridging anaphora resolution compared to the SQuAD corpus (Rajpurkar et al., 2016). Moreover, we show that our model pre-trained on this dataset and fine-tuned on a small amount of in-domain dataset achieves new state-of-the-art results for bridging anaphora resolution on two bridging corpora (i.e., ISNotes (Markert et al., 2012) and BASHI (Rösiger, 2018)).

To summarize, the main contributions of our work are: (1) we formalize bridging anaphora resolution as a question answering problem and propose a QA model to solve the task; (2) we explore a new method to generate a large amount of quasi-bridging training dataset and demonstrate its value for bridging anaphora resolution; and (3) we carefully carry out a series of experiments on two referential bridging corpora and provide some error analysis to verify the effectiveness of our QA model to resolve the context-dependent bridging anaphors in ISNotes. We release the code and all experimental datasets at `https://github.com/IBM/bridging-resolution`.

2 Related Work

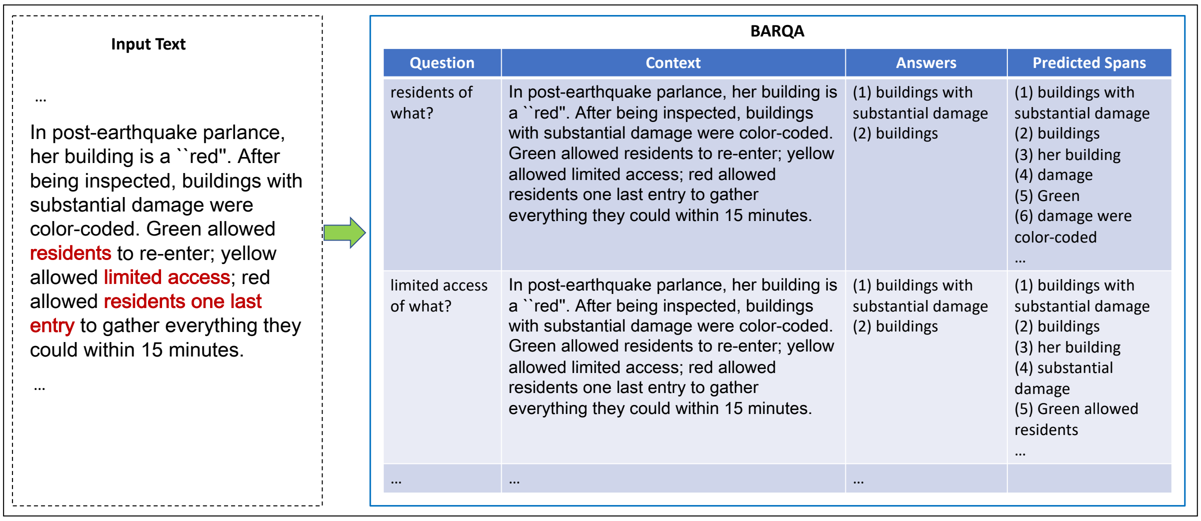
Bridging Anaphora Resolution. Since the '90s, the empirical corpus studies related to bridging have been carried out on various genres and different languages (Fraurud, 1990; Poesio and Vieira, 1998; Poesio, 2004; Nissim et al., 2004; Gardent and Manuélian, 2005; Nedoluzhko et al., 2009; Eckart et al., 2012; Markert et al., 2012; Rösiger, 2018; Poesio et al., 2018). Among those datasets, ISNotes (Markert et al., 2012), BASHI (Rösiger, 2018) and ARRAU (Poesio et al., 2018) are recent three public English corpora which contain medium- to large-sized bridging annotations and have been used to evaluate systems' performance on bridging anaphora recognition (Hou et al., 2013a; Hou, 2016; Rösiger et al., 2018), bridging anaphora resolution (Poesio et al., 2004; Lassalle and Denis, 2011; Hou et al., 2013b; Hou, 2018a), as well as full bridging resolution (Hou et al., 2014, 2018; Rösiger et al., 2018). In this paper, we focus exclusively on the task of antecedent selection.

It is worth noting that the bridging definition in the ARRAU corpus is different from the one used in the other two datasets. Rösiger et al. (2018) pointed out that ISNotes and BASHI contain "referential bridging" where bridging anaphors are truly anaphoric and bridging relations are context-dependent, while in ARRAU, most bridging links are purely lexical bridging pairs which are not context-dependent (e.g., *Europe* – **Spain** or *Tokyo* – **Japan**). In this paper, we focus on resolving referential bridging anaphors.

