Multitask Learning-Based Neural Bridging Reference Resolution

Juntao Yu and **Massimo Poesio**
Queen Mary University of London
{juntao.yu, m.poesio}@qmul.ac.uk

Abstract

We propose a multi task learning-based neural model for resolving bridging references tackling two key challenges. The first challenge is the lack of large corpora annotated with bridging references. To address this, we use multi-task learning to help bridging reference resolution with coreference resolution. We show that substantial improvements of up to 8 p.p. can be achieved on full bridging resolution with this architecture. The second challenge is the different definitions of bridging used in different corpora, meaning that hand-coded systems or systems using special features designed for one corpus do not work well with other corpora. Our neural model only uses a small number of corpus independent features, thus can be applied to different corpora. Evaluations with very different bridging corpora (ARRAU, ISNOTES, BASHI and SCICORP) suggest that our architecture works equally well on all corpora, and achieves the SoTA results on full bridging resolution for all corpora, outperforming the best reported results by up to 36.3 p.p..[1]

1 Introduction

Anaphora resolution (Karttunen, 1976; Webber, 1979; Kamp and Reyle, 1993; Garnham, 2001; Poesio et al., 2016) is the aspect of language interpretation concerned with linking nominal expressions to entities in the context of interpretation (or **discourse model**). As illustrated by (1) (adapted from (Hou et al., 2018)), nominal expressions can be linked to the context in several ways: **corefererence** (linking [The Bakersfield Supermarket], [The business], [its]), **bridging or associative reference** (linking [the customers] to [the supermarket]) (Clark, 1975; Prince, 1981; Poesio et al., 2004; Hou et al., 2018), and **discourse deixis** (linking [the murder] to the event of murdering) (Webber, 1991; Kolhatkar et al., 2018).

(1) [The Bakersfield Supermarket] went bankrupt last May. [The business] closed when [[its] old owner] was murdered by [robbers]. [The murder] saddened [the customers].

Bridging reference resolution is the sub-task of anaphora resolution concerned with identifying and resolving bridging references, i.e., anaphoric references to non-identical associated antecedents. Bridging resolution is much less studied than the closely related sub-task of coreference resolution, which has received a lot of attention ((Pradhan et al., 2012; Wiseman et al., 2015; Lee et al., 2017; Lee et al., 2018), to mention just a few recent proposals). One reason for this is the lack of training data. Several corpora have been annotated with bridging reference, including e.g. GNOME (Poesio, 2004), ISNOTES (Markert et al., 2012), SCICORP (Rösiger, 2016) and BASHI (Rösiger, 2018), but they are all rather small, with at most around 1k examples of bridging reference. ARRAU (Poesio and Artstein, 2008; Uryupina et al., 2019) is much larger, but still contains only 5.5k bridging pairs. It is challenging to train a learning based system on that amount of data, particularly the new neural models. As a result, the current SoTA systems for full bridging resolution are still rule-based, employing a number of heuristic rules many of which are corpus-dependent (Hou et al., 2014; Roesiger et al., 2018). This is problematic at the light of the second challenge for work in this area: namely, that the definitions of bridging are different in these

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/.

[1]The code is available at https://github.com/juntaoy/dali-bridging

Proceedings of the 28th International Conference on Computational Linguistics, pages 3534–3546
Barcelona, Spain (Online), December 8-13, 2020

different corpora (Roesiger et al., 2018). Existing corpora differ in whether they attempt to annotate only what Roesiger et al call **referential bridging** (as in ISNOTES), or the full range of bridging references, as in ARRAU.[2] The ISNOTES, BASHI and SCICORP corpora consist mostly of referential bridging examples, while the ARRAU corpus contains both types of bridging references. As a consequence, a system designed for one corpus (e.g. ISNOTES) works poorly when applied to other corpora (e.g. ARRAU), and significant modifications are needed to make the system works equally well on different corpora (Roesiger, 2018).

To tackle these challenges, we introduce a multi task learning-based neural model that learns bridging resolution together with coreference resolution. Multi task architectures have proven effective at exploiting the synergies between distinct but related tasks to in cases when only limited amounts of data are available for one or more of the tasks (Clark et al., 2019). Such an architecture should therefore be especially suited for our context, given that, linguistically, bridging reference resolution and coreference resolution are two distinct but closely related aspects of anaphora resolution, and indeed were often tackled together in early systems (Sidner, 1979; Vieira and Poesio, 2000). Using a neural network-based approach that minimises feature engineering enables the system to be more flexible on the choices of corpora. We mainly evaluate our system on the RST portion of the ARRAU corpus since it is the largest available resource, but we additionally evaluate it on the TRAINS and PEAR portion of the ARRAU corpus, ISNOTES, BASHI and SCICORP corpus to demonstrate its tolerance of diversity.

We start with a strong baseline for bridging adapted from the SoTA coreference architecture (Lee et al., 2018; Kantor and Globerson, 2019) enhanced by BERT embeddings (Devlin et al., 2019). We extend the system to multi-task learning by adding a coreference classifier that shares part of the network with the bridging classifier. In this way, we improve full bridging resolution and its subtasks (anaphor recognition and antecedent selection) by 6.5-7.3 p.p. respectively. But because the number of coreference examples is much larger than the number of bridging pairs, the dataset is highly imbalanced. We achieve further improvements of 1.7 p.p. and 6.6 p.p. on full bridging resolution and anaphor recognition by using undersampling during the training. This final system achieves SoTA results on both full bridging resolution and its subtasks, i.e. 4.5, 6 and 9.5 p.p. higher than the best reported results (Roesiger, 2018) on full bridging resolution, anaphor recognition and antecedent selection respectively. Evaluation on TRAINS, PEAR, ISNOTES, BASHI and SCICORP shows the same trend. Although the datasets are much smaller and the annotation schemes for ISNOTES, BASHI and SCICORP are different from ARRAU, our system works equally well, achieving the new SoTA results on full bridging resolution and anaphor recognition for all six corpora as well as five corpora on antecedent selection.

2 Related Work

2.1 Bridging Reference Resolution

Bridging reference resolution involves two subtasks: anaphor recognition and antecedent selection (Hou et al., 2018). Early work on bridging resolution mostly focused on definite bridging anaphors (Sidner, 1979; Vieira and Poesio, 2000; Lassalle and Denis, 2011), but later systems covered unrestricted antecedent selection (Poesio et al., 2004; Hou et al., 2013). Hou et al. (2013) introduced a model based on Markov logic networks and using an extensive set of features and constraints. They evaluated the system with both local and global features on ISNOTES, and showed that global features can greatly improve performance. The system was later extended in (Hou, 2018b; Hou, 2018a; Hou et al., 2018) to explore additional features from embeddings tailored for bridging resolution, to advanced antecedent candidate selection using the Penn Discourse Treebank (*d-scope-salience*). Recently, Hou (2020) framed the antecedent selection task as question answering, and pre-trains the system with a large synthetic bridging corpus; this system achieves SoTA results on ISNOTES. But those systems are highly specialized for the

[2]Roesiger et al use 'referential bridging' for the cases in which the bridging reference needs an antecedent in order to be interpretable, such as the door in *John walked towards the house. The door was open.*. 'Lexical bridging' is when the bridging reference could also be interpreted autonomously, such as Madrid in *I went to Spain last year. I particularly liked Madrid.* See (Poesio, 2004; Baumann and Riester, 2012; Markert et al., 2012; Hou et al., 2018; Uryupina et al., 2019) for a detailed discussion of the annotation schemes, and (Roesiger et al., 2018; Roesiger, 2018) for a discussion of the implications.

ISNOTES corpus, hence perform less well on other corpora. The anaphor recognition subtask is usually solved as a part of the information status task (Markert et al., 2012; Hou, 2016; Hou et al., 2018).

Recent systems for full bridging resolution include (Hou et al., 2014; Hou et al., 2018; Roesiger et al., 2018; Roesiger, 2018). Hou et al. (2014) proposed a rule-based system for full bridging resolution with the ISNOTES corpus, consisting of a rich system of rules motivated by linguistic knowledge. They also evaluated a learning-based system that uses the rules as features, but the learning-based system only outperforms the rule-based system's F1 score by 0.1 percentage points. The rule-based system was later adapted by Roesiger et al. (2018); Roesiger (2018) for full bridging resolution on ARRAU, but since ARRAU follows a more general definition of bridging, most of the rules had to be changed.

2.2 Multi-task Learning for Under-Resourced Tasks

Multi-task learning has been successfully used in several NLP applications (Collobert and Weston, 2008; Luong et al., 2016; Kiperwasser and Ballesteros, 2018; Clark et al., 2019). Normally, the goal of multi-task learning is to improve performance on all tasks; but in an under-resourced setting, the aim often is only to improve performance on the low resource task/language/domains (the **target task**). This is sometimes known as **shared representation based transfer learning**. Yang et al. (2017) applied transfer learning to sequence labelling tasks; the deep hierarchical recurrent neural network used in their work is fully/partially shared between the source and the target tasks. They demonstrated that SoTA performance can be achieved by using models trained on multi-tasks. Cotterell and Duh (2017) trained a neural NER system on a combination of high-/low-resource languages to improve NER for the low-resource languages. In their work, character-based embeddings are shared across the languages. Recently, Zhou et al. (2019) introduced a multi-task network together with adversarial learning for under-resourced NER. The evaluation on both cross-language and cross-domain settings shows that partially sharing the BiLSTM works better for cross-language transfer, while for cross-domain setting, the system performs better when the LSTM layers are fully shared.

2.3 Neural Coreference Resolution

By contrast with bridging reference, coreference resolution has been extensively studied. Wiseman et al. (2015; Wiseman et al. (2016) first introduced a neural network-based approach to solving coreference in a non-linear way. Clark and Manning (2016) integrated reinforcement learning to let the model, optimized directly on the B^3 scores. Lee et al. (2017) proposed a neural joint approach for mention detection and coreference resolution. Their model does not rely on parse trees; instead, the system learns to detect mentions by exploring the outputs of a BiLSTM. After the introduction of context dependent word embeddings such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019), the Lee et al. (2017) system has been greatly improved by those embeddings (Lee et al., 2018; Kantor and Globerson, 2019) to achieve SoTA results. We use a simplified version of the model by (Lee et al., 2018; Kantor and Globerson, 2019) as baseline.