Regarding the algorithm for bridging anaphora resolution, most previous work uses the pairwise model for the task. The model assumes gold or system mention information (NPs) is given beforehand. It creates (positive/negative) training instances by pairing every anaphor a with its preceding mention m. Usually, m is from a set of antecedent candidates which is formed using a fixed window size. Poesio et al. (2004) and Lassalle and Denis (2011) trained such pairwise models to resolve mereological bridging anaphors in the English GNOME corpus[2] and the French DEDE corpus (Gardent and Manuélian, 2005), respectively. One exception is Hou et al. (2013b), which proposed a joint inference framework to resolve bridging anaphors in ISNotes. The framework is built upon the pairwise model and predicts all semantically related bridging anaphors in one document together.

Recently, Hou (2018a) generated a word representation resource for bridging (i.e., *embeddings_bridging*) and proposed a simple deterministic algorithm to find antecedents for bridging anaphors in ISNotes and BASHI. The word representation resource is learned from a large corpus

[2] The GNOME corpus is not publicly available.



Input Text

...

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

...

BARQA

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

Figure 1: Resolving bridging anaphors in Example 1 using *BARQA*.

and it captures the common-sense knowledge (i.e., semantic relatedness) between NPs.

Different from the algorithms mentioned above, our QA model does not require the extracted or gold mentions (NPs) as the input, and it predicts the span of the antecedent for a bridging anaphor directly.

Question Answering. *Reading comprehension* or *question answering based on context* has attacted much attention within the NLP community, in particular since Rajpurkar et al. (2016) released a large-scale dataset (SQuAD) consisting of 100,000+ questions on a set of paragraphs extracted from Wikipedia articles. Previous work has cast a few traditional NLP tasks as question answering, such as textual entailment (McCann et al., 2018), entity–relation extraction (Li et al., 2019), and coreference resolution (Wu et al., 2020). However, unlike these tasks, we do not have large scale training datasets for bridging. As a result, we form the questions for our task in a more natural way in order to leverage the existing QA datasets (e.g., SQuAD) that require common-sense reasoning. In addition, we generate a large-scale training dataset of quasi-bridging and demonstrate that it is a good pre-training corpus for bridging anaphora resolution.

Recently, Gardner et al. (2019) argue that we should consider question answering as a *format* instead of a *task* in itself. From this perspective, our work can be seen as a specific probing task to test a QA model's ability to understand bridging anaphora based on context.

Winograd Schema Challenge. Bridging anaphora resolution shares some similarities with Winograd Schema Challenge (WSC). Specifically, in both tasks, one has to understand the context to find the antecedents for anaphors. However, the antecedent search space in bridging anaphora resolution is much bigger than the one in WSC. This is because an anaphor (pronoun) and its antecedent in WSC are usually from the same sentence, while bridging pairs usually require cross-sentence inference. For instance, in ISNotes, only around 26% of anaphors have antecedents occurring in the same sentence, and 23% of anaphors have antecedents that are more than two sentences away.

Recently, Kocijan et al. (2019) use some heuristics to generate a large-scale WSC-like dataset and report that the model pre-trained on this dataset achieves the best results on several WSC datasets after being fine-tuned on a small in-domain dataset. We find similar patterns of results for bridging anaphora resolution (see Section 5.3).

3 BARQA: A QA System for Bridging Anaphora Resolution

In this section, we describe our QA system (called *BARQA*) for bridging anaphora resolution in detail. Figure 1 illustrates how *BARQA* predicts antecedents for bridging anaphors in Example 1.

3.1 Problem Definition

We formulate bridging anaphora resolution as a context-based QA problem. More specifically, given a bridging anaphor a and its surrounding context c_a, we rephrase a as a question q_a. The goal is to predict a text span s_a from c_a that is the antecedent of a. We propose to use the span-based QA framework to extract s_a. In general, our *BARQA* system is built on top of the vanilla BERT QA framework (Devlin et al., 2019). We further modify the inference algorithm to guarantee that the answer span s_a should always appear before the bridging anaphor a (see Section 3.4 for more details).

Following Devlin et al. (2019), we present the input question q_a and the context c_a as a single packed sequence "$[cls]$ q_a $[sep]$ c_a" and calculate the probabilities of every word in c_a being the start and end of the answer span. The training objective is the log-likelihood of the correct start and end positions.

3.2 Question Generation

In English, the preposition "***of***" in the syntactic structure "np_1 ***of*** np_2" encodes different associative relations between noun phrases that cover a variety of bridging relations. For instance, "*the chairman of IBM* " indicates *a professional function in an organization*, and "*the price of the stock*" indicates *an attribute of an object*. Poesio et al. (2004) also used such patterns to estimate the part-of bridging relations. These patterns reflect how we explain bridging anaphora as human beings. It seems that the most natural way to understand the meaning of a bridging anaphor a is to find the answer for the question "*a of what?*" from the surrounding context of a.

As a result, in order to generate the corresponding question q_a for a bridging anaphor a, we first create a' by removing all words appearing after the head of a, we then concatenate a' with "*of what?*" to form the question. This is because, as pointed by Hou (2018a), premodifiers of bridging anaphors are essential elements to understand bridging relations. For instance, for the bridging anaphor "**a painstakingly documented report, based on hundreds of interviews with randomly selected refugees**", the corresponding question is "*a painstakingly documented report of what?*".

3.3 Answer Generation

For each bridging anaphor a together with its corresponding question q_a and context c_a described above, we construct a list of answers A that contains all antecedents of a occurring in the context c_a.[3] In addition, for every NP antecedent n from A, we add the following variations which represent the main semantics of n into the answer list:

- the head of n (e.g., *last week's earthquake*)
- n' which is created by removing all postmodifiers from n (e.g., *the preliminary conclusion from a survey of 200 downtown high-rises*)
- n'' which is created by removing all postmodifiers and the determiner from n (e.g., *the total potential claims from the disaster*)

It is worth noting that if the context c_a does not contain any antecedent for the bridging anaphor a (e.g., some anaphors do not have antecedents occurring in c_a if we use a small window size to construct it), we put "*no answer*" into the answer list A.

3.4 Inference

Different from the SQuAD-style question answering where there is no specific requirement for the position of the predicted span, in bridging anaphora resolution, an anaphor must appear after its antecedent. Therefore in the inference stage, for each bridging anaphor a, we first identify the position of a in its context c_a, then we only predict text spans which appear before a. We further prune the list of predicted text spans by only keeping the top k span candidates that contain at most l words (k and l are empirically set to 20 and 5, respectively). We also prune span predictions that are function words (e.g., *a, an, the, this, that*).

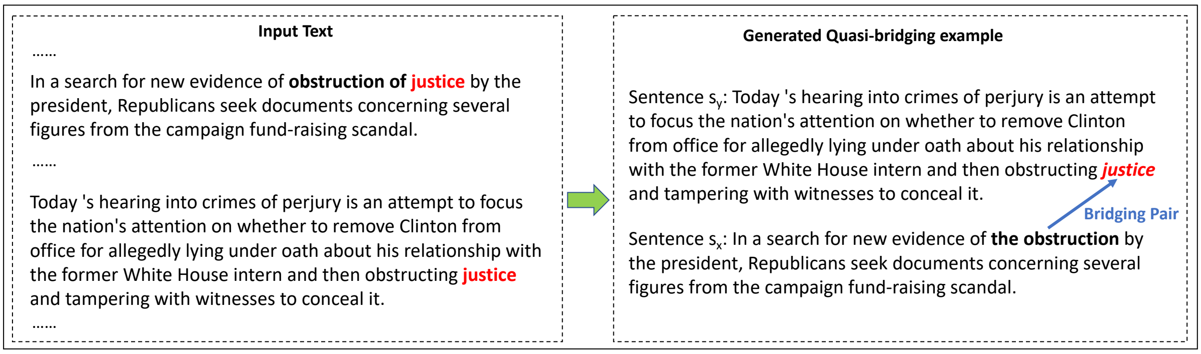
3.5 Training

During the training process, we first use SpanBERT (Joshi et al., 2019) to initialize our BARQA model because it shows promising improvements on SQuAD 1.1 compared to the vanilla BERT embeddings. We then continue to train our model using different pre-training and fine-tuning strategies. Section 5.3 describes different training strategies in detail.

For every training strategy, we train BARQA for five epochs with a learning rate of 3e-5 and a batch size of 24.[4] During training and testing, the maximum text length is set to 128 tokens.

[3] In ISNotes and BASHI, we use gold coreference annotations from OntoNotes (Weischedel et al., 2011) to identify all possible antecedents for every bridging anaphor.