3 Methods

3.1 The Single-Task Baseline System

We use as our single-task baseline for bridging reference a simplified version of the SoTA coreference systems by Lee et al. (2018; Kantor and Globerson (2019), since bridging resolution is closely related to coreference: like coreference it requires establishing a link to an entity in the discourse model, but through a non-identity relation. The Kantor and Globerson (2019) model is an extended version of (Lee et al., 2018); the main difference is that Kantor et al use BERT embeddings (Devlin et al., 2019) instead of the ELMo embeddings (Peters et al., 2018) used by Lee et al. These systems have similar architecture and both do mention detection and coreference jointly. We only use the coreference part of the system, since for bridging resolution evaluation is usually on gold mentions.

Our baseline system first creates representations for mentions using the output of a BiLSTM. The sentences of a document are encoded from both directions to obtain a representation for each token in the sentence. The BiLSTM takes as input the concatenated embeddings ($(emb_t)_{t=1}^{T}$) of both word and

character levels. For word embeddings, GloVe (Pennington et al., 2014) and BERT (Devlin et al., 2019) embeddings are used. Character embeddings are learned from a convolution neural network (CNN) during training. The tokens are represented by concatenated outputs from the forward and the backward LSTMs. The token representations $(x_t)_{t=1}^{T}$ are used together with head representations (h_i) to represent mentions (M_i). The h_i of a mention is obtained by applying attention over its token representations ($\{x_{b_i}, ..., x_{e_i}\}$), where b_i and e_i are the indices of the start and the end of the mention respectively. Formally, we compute h_i, M_i as follows:

$$\alpha_t = \text{FFNN}_\alpha(x_t) \ , \ a_{i,t} = \frac{exp(\alpha_t)}{\sum_{k=b_i}^{e_i} exp(\alpha_k)}$$

$$h_i = \sum_{t=b_i}^{e_i} a_{i,t} \cdot x_t \ , \ M_i = [x_{b_i}, x_{e_i}, h_i, \phi(i)]$$

where $\phi(i)$ is the mention width feature embeddings. Next, we pair the mentions with candidate antecedents to create a pair representation ($P_{(i,j)}$):

$$P_{(i,j)} = [M_i, M_j, M_i \circ M_j, \phi(i,j)]$$

where M_i, M_j is the representation of the antecedent and anaphor respectively, $\circ$ denotes element-wise product, and $\phi(i,j)$ is the distance feature between a mention pair. To make the model computationally tractable, we consider for each mention a maximum 150 candidate antecedents[3].

The next step is to compute the pairwise score ($s(i,j)$). Following Lee et al. (2018), we add an artificial antecedent ϵ to deal with cases of non-bridging anaphor mentions or cases when the antecedent does not appear in the candidate list. We compute $s(i,j)$ as follows:

$$s(i,j) = \begin{cases} 0 & i = \epsilon \\ \text{FFNN}(P_{(i,j)}) & i \neq \epsilon \end{cases}$$

For each mention the predicted antecedent is the one has the highest $s(i,j)$, a bridging link will be only created if the predicted antecedent is not ϵ.

3.2 Our Multi-task Learning Architecture

Choosing a source task that is closely related to bridging resolution is crucial to the success of our multi-task learning model. In this work, we use coreference as the source task. The key intuitions behind this choice are: (i) from a language interpretation point of view, resolving anaphoric coreference and anaphoric bridging reference are closely related tasks in that they both involve trying to identify relations between anaphors and antecedents (Poesio, 2016)–indeed, the two tasks were typically tackled jointly by non ML-based anaphora resolution systems (Sidner, 1979; Hobbs et al., 1993; Vieira and Poesio, 2000); (ii) from the point of view of the model, both tasks rely on a good mention representations and can be solved by neural mention pair models.

We turn our model into a multi-task model by adding to the architecture an additional classifier for coreference, and jointly predicting coreference and bridging (Figure 1). We use the same candidate antecedents for both bridging and coreference tasks. As shown in Figure 1, our model uses shared mention representations (i.e. the word embeddings and the BiLSTM) with additional options to share some/all hidden layers of the FFNN. By sharing most of the network structure, the mention representations learned by the coreference task become accessible by bridging resolution.

3.3 Learning with Imbalanced Data

Following Lee et al. (2018) we optimise our system on the marginal log-likelihood of all correct antecedents. For bridging, we consider a bridging antecedent correct if it is from the same gold coreference cluster $\text{GOLD}(i)$ of the gold bridging antecedent. For coreference, the correct antecedents is implied from the gold coreference cluster $\text{GOLD}(j)$ the mention belongs to. We compute both bridging and coreference losses as follows:

[3]The number of maximum antecedents was tuned on the dev set.

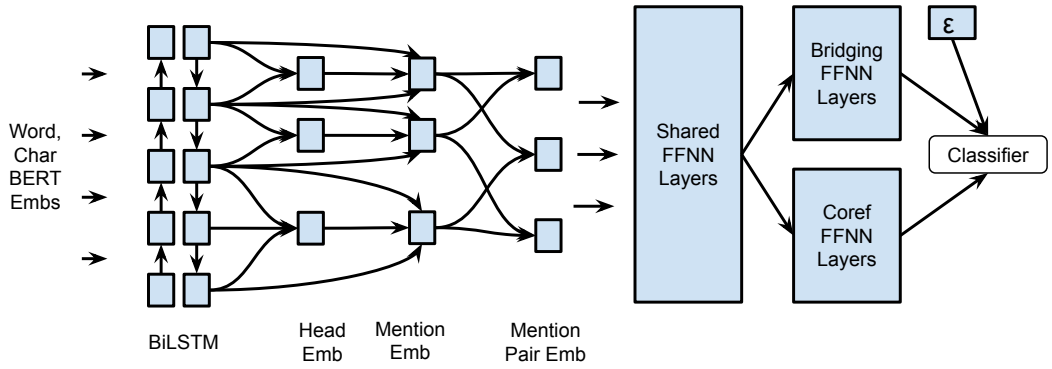

Figure 1: The proposed multi-task architecture.