[4] In general, the small learning rate (i.e., 3e-5, 4e-5, and 5e-5) and small fine-tuning epochs are common practices for fine-tuning BERT models. We test the combination of these



Input Text

......

In a search for new evidence of **obstruction of** justice by the president, Republicans seek documents concerning several figures from the campaign fund-raising scandal.

......

Today 's hearing into crimes of perjury is an attempt to focus the nation's attention on whether to remove Clinton from office for allegedly lying under oath about his relationship with the former White House intern and then obstructing **justice** and tampering with witnesses to conceal it.

......

Generated Quasi-bridging example

Sentence s_y: Today 's hearing into crimes of perjury is an attempt to focus the nation's attention on whether to remove Clinton from office for allegedly lying under oath about his relationship with the former White House intern and then obstructing ***justice*** and tampering with witnesses to conceal it.

Sentence s_x: In a search for new evidence of **the obstruction** by the president, Republicans seek documents concerning several figures from the campaign fund-raising scandal.

Figure 2: Examples of generating "quasi-bridging" training data.

4 Generate "Quasi-bridging" Training Data

Bridging anaphora is a complex phenomenon, and there are no large-scale corpora available for referential bridging. In this section, we describe how we generate a large scale "quasi-bridging" dataset.

Hou (2018b) explores the syntactic prepositional and possessive structures of NPs to train word embeddings for bridging. Inspired by this work, we first use these structures to identify "bridging anaphors" and the corresponding "antecedents". Next, we map them back to the discourse to create bridging-like examples.

More specifically, given a text, we first extract NPs containing the prepositional structure (e.g., **X** *preposition* **Y**) or the possessive structure (e.g., **Y** *'s* **X**). In order to have a high-quality set of automatically generated bridging annotations, we apply an additional constraint to the above NPs, i.e., **X** and **Y** should not contain any other NP nodes in the constituent tree. For instance, we do not consider NPs such as "*the political value of imposing sanctions against South Africa*" or "*the cost of repairing the region's transportation system*".

Figure 2 illustrates how we generate a bridging annotation with a sentence pair $\{s_y, s_x\}$ from a raw text[5]: we first extract the NP "*obstruction of justice*" from the sentence s_i and identify **X/Y** in this extracted NP (i.e., **X** = obstruction, **Y** = justice). Next, we collect a list of sentences S from the whole text. Every sentence in S contains **Y** but does not contain **X**. If S contains more than one sentence, we choose the one which is the closest to s_i as s_y. This is because close sentences are more likely semantically related. Finally, we generate the sentence s_x by replacing "*obstruction of justice*" in the original sentence s_i with "***the obstruction***". This gives us a quasi-bridging example with two adjacent sentences (i.e., s_y and s_x) and a bridging link (i.e., *justice* - **the obstruction**).

As a result, we obtain a large amount of "quasi-bridging" training data (i.e., around 2.8 million bridging pairs) by applying the method described above to the NYT19 section of the automatically parsed Gigaword corpus.

In order to understand the quality of our "quasi-bridging" training dataset, we randomly sample 100 quasi-bridging sentence pairs and manually check bridging annotations in these instances. We score each bridging annotation using a scale of 0-2: "2" means that the bridging annotation is correct and the sentence pair sounds natural; "1" indicates that the example makes sense, but it does not sound natural in English; and "0" denotes that the annotation is unacceptable. Overall, we find that 25% of instances and 37% of instances have a score of 2 and 1, respectively. And the remaining 38% of instances are scored as zero. In general, our noisy "quasi-bridging" training dataset does contain a large number of diverse bridging pairs.

5 Experiments

5.1 Datasets

We use four datasets for experiments. The first dataset is ISNotes[6] released by Markert et al.

parameters for various training configurations on a small set (10 documents) of the ISNotes corpus and the BASHI corpus, respectively. On both corpora, we observed that a learning rate of 3e-5, 4e-5, or 5e-5 has minimal impact on results; and for each learning rate, the result continues improving at the beginning (epochs = 1,2,3,4,5), but the performances stays more or less the same after epochs>5.

[5] The raw text is from the Gigaword corpus (Parker et al., 2011; Napoles et al., 2012).