Corpus	Genre	Bridging Type	Mention Type	Number of Bridging
ARRAU RST	WSJ news	lexical, referential bridging	Gold	3303
ARRAU TRAINS	Dialogues	lexical, referential bridging	Gold	558
ARRAU PEAR	Narrative	lexical, referential bridging	Gold	303
ISNOTES	WSJ news	referential bridging	Gold	663
BASHI	WSJ news	referential bridging	Predicted	459
SCICORP	Scientific texts	referential bridging	Predicted	1366

Table 1: The statistics of bridging corpora used in our experiments.

$$log \prod_{j=1}^{N} \sum_{\hat{y} \in Y(j) \cap \text{GOLD}(i/j)} s(\hat{y}, j)$$

in case mention i is not a bridging/coreference anaphor or $Y(j)$ (the candidate antecedents) does not contain mentions from $\text{GOLD}(i)$ for bridging or $\text{GOLD}(j)$ for coreference, we set $\text{GOLD}(i/j) = \{\epsilon\}$.

When training with coreference one of the problem we need to face is class imbalance. Consider the RST portion of ARRAU as an example (mostly WSJ text). The corpus contains 72k mentions in total: of these, 45k (63%) are discourse-new (DN), 24k (33%) are discourse-old (DO), and the remaining 3k (4%) are bridging anaphors. Training the model on such an imbalanced corpus may significantly harm recall with bridging anaphors.

To reduce the negative effect of this imbalance, we use undersampling during to training by randomly removing DN and DO examples to make the corpus more balanced. More precisely, we use a heuristic negative example ratio γ to control the total number of negative examples during the training, so that, e.g., $\gamma = 2$ means we keep 6k DN and 6k DO examples during training. We set a value for γ by trying a few small values in preliminary experiments; they all gave very similar results, hence we set $\gamma = 2$ in the experiments below.

4 Experiments

Datasets We evaluated our systems on ARRAU (Poesio and Artstein, 2008; Uryupina et al., 2019), ISNOTES (Markert et al., 2012), BASHI (Rösiger, 2018) and SCICORP (Rösiger, 2016) with ARRAU RST as our primary dataset as it is substantially larger than other datasets (see Table 1 for more detail).

Bridging references are annotated in ARRAU according to the scheme in (Uryupina et al., 2019), which covers both what Roesiger (2018) call 'lexical' and 'referential' bridging. The corpus was used for Task 2 of the CRAC 2018 shared task (Poesio et al., 2018), focused on bridging resolution. As done in the CRAC shared task, we evaluate our system on all three subcorpus RST,TRAINS and PEAR stories[4]. The RST

[4]We followed Roesiger (2018) in excluding a small portion of bridging links that are problematic (e.g. empty antecedent, split antecedent).

Parameter	Value
BiLSTM layers, size, dropout	3, 200, 0.4
FFNN layers, size, dropout	2, 150, 0.2
CNN filter widths, size	[3,4,5], 50
Char,GloVe,Feature embedding size	8, 300, 20
BERT layer, size	Last 4, 1024
Embedding dropout	0.5
Max num of antecedent	150
Negative example ratio (γ)	2
Optimiser, Learning rate	Adam, 1e-3

Table 2: Hyperparameters for our models.

portion of the corpus, consisting of 413 news documents (1/3 of the WSJ section of the Penn Treebank). We used the default train/dev and test subdivisions. The TRAINS and PEAR portion of the corpus contains 114 dialogues and 20 fictions respectively. Since the TRAINS and PEAR are much smaller we use 10-fold cross validation and report the results on test set to compare with previous work.

The ISNOTES corpus consists of 50 documents from the WSJ portion of ONTONOTES, with 663 bridging pairs annotated as well as fine-grained information status according to the scheme in (Markert et al., 2012). Bridging is annotated as one of the information status subclasses. Like ISNOTES, the BASHI corpus contains 50 documents from the WSJ portion of ONTONOTES. The dataset has 459 bridging pairs annotated according to a annotation scheme similar to that of the ISNOTES corpus (Rösiger, 2018). We follow Hou (2020) to mix use the ISNOTES and BASHI corpora[5]. The SCICORP corpus uses text from a very different domain of scientific texts. The corpus has in total 1366 bridging pairs annotated, again according to its own annotation scheme (Rösiger, 2016). Since those corpora are rather small, we used 10-fold cross-validation to evaluate on them.

Evaluation metrics Following, e.g., Hou et al., we evaluate on both full bridging resolution and its subtasks (anaphor recognition/antecedent selection). For full bridging resolution and anaphor recognition we report F1 scores.[6] For antecedent selection we report accuracy as it uses gold bridging anaphors.

Hyperparameters Apart from the two parameters introduced by our model (maximum number of antecedents and negative example ratio λ), we use the default settings from Lee et al. (2018), and replace their ELMo settings with the BERT settings from Kantor and Globerson (2019). Table 2 summarizesthe hyperparameter settings of our model. We train the models evaluated on the ARRAU RST for 200 epochs, and for 50 epochs the models trained on the other corpora.

5 Results and Discussions

5.1 Evaluation on the ARRAU RST corpus

We first evaluated our multi-task learning based system on the antecedent selection subtask, to assess the suitability of our model for bridging. The antecedent selection subtask uses gold bridging anaphors, hence it is simpler than full bridging resolution which additionally involves identifying the bridging references. Focusing on a simpler task also allowed us to have a clearer view of the effects of multi-task learning. In this experiment, we configured the system to share only the mention representations (the word embeddings and BiLSTM). As illustrated in Table 3a, the baseline system already achieved a pretty good accuracy for this type of task. Although starting from a strong baseline, our multi-task learning based system achieved an improvement of 3.5 percentage points, confirming our hypothesis that coreference is a good source task for bridging.

Sharing The Feed-forward Layers We further extended our model to share the FFNN layers in addition to the mention representations. The FFNN layers have access to pairwise representations to learn

[5]For each fold of the cross-validation, we train the system with 90% of the main corpus plus the full auxiliary corpus and tested on the rest 10% of the main corpus.

[6]Following (Roesiger, 2018) we consider a predicted bridging antecedent is correct when it belongs to the same gold corefernce cluster as the gold bridging antecedent.

System	Shared Network	RST	ISNOTES
bridging only		47.4	33.8
multi-task	embeddings, LSTM	50.9	38.7
	+ 1 FFNN Layer	**54.7**	**43.7**
	+ 2 FFNN Layer	51.7	40.1

(a) antecedent selection

	RST						ISNOTES					
	anaphor rec.			full bridging res.			anaphor rec.			full bridging res.		
System	P	R	F1	P	R	F1	P	R	F1	P	R	F1
bridging only	50.0	12.5	20.0	34.5	8.6	13.8	63.9	16.2	25.8	33.3	8.5	13.5
multi-task	47.3	19.0	27.1	35.5	14.2	20.3	59.3	22.5	32.7	31.5	12.0	17.4
+ undersampling	31.5	36.2	**33.7**	20.6	23.7	**22.0**	45.6	47.2	**46.4**	19.1	19.7	**19.4**

(b) full bridging resolution

Table 3: Parameter tuning on the dev set of ARRAU RST and ISNOTES.

	Number	Baseline	Multi-task
SUBSET	113	49.6	57.5
ELEMENT	89	50.6	52.8
POSS	8	0.0	12.5
OTHER	11	54.6	54.6
UNDERSP-REL	11	27.3	72.7

Table 4: The accuracy comparison between our best multi-task model and the baseline on antecedent selection for different bridging relations. Evaluated on the dev set of ARRAU RST

the relations between the anaphors and the antecedents, hence contain useful information regarding how likely two mentions are to be related. As expected, this additional sharing of the FFNN layers resulted in additional improvements. A further improvement of 3.8 percentage points was achieved by sharing 1 additional FFNN layer. The accuracy drops when both hidden layers are shared between coreference and bridging, but the performance is still higher when compared with the model that only shares mention representations. Overall, the multi-task model achieved a substantial gain of 7.3 percentage points when compared with the system only carrying out bridging reference resolution (see Table 3a).

The ARRAU RST corpus also contains information about the semantic relation between the bridging reference and its antecedent. Five different relations are annotated: subset (SUBSET), set membership (ELEMENT), a generalised possession relation including part-of relations (POSS), *other* NPs such as the other in *Two men entered the pub. One man ordered a beer, the other a glass of wine* (OTHER) and bridging relations that do not fit into any defined classes (UNDERSP-REL). We further compare our best multi-task model with the single-task baseline on different bridging relations to find out which relation was helped the most by multi-task learning. Table 4 shows the result of our comparison on the development set. The development set contains mostly bridging references of type ELEMENT and SUBSET, and a small number of the other three relations. For the two main classes, the multi-task system improved strongly the SUBSET bridges with a large gain of 8%. The improvement on ELEMENT bridges is smaller (2.2%). One possible explanation for the larger improvement with SUBSET is that many SUBSET bridging references in ARRAU RST denote a subtype of a previously introduced type, as in *computers ... personal computers*, and knowledge about coreference may be especially helpful for such cases.

Full Bridging Resolution Having ascertained the benefits of our multi-task model for antecedent selection, we applied the best settings (sharing mention representations and the 1st hidden layer of the

	RST	TRAINS	PEAR	ISNOTES[7]	BASHI	SCICORP
Hou et al. (2013)	-	-	-	41.3	-	-
Hou (2018a)	32.4	-	-	46.5	27.4	-
Roesiger (2018)	39.8	48.9	28.2	-	-	
Hou (2020)	34.6	-	-	**50.1**	-	-
Our model	**49.3**	**50.9**	**61.2**	43.7	**36.0**	**33.4**

Table 5: Comparing our model with the SoTA for antecedent selection.