[6] http://www.h-its.org/en/research/nlp/isnotes-corpus

(2012). This dataset contains 50 texts with 663 referential bridging NPs from the World Street Journal (WSJ) portion of the OntoNotes corpus (Weischedel et al., 2011). The second dataset is called BASHI from Rösiger (2018). It contains 459 bridging NPs[7] with 344 referential anaphors from 50 WSJ texts[8]. Note that bridging anaphors in these two corpora are not limited to definite NPs as in previous work (Poesio et al., 1997, 2004; Lassalle and Denis, 2011) and bridging relations are not limited to the prototypical *whole – part* relation or *set – element* relation. We consider these two corpora as expert-annotated in-domain datasets.

We assume that some reasoning skills (e.g., world knowledge, word relatedness) required to answer questions in SQuAD can also be applied for bridging anaphora resolution. Therefore we include the SQuAD 1.1 training data (Rajpurkar et al., 2016) as one training dataset. Another training dataset is the large scale quasi-bridging corpus (*QuasiBridging*) described in Section 4.

Table 1 summarizes the four datasets mentioned above. Note that in ISNotes and BASHI, the number of QA pairs is more than the number of bridging anaphors. This is because an anaphor can have multiple antecedents (e.g., coreferent mentions of the same antecedent entity).

5.2 Experimental Setup

Following Hou (2018a), we use *accuracy* on the number of bridging anaphors to measure systems' performance for resolving bridging anaphors on ISNotes and BASHI. It is calculated as the number of the correctly resolved bridging anaphors divided by the total number of bridging anaphors.

We measure two types of *accuracy*: *lenient accuracy* and *strict accuracy*. In *strict accuracy*, only the original gold antecedent annotations are counted as the correct answers. For *lenient accuracy*, we add the additional variations of the original antecedent annotations (described in Section 3.3) into the correct answer list. For instance, suppose that the gold antecedent annotation is "*the Four Seasons restaurant*", and the predicted span is "*Four Seasons restaurant*", we count this prediction as an incorrect prediction in *strict accuracy* evaluation. However, it is a correct prediction in *lenient accuracy* evaluation.

It is worth noting that our lenient accuracy corresponds to the "exact match" metric in SQuAD (Rajpurkar et al., 2016). The correct answer lists that are generated as described in Section 3.3 can partially address the evaluation problem of imperfect system mention predictions. We do not report F1 score because it will give partial credit for a prediction that does not capture the main semantics of the original gold annotation, such as "*the Four Seasons*".

During evaluation, for every bridging anaphor a, let s_a be the sentence containing a, we use the first sentence of the text, the previous two sentences of s_a, as well as s_a to form a's surrounding context c_a. This is in line with Hou (2018a)'s antecedent candidate selection strategy.

5.3 Results on ISNotes and BASHI Using Different Training Strategies

In this section, we carry out experiments using our *BARQA* system with different training strategies. For every bridging anaphor a, we choose the span with the highest confidence score from its context c_a as the answer for the question q_a and use this span as the predicted antecedent. We report results on ISNotes and BASHI using *lenient accuracy* (see Table 2).

Looking at the results on ISNotes, we find that *BARQA* trained on a small number of in-domain dataset (*BASHI*) achieves an accuracy of 38.16% on ISNotes, which is better than the model trained on the other two large-scale datasets (*SQuAD 1.1* and *QuasiBridging*). However, when using these two datasets to pre-train the model then fine-tuning it with the small in-domain dataset (*BASHI*), both settings (i.e., *SQuAD 1.1* + *BASHI* and *QuasiBridging* + *BASHI*) achieve better results compared to using *BASHI* as the only training dataset. This verifies the value of the *pre-training* + *fine-tuning* strategy, i.e., pre-training the model with large scale out-of-domain or noisy dataset, then fine-tuning it with a small in-domain dataset.

Particularly, we notice that the performance of using *QuasiBridging* alone is worse than the one using *SQuAD 1.1* only. However, combining *QuasiBridging* and *BASHI* achieves the best result on ISNotes, with an accuracy of 47.21%. It seems that the large-scale in-domain noisy training data (*QuasiBridging*) brings more value than the large-scale out-of-domain training data (*SQuAD 1.1*).

We observe similar patterns on the results on

[7] BASHI considers comparative anaphora as bridging anaphora. We exclude them from this study.

[8] Note that these WSJ articles are different from the ones in ISNotes.

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

Table 1: Four datasets used for experiments.

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

Table 2: Results of *BARQA* on ISNotes and BASHI using different training strategies. ⋆ indicates statistically significant differences over the other models (two-sided paired approximate randomization test, $p < 0.01$).