FFNN) to full bridging resolution as well. We also report the F1 scores for bridging anaphor recognition, a byproduct of full bridging resolution. Table 3b shows a comparison between the single-task baseline and the multi-task models. The baseline model trained without multi-task learning achieved F1 scores of 13.8% and 20% on full bridging resolution and anaphor recognition, respectively. The low F1 scores are mainly due to a poor recall in both tasks, a well known problem with bridging reference resolution. When applying multi-task learning, the F1 scores improve substantially (6.5% and 7.1% for full bridging resolution and anaphor recognition respectively). These F1 improvements are mainly a result of better recall; the precisions of the two models are similar. This suggests that learning with coreference does help the model to capture more correct bridging pairs. However, recall is still much lower than precision. As this might a result of data imbalance, we applied undersampling during training, to train the model on a more balanced dataset. As shown in Table 3b, with undersampling the model has a more balanced precision and recall, and also achieves better F1 scores on both full bridging resolution and anaphor recognition. The new model achieved improvements of 6.6% and 1.7% on anaphor recognition and full bridging resolution, respectively, when compared with the model without undersampling. Overall, our multi-task models showed their merit on both tasks and achieved considerable gains of 8.2% and 13.7% when compared with the single-task system.

Comparison with the State of the Art We then evaluated our model on the test set of ARRAU RST to compare it with the previously reported state of the art on the same dataset. Table 5 shows the comparison on antecedent selection. The best reported system on this task, (Roesiger, 2018), is a modified version of the original rule-based system designed for ISNOTES by Hou et al. (2014). Our system outperforms the current state of the art by nearly 10 percentage points. Table 6 presents the comparison on the full bridging resolution and anaphor recognition. Since the only reported full bridging resolution results on ARRAU (Roesiger, 2018) are evaluated with gold coreferent anaphors removed, we follow the same method to remove gold coreferent anaphors from the evaluation, but we also report the results with coreferent anaphors included for future reference. Filtering out the gold coreferent anaphors the task is easier, resulting in better F1 scores. After filtering out gold coreferent anaphors, our system achieved F1 scores of 24% and 36.7% on full bridging resolution and anaphor recognition respectively, which is 4.5% and 6% higher than the scores reported in Roesiger (2018). Overall, our model achieved the new SoTA results on both full bridging resolution and its subtasks.

5.2 Evaluation on the ARRAU TRAINS and PEAR corpora

We then evaluate our system on the TRAINS and PEAR sub-corpora of ARRAU. For both corpora, the only reported results are by Roesiger (2018). For antecedent selection our system achieved scores 2% and 33% better than theirs on TRAINS and PEAR respectively (see Table 5). For the other two tasks, they only report scores after filtering out the gold coreferent anaphors, when evaluate in the same setting, our system achieved substantial improvements of up to 35.1% and 36.3% for anaphor recognition and full bridging resolution respectively. Overall, our system is substantially better than the Roesiger (2018) system on both TRAINS and PEAR corpora.

[7]Hou et al. (2018) report a score of 50.7%, but their system relies on hand-coded information from the Penn Discourse Treebank for antecedent candidate selection, so that systems is not really comparable to systems using only automatic information. Hence, we follow Hou (2020) and exclude that result from the table.

[8]The results of Hou et al. (2014) are from Roesiger et al. (2018), as they were obtained on an unknown subset of the corpus.

Corpus	Gold Coreference Anaphors Setting	Models	anaphor rec.			full bridging res.		
			P	R	F1	P	R	F1
RST	Keep	Our model	31.8	29.8	30.8	20.2	18.9	19.5
	Remove	Roesiger (2018)	29.2	32.5	30.7	18.5	20.6	19.5
		Our model	37.6	35.9	**36.7**	24.6	23.5	**24.0**
TRAINS	Keep	Our model	49.4	36.0	41.6	33.7	24.6	28.4
	Remove	Roesiger (2018)	39.3	21.8	24.2	27.1	21.8	24.2
		Our model	62.2	40.4	**48.9**	39.2	25.4	**30.9**
PEAR	Keep	Our model	74.7	48.8	59.0	68.4	44.6	54.0
	Remove	Roesiger (2018)	75.0	16.0	26.4	57.1	12.2	20.1
		Our model	81.1	49.6	**61.5**	74.3	45.5	**56.4**
ISNOTES	Keep	Hou et al. (2014) [8]	65.9	14.1	23.2	57.7	10.1	17.2
		Roesiger et al. (2018)	45.9	18.3	26.2	32.0	12.8	18.3
		Our model	53.9	33.6	**41.4**	31.6	19.8	**24.3**
	Remove	Roesiger et al. (2018)	71.6	18.3	29.2	50.0	12.8	20.4
		Our model	58.3	35.1	**43.8**	33.5	20.2	**25.2**
BASHI	Keep	Our model	34.4	34.2	34.3	17.7	17.5	17.6
	Remove	Roesiger et al. (2018)	49.4	20.2	28.7	24.3	10.0	14.1
		Our model	35.3	34.9	**35.1**	18.2	18.0	**18.1**
SCICORP	Keep	Our model	45.0	35.7	39.8	21.5	17.1	19.0
	Remove	Roesiger et al. (2018)	17.7	0.9	8.1	3.2	0.9	1.5
		Our model	52.9	41.2	**46.3**	25.0	19.4	**21.9**

Table 6: Comparing our model with the SoTA for full bridging resolution.

5.3 Evaluation on the ISNOTES corpus

Most recent work on bridging reference resolution was evaluated on ISNOTES; a number of systems were developed for both full bridging resolution (Hou et al., 2014; Roesiger et al., 2018) and antecedent selection (Hou et al., 2013; Hou, 2018b; Hou, 2018a; Hou, 2020). Since the ISNOTES follows a very different annotation scheme than that of ARRAU, to confirm the suitability of our best setting for ARRAU on corpora only containing referential bridging examples (ISNOTES, BASHI and SCICORP) we run additional parameter tuning on the ISNOTES corpus. For parameter tuning we use the same 10 documents used by Roesiger et al. (2018) as a development set, and use the rest 40 documents for training. As shown in Table 3a and Table 3b the results on ISNOTES follows the same trend as for ARRAU RST, the best settings for two corpora remain the same.