BASHI. Pre-training the model on *QuasiBridging* then fine-tuning it on *ISNotes* achieves the best result with an accuracy of 37.79%. Furthermore, when evaluating on BASHI, it seems that using *SQuAD 1.1* as the pre-training dataset does not bring additional values when combining it with *ISNotes*.

5.4 Results on ISNotes and BASHI Compared to Previous Approaches

Previous work for bridging anaphora resolution on ISNotes and BASHI use gold/system mentions as antecedent candidates and report results using *strict accuracy* (Hou et al., 2013b; Hou, 2018a).

In order to fairly compare against these systems, for every bridging anaphor a, we first map all top 20 span predictions of our system *BARQA* to the gold/system mentions, then we choose the gold/system mention with the highest confidence score as the predicted antecedent. Specifically, we map a predicted span s to a mention m if they share the same head and s is part of m' (m' is created by removing all postmodifiers from m). For instance, "*total potential claims*" is mapped to the mention "*the total potential claims from the disaster*". If a predicted span can not be mapped to any gold/system mentions, we filter it out. Following Hou (2018a), we only keep the predictions whose semantic types are "time" if a is a time expression. The above process is equal to using gold/system mentions and their semantic information to further prune *BARQA*'s span predictions.

Table 3 and Table 4 compare the results of our system *BARQA* against previous studies for bridging anaphora resolution on ISNotes and BASHI, respectively. For both datasets, the *BARQA* model is trained using the best strategy reported in Table 2 (pre-training with *QuasiBridging* + fine-tuning with small in-domain data).

On ISNotes, previously Hou (2018a) reported the best result by adding the prediction from a deterministic algorithm (*embeddings_bridging (NP head + modifiers)*) as an additional feature into the global inference model (*MLN II*) proposed by Hou et al. (2013b). The deterministic algorithm is based on word embeddings for bridging and models the meaning of an NP based on its head noun and modifications.

Our system *BARQA*, when using the gold mentions together with their semantic information to further prune the span predictions, achieves the new state-of-the-art result on ISNotes, with a strict accuracy of 50.08% (see *BARQA with gold mentions/semantics, strict accuracy* in Table 3). How-

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

Table 3: Results of different systems for bridging anaphora resolution in ISNotes. Bold indicates statistically significant differences over the other models (two-sided paired approximate randomization test, $p < 0.01$).

System	Use System Mentions	Accuracy
Model from Hou (2018a)		
embeddings_bridging (NP head + modifiers)	yes	29.94
This work		
BARQA with system mentions/semantics, strict accuracy	yes	**38.66**
BARQA without mention information, strict accuracy	no	32.27
BARQA without mention information, lenient accuracy	no	**37.79**

Table 4: Results of different systems for bridging anaphora resolution in BASHI. Bold indicates statistically significant differences over the other models (two-sided paired approximate randomization test, $p < 0.01$).

ever, we argue that using gold mention information to construct the set of antecedent candidates is a controlled experiment condition, and our experiment setup *BARQA without mention information, lenient accuracy* is a more realistic scenario in practice.

On BASHI, Hou (2018a) reported an accuracy of 29.94% (*strict accuracy*) using automatically extracted mentions from the gold syntactic tree annotations. Our system *BARQA* without any mention/semantic information achieves an accuracy of 32.27% using the same *strict accuracy* evaluation. The result of *BARQA* is further improved with an accuracy of 38.66% when we integrate mention/semantic information into the model.

Note that Hou (2018a) also adapted their deterministic algorithm to resolve lexical bridging anaphors on ARRAU (Poesio et al., 2018) and reported an accuracy of 32.39% on the *RST Test* dataset. Although in this paper we do not focus on lexical bridging, our model *BARQA* can also be applied to resolve lexical bridging anaphors. We found that *BARQA* trained on the *RST Train* dataset alone with around 2,000 QA pairs achieves an accuracy of 34.59% on the *RST Test* dataset.

6 Error Analysis

In order to better understand our model, we automatically label bridging anaphors in ISNotes as either "*referential bridging/world-knowledge*" or "*referential bridging/context-dependent*". We then analyze the performance of *BARQA* and the best model from Hou (2018a) on these two categories.

Rösiger et al. (2018) pointed out that although lexical and referential bridging are two different concepts, sometimes they can co-occur within the same pair of expressions. In Example 2, "**Employees**" is an anaphoric expression. At the same time, the relation between the antecedent entity "{*Mobil Corp./the company's*}" and the bridging anaphor "**Employees**" corresponds to the common-sense world knowledge which is true without any specific context. We call such cases as "*referential bridging/world-knowledge*". Differently, we call a bridging anaphor as "*referential bridging/context-dependent*" if it has multiple equally plausible antecedent candidates according to the common-sense world knowledge about the NP pairs and we have to analyze the context to choose the antecedent (see Example 1). One may

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

Table 5: Comparison of the percentage of correctly resolved anaphors between *BARQA* and the best model from Hou (2018a) on two bridging categories.

argue that "{the exploration and production division – **Employees**}" in Example 2 is also a valid common-sense knowledge fact, however, we consider that it is less prominent than "{*the company's* – **Employees**}".

(2) *Mobil Corp.* is preparing to slash the size of its workforce in the U.S., possibly as soon as next month, say individuals familiar with *the company's* strategy. The size of the cuts isn't known, but they'll be centered in the exploration and production division, which is responsible for locating oil reserves, drilling wells and pumping crude oil and natural gas. **Employees** haven't yet been notified.

For a bridging anaphor a, the deterministic algorithm (*embeddings_bridging*) from Hou (2018a) uses a word representation resource learned from a large corpus to predict the most semantically related NP among all NP candidates as the antecedent. The predictions from this system reflect the common-sense world knowledge about the NP pairs. We thus use this algorithm to label bridging anaphors in ISNotes: if a bridging anaphor is correctly resolved by *embeddings_bridging*, we label it as "*referential bridging/world-knowledge*", otherwise the label is "*referential bridging/context-dependent*".

Table 5 compares the percentage of correctly resolved anaphors between *BARQA* with gold mentions and the best model from Hou (2018a) (*MLN II + emb*) on the two bridging categories. Note that *MLN II + emb* contains several context-level features (e.g., document span, verb pattern). Overall, it seems that our *BARQA* model is better at resolving context-dependent bridging anaphors.

7 Conclusions

In this paper, we model bridging anaphora resolution as a question answering problem and propose a QA system (*BARQA*) to solve the task.

We also propose a new method to automatically generate a large scale of "quasi-bridging" training data. We show that our QA system, when trained on this "quasi-bridging" training dataset and fine-tuned on a small amount of in-domain dataset, achieves the new state-of-the-art results on two bridging corpora.

Compared to previous systems, our model is simple and more realistic in practice: it does not require any gold annotations to construct the list of antecedent candidates. Moreover, under the proposed QA formulation, our model can be easily strengthened by adding other span-based text understanding QA corpora as pre-training datasets.

Finally, we will release our experimental QA datasets (in the SQuAD json format) for bridging anaphora resolution on ISNotes and BASHI. They can be used to test a QA model's ability to understand a text in terms of bridging inference.

Acknowledgments

The author appreciates the valuable feedback from the anonymous reviewers.

References

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,* Minneapolis, USA, 2–7 June 2019, pages 4171–4186.

Kerstin Eckart, Arndt Riester, and Katrin Schweitzer. 2012. A discourse information radio news database for linguistic analysis. In Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann, editors, *Linked Data in Linguistics*, pages 65–76. Springer Berlin Heidelberg.

Kari Fraurud. 1990. Definiteness and the processing of noun phrases in natural discourse. *Journal of Semantics*, 7:395–433.

Claire Gardent and Hélène Manuélian. 2005. Création d'un corpus annoté pour le traitement des descriptions définies. *Traitement Automatique des Langues*, 46(1):115–140.

Matt Gardner, Jonathan Berant, Hannaneh Hajishirzi, Alon Talmor, and Min Sewon. 2019. Question answering is a format; when is it useful? *arXiv preprint arXiv:909.11291.*

Yufang Hou. 2016. Incremental fine-grained information status classification using attention-based LSTMs. In *Proceedings of the 26th International Conference on Computational Linguistics,* Osaka, Japan, 11–16 December 2016, pages 1880–1890.

Yufang Hou. 2018a. A deterministic algorithm for bridging anaphora resolution. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing,* Brussels, Belgium, 31 October– 4 November 2018, pages 1938–1948.

Yufang Hou. 2018b. 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,* New Orleans, Louisiana, 1–6 June 2018, pages 1–7.

Yufang Hou, Katja Markert, and Michael Strube. 2013a. 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,* Seattle, Wash., 18–21 October 2013, pages 814–820.