To compare with the SoTA systems, we use 10-fold cross-validation to obtain predictions for the whole corpus. On the full bridging resolution task, our system outperforms all the previous results both when coreferent anaphors are included (6%) and when they are excluded (4.8%). The improvements on anaphor recognition are larger, and our system is more than 14% better in both settings.

For antecedent selection, however, our system achieved a result 2.4% better than the best reported system with out the access to large synthetic bridging corpora (Hou et al., 2013), but lower than those obtained with the help of such corpora (Hou, 2018a; Hou, 2020). The synthetic bridging corpora extracted by Hou (2018a) and Hou (2020) are based on the prepositional (e.g., X preposition Y) or the possessive structures (e.g., Y 's X) which is common in the ISNOTES corpus but not in the other corpus (e.g. ARRAU). Models enhanced by such corpora are overfitted to the ISNOTES corpus, the performances are much lower when evaluated on other corpora. For example, those systems achieved a much lower

results on ARRAU RST (14.7% - 16.9%) when compared with ours. In additional, both systems rely on the gold semantic information for selecting candidate antecedents during test time, which is less realistic.

Overall, on the ISNOTES corpus our system achieved a competitive result on antecedent selection and the SoTA on full bridging resolution and anaphor recognition.

5.4 Evaluation on the BASHI corpus

We next compare our system with previous models on the BASHI corpus. Since gold mentions are not annotated in BASHI, we use NPs as our predicted mentions without filtering[9]. For antecedent selection the only reported result on BASHI corpus is from Hou (2018a)[10] (see Table 5). Our system achieved a accuracy of 36%, which is 8.6% better than that of Hou (2018a). Roesiger et al. (2018) reported the only results for full bridging resolution and anaphor recognition. Our system achieves F1 scores that are 6.4% and 4% better than their results on anaphor recognition and full bridging resolution respectively (see Table 6). Overall, our model achieves new SoTA on all three tasks.

5.5 Evaluation on the SCICORP corpus

Finally, we evaluated our system on SCICORP corpus, in which, like in the BASHI corpus, gold mentions are not annotated, so again we used NPs as our predicted mentions. SCICORP consists of scientific documents that are very different from the BASHI (news). As a result, the only reported result on SCICORP, (Roesiger et al., 2018), is rather poor. Roesiger et al.'s rule-based system only achieved 1.5% and 8.1% (F1) for full bridging resolution and anaphor recognition respectively. The poor result is manly due to the system only recognizing less than 1% of the bridging anaphors, which is another example of the sensitivity of rule-based systems to domain shifting. By contrast, our system achieved on this corpus F1 scores of 21.8% and 46.3% for full bridging resolution and anaphor recognition, respectively (Table 6). These scores on SCICORP are broadly in the same range to the scores achieved by our system on the other three corpora, which indicates that our system's performance doesn't deteriorate so badly with domain shifting. In terms of the antecedent selection task, our system achieved an accuracy of 33.4%; to the best of our knowledge, this is the first result for antecedent selection on SCICORP .

6 Conclusions

In this paper we proposed a multi-task neural architecture tackling two major challenges for bridging reference resolution. The first challenge is the lack of very large training datasets, as the largest corpus for bridging reference, ARRAU (Uryupina et al., 2019), only contains 5.5k examples, and other corpora are much smaller (the most used corpus for bridging, ISNOTES (Markert et al., 2012), only contains 663 bridging pairs). The second challenge is that different annotation schemes for bridging are used in different corpora, so designing a system that can be applied to different corpora is complicated. Our results on the ARRAU RST corpus demonstrate that the performance on full bridging resolution and its subtasks can be significantly improved by learning with additional coreference annotations. Our multi-task model achieved substantial improvements of 7.3%-13.7% for full bridging resolution and its subtasks when compared with the single task baseline that learns solely on bridging annotations. As a result, our final system achieved SoTA results in all three tasks. Further evaluation on TRAINS, PEAR, ISNOTES, BASHI and SCICORP demonstrates the robustness of our system under changes of annotation scheme and domain. The very same architecture used for ARRAU RST again achieved SoTA results on full bridging resolution for all five corpora.

Overall, our results suggest that coreference is a useful source task for bridging reference resolution, and our neural bridging architecture is applicable to bridging corpora based on different domain or definitions of bridging.

Acknowledgements

This research was supported in part by the DALI project, ERC Grant 695662.

[9]The NPs do not belong to coreference clusters or bridging relations are treated as non-mention during training.

[10]Hou (2020) only reported results on a subset of the BASHI corpus but not the full corpus.

References

Stefan Baumann and Arndt Riester. 2012. Referential and lexical givenness: semantic, prosodic and cognitive aspects. In *Prosody and Meaning*.

Kevin Clark and Christopher D. Manning. 2016. Improving coreference resolution by learning entity-level distributed representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653, Berlin, Germany, August. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Urvashi Khandelwal, Christopher D. Manning, and Quoc V. Le. 2019. BAM! born-again multi-task networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5931–5937, Florence, Italy, July. Association for Computational Linguistics.