Yufang Hou, Katja Markert, and Michael Strube. 2013b. Global inference for bridging anaphora resolution. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* Atlanta, Georgia, 9–14 June 2013, pages 907–917.

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,* Doha, Qatar, 25–29 October 2014, pages 2082–2093.

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. 2019. Spanbert: Improving pre-training by representing and predicting spans. *arXiv preprint arXiv:1907.10529*.

Vid Kocijan, Ana-Maria Cretu, Oana-Maria Camburu, Yordan Yordanov, and Thomas Lukasiewicz. 2019. A surprisingly robust trick for the Winograd Schema Challenge. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,* Florence, Italy, 28 July–2 August 2019, pages 4837–4842.

Emmanuel Lassalle and Pascal Denis. 2011. Leveraging different meronym discovery methods for bridging resolution in French. In *Proceedings of the 8th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2011),* Faro, Algarve, Portugal, 6–7 October 2011, pages 35–46.

Xiaoya Li, Fan Yin, Zijun Sun, Xiayu Li, Arianna Yuan, Duo Chai, Mingxin Zhou, and Jiwei Li. 2019. Entity-relation extraction as multi-turn question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,* Florence, Italy, 28 July–2 August 2019, pages 1340–1350.

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,* Jeju Island, Korea, 8–14 July 2012, pages 795–804.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*.

Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. 2012. Annotated Gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction & Web-scale Knowledge Extraction (AKBC-WEKEX)* Montréal, Québec, Canada, 7-8 June 2012, pages 95–100.

Anna Nedoluzhko, Jiří Mírovský, and Petr Pajas. 2009. The coding scheme for annotating extended nominal coreference and bridging anaphora in the Prague dependency treebank. In *Proceedings of the Third Linguistic Annotation Workshop at ACL-IJCNLP 2009,* Suntec, Singapore, 6–7 August 2009, pages 108–111.

Malvina Nissim, Shipara Dingare, Jean Carletta, and Mark Steedman. 2004. An annotation scheme for information status in dialogue. In *Proceedings of the 4th International Conference on Language Resources and Evaluation,* Lisbon, Portugal, 26–28 May 2004, pages 1023–1026.

Juri Opitz and Anette Frank. 2018. Addressing the Winograd Schema Challenge as a sequence ranking task. In *Proceedings of the First International Workshop on Language Cognition and Computational Models*, pages 41–52.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English Gigaword Fifth Edition. LDC2011T07.

Massimo Poesio. 2004. The MATE/GNOME proposals for anaphoric annotation, revisited. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue,* Cambridge, Mass., 30 April – 1 May 2004, pages 154–162.

Massimo Poesio, Yulia Grishina, Varada Kolhatkar, Nafise Sadat Moosavi, Ina Rösiger, Adam Roussel, Fabian Simonjetz, Alexandra Uma, Olga Uryupina, Juntao Yu, and Heike Zinsmeister. 2018. Anaphora resolution with the ARRAU corpus. In *Proceedings of the Workshop on Computational Models of Reference, Anaphora and Coreference.* New Orleans, Louisiana, June 6, 2018, pages 11–22.

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,* Barcelona, Spain, 21–26 July 2004, pages 143–150.

Massimo Poesio and Renata Vieira. 1998. A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216.

Massimo Poesio, Renata Vieira, and Simone Teufel. 1997. Resolving bridging references in unrestricted text. In *Proceedings of the ACL Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Text, Madrid, Spain, July 1997*, pages 1–6.

Altaf Rahman and Vincent Ng. 2012. Resolving complex cases of definite pronouns: The Winograd Schema Challenge. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing and Natural Language Learning,* Jeju Island, Korea, 12–14 July 2012, pages 777–789.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100, 000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing,* Austin, Texas, USA, 1–4 November 2016, pages 2383–2392.

Ina Rösiger. 2018. BASHI: A corpus of wall street journal articles annotated with bridging links. In *Proceedings of the 11th International Conference on Language Resources and Evaluation,* Miyazaki, Japan, 7–12 May 2018, pages 382–388.

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,* Santa Fe, New-Mexico, USA, 20–26 August 2018, pages 3516–3528.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2011. OntoNotes release 4.0. LDC2011T03, Philadelphia, Penn.: Linguistic Data Consortium.

Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li. 2020. Coreference resolution as query-based span prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* Seattle, Wash., 5–10 July 2020.