Herbert H. Clark. 1975. Bridging. In *Theoretical Issues in Natural Language Processing*.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *ICML*.

Ryan Cotterell and Kevin Duh. 2017. Low-resource named entity recognition with cross-lingual, character-level neural conditional random fields. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 91–96, Taipei, Taiwan, November. Asian Federation of Natural Language Processing.

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, June. Association for Computational Linguistics.

Alan Garnham. 2001. *Mental models and the interpretation of anaphora*. Psychology Press.

Jerry R. Hobbs, Mark Stickel, Doug Appelt, and Paul Martin. 1993. Interpretation as abduction. *Artificial Intelligence Journal*, 63:69–142.

Yufang Hou, Katja Markert, and Michael Strube. 2013. 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*, pages 907–917, Atlanta, Georgia, June. 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, October. Association for Computational Linguistics.

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

Yufang Hou. 2016. Incremental fine-grained information status classification using attention-based LSTMs. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1880–1890, Osaka, Japan, December. The COLING 2016 Organizing Committee.

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

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, Volume 2 (Short Papers)*, pages 1–7, New Orleans, Louisiana, June. Association for Computational Linguistics.

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

Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic*. D. Reidel, Dordrecht.

Ben Kantor and Amir Globerson. 2019. Coreference resolution with entity equalization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 673–677, Florence, Italy, July. Association for Computational Linguistics.

Lauri Karttunen. 1976. Discourse referents. In *Syntax and Semantics 7 - Notes from the Linguistic Underground*, pages 363–385. Academic Press, New York.

Eliyahu Kiperwasser and Miguel Ballesteros. 2018. Scheduled multi-task learning: From syntax to translation. *Transactions of the Association for Computational Linguistics*, 6:225–240.

Varada Kolhatkar, Adam Roussel, Stefanie Dipper, and Heike Zinsmeister. 2018. Anaphora with non-nominal antecedents in computational linguistics: a Survey. *Computational Linguistics*, 44(3):547–612.

Emmanuel Lassalle and Pascal Denis. 2011. Leveraging different meronym discovery methods for bridging resolution in french. In *Proc. of 8th DAARC*, pages 35–46, Faro, Portugal.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark, September. Association for Computational Linguistics.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692, New Orleans, Louisiana, June. Association for Computational Linguistics.

Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task sequence to sequence learning. *ICLR*.

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, July. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June. Association for Computational Linguistics.

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

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, July.

Massimo Poesio, Roland Stuckardt, and Yannick Versley. 2016. *Anaphora Resolution: Algorithms, Resources and Applications*. Springer, Berlin.

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

Massimo Poesio. 2004. Discourse annotation and semantic annotation in the GNOME corpus. In *Proc. of the ACL Workshop on Discourse Annotation*.

Massimo Poesio. 2016. Linguistic and cognitive evidence about anaphora. In M. Poesio, R. Stuckardt, and Y. Versley, editors, *Anaphora Resolution: Algorithms, Resources and Applications*, chapter 2. Springer.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Proceedings of the Sixteenth Conference on Computational Natural Language Learning (CoNLL 2012)*, Jeju, Korea.

Ellen F. Prince. 1981. Toward a taxonomy of given-new information. In P. Cole, editor, *Radical Pragmatics*, pages 223–256. Academic Press, New York.

Ina Roesiger, 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, August. Association for Computational Linguistics.

Ina Roesiger. 2018. Rule- and learning-based methods for bridging resolution in the ARRAU corpus. In *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 23–33, New Orleans, Louisiana, June. Association for Computational Linguistics.

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

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

Candace L. Sidner. 1979. *Towards a computational theory of definite anaphora comprehension in English discourse*. Ph.D. thesis, MIT.

Olga Uryupina, Ron Artstein, Antonella Bristot, Federica Cavicchio, Francesca Delogu, Kepa J. Rodriguez, and Massimo Poesio. 2019. Annotating a broad range of anaphoric phenomena, in a variety of genres: the ARRAU corpus. *Journal of Natural Language Engineering*.

Renata Vieira and Massimo Poesio. 2000. An empirically-based system for processing definite descriptions. *Computational Linguistics*, 26(4):539–593, December.

Bonnie L. Webber. 1979. *A Formal Approach to Discourse Anaphora*. Garland, New York.

Bonnie L. Webber. 1991. Structure and ostension in the interpretation of discourse deixis. *Language and Cognitive Processes*, 6(2):107–135.

Sam Wiseman, Alexander M. Rush, Stuart Shieber, and Jason Weston. 2015. Learning anaphoricity and antecedent ranking features for coreference resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1416–1426, Beijing, China, July. Association for Computational Linguistics.

Sam Wiseman, Alexander M. Rush, and Stuart M. Shieber. 2016. Learning global features for coreference resolution. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 994–1004, San Diego, California, June. Association for Computational Linguistics.

Zhilin Yang, Ruslan Salakhutdinov, and William W Cohen. 2017. Transfer learning for sequence tagging with hierarchical recurrent networks. *ICLR*.

Joey Tianyi Zhou, Hao Zhang, Di Jin, Hongyuan Zhu, Meng Fang, Rick Siow Mong Goh, and Kenneth Kwok. 2019. Dual adversarial neural transfer for low-resource named entity recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3461–3471, Florence, Italy, July. Association for Computational Linguistics.