Constrained Multi-Task Learning for Bridging Resolution

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

Abstract

We examine the extent to which supervised bridging resolvers can be improved without employing additional labeled bridging data by proposing a novel constrained multi-task learning framework for bridging resolution, within which we (1) design cross-task consistency constraints to guide the learning process; (2) pretrain the entity coreference model in the multi-task framework on the large amount of publicly available coreference data; and (3) integrate prior knowledge encoded in rule-based resolvers. Our approach achieves state-of-the-art results on three standard evaluation corpora.

1 Introduction

Bridging (Clark, 1975) plays an important role in establishing entity coherence in a text. In contrast to *direct anaphors*, which indicate the coreference relation between a nominal expression and its antecedent, *bridging anaphors* or *associative anaphors* link to their antecedents via non-identical relations. *Bridging resolution* is the task of recognizing and resolving bridging anaphors in a text.

Bridging resolution and coreference resolution are closely related to *Information Status* (IS henceforth) classification, the goal of which is to assign an IS to each discourse entity that indicates how these entities are referred to in a text (Prince, 1981; Nissim et al., 2004; Markert et al., 2012). In general, an entity is *old* if it is coreferent with an entity that has been mentioned before (e.g., "[The business]" and "[its]" in Figure 1). *Bridging anaphors* are discourse-new but hearer-old. They have not been introduced in the discourse directly, but are inferrable from previously mentioned entities (e.g., "[the customers]" in Figure 1). *New* entities are introduced into the discourse for the first time and are not known to the hearer before (e.g. "[The Bakersfield Supermarket]" in Figure 1).

Progress on bridging resolution research is limited in part by the scarcity of annotated training data. While one of the largest annotated entity coreference resolution datasets, OntoNotes, is composed of 2802 English documents in its training split, the two most commonly used English corpora for bridging resolution research, ISNotes (Markert et al., 2012) and BASHI (Rösiger, 2018), are composed of 50 WSJ documents each. Perhaps the most straightforward way to mitigate this data scarcity problem is to combine existing annotated bridging datasets to create a larger training set (Yu and Poesio, 2020). While it makes sense to combine corpora that are created using the same annotation guidelines (e.g., ISNotes and BASHI), attempting to combine corpora created using different guidelines (e.g., ARRAU (Poesio and Artstein, 2008) and ISNotes) will likely confuse the learner, thus limiting the applicability of this method. Some researchers have instead attempted to create *automatically* labeled data via lexico-syntactic patterns (Hou, 2018) and distant supervision (Hou, 2020), but a manual analysis of the resulting data instances reveals that they may be too noisy for training: on average only one-fourth of them are correctly labeled (Hou, 2020).

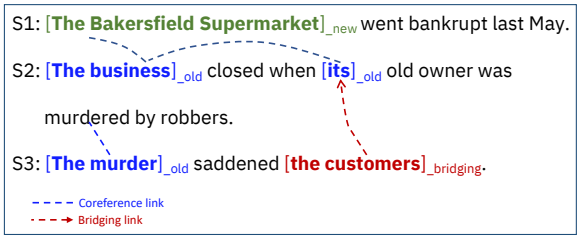

Figure 1: Illustration of information status, bridging and coreference. Example is from Yu and Poesio (2020).

By contrast, we aim to investigate the extent to which supervised bridging resolvers can be improved *without* increasing the amount of labeled bridging data. To this end, we begin by proposing a novel constrained multi-task learning (MTL) framework for bridging resolution. While Yu and Poesio (2020) develop a standard MTL model for

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 759 - 770

bridging resolution and use coreference resolution as the only auxiliary task, we propose to (1) exploit the close connection between IS and bridging/coreference resolution by introducing IS classification as the third task into the MTL framework and (2) guide the learning process by designing cross-task consistency constraints. For instance, in Figure 1, the prediction from the coreference resolution module indicating that both "[The business]" and "[The murder]" are *old* entities can help the bridging resolution module to avoid misclassifying these two mentions as bridging anaphors. Similarly, if the IS classification module predicts "[the customers]" as a bridging anaphor, then the bridging resolution module should find an antecedent for it. We hypothesize that such constraints can guide the training of a complex model to produce a more coherent output across different tasks, thereby improving bridging resolution performance.

While the cross-task consistency constraints could improve performance, they could also hurt performance. Returning to our example in Figure 1, if the IS classification module misclassifies "[the customers]" as non-bridging, the constraints will propagate this error to the bridging resolution module, causing it *not* to resolve the mention. To address this problem, we (1) formulate these constraints as *soft* rather than hard constraints, and (2) improve entity coreference resolution performance by leveraging the large amount of publicly-available coreference-annotated data in OntoNotes to *pre-train* the coreference module.

Finally, since previous work (Hou et al., 2014; Roesiger et al., 2018) has shown that manually defined rules based on various syntactic and semantic properties are valuable to recognize and resolve bridging anaphors, we integrate such prior knowledge about bridging into our MTL framework. Note that the only *hybrid* rule-based and learning-based approach to bridging resolution (Kobayashi and Ng, 2021) merely applies the rule-based resolver and the learning-based resolver in a sequential manner, without combining them into a single model.

In sum, our contributions are two-fold. First, we propose a novel constrained MTL framework that jointly learns three tasks, bridging resolution, coreference resolution, and IS classification, via the use of soft cross-task consistency constraints, prior knowledge provided by rule-based approaches, and pre-training on coreference data. Second, experimental results demonstrate that our framework achieves new state-of-the-art results for full bridging resolution on three datasets (ISNotes, BASHI, and ARRAU).

The rest of the paper is structured as follows. Section 2 describes related work on bridging resolution and constrained multi-task learning with deep neural networks. Section 3 describes our model, including our multi-task framework for jointly learning IS classification, entity coreference resolution and bridging resolution, our cross-task consistency constraints, and how we integrate rule knowledge into the framework. We present evaluation results in Section 4 and our conclusions in Section 5.

2 Related Work

Bridging resolution. Bridging resolution is composed two sub-tasks: *bridging anaphora recognition* and *antecedent selection*. Most previous work tackles them separately. One line of research models bridging recognition as part of IS classification (Rahman and Ng, 2011; Markert et al., 2012; Cahill and Riester, 2012; Rahman and Ng, 2012; Hou, 2021), while others have focused on antecedent selection based on gold bridging anaphors (Poesio et al., 2004; Lassalle and Denis, 2011; Hou et al., 2013; Hou, 2020).

There are a few studies tackling the challenging task of full bridging resolution (i.e., bridging anaphor recognition *and* resolution). Hou et al. (2014) and Roesiger et al. (2018) develop rules to identify bridging links based on syntactic and semantic constraints. Hou et al. (2018) propose a pipeline system built on top of complex manually designed features. Yu and Poesio (2020) design a MTL neural model for bridging resolution that uses coreference resolution as an auxiliary task. Recently, Kobayashi and Ng (2021) show the effectiveness of a hybrid rule-based and MTL approach for bridging resolution. For a detailed overview of these approaches, we refer the reader to a recent survey by Kobayashi and Ng (2020).

Constrained multi-task learning with deep neural networks. Multi-task learning has been widely adopted in various NLP applications to improve the performance of individual tasks (Ruder, 2017). Recently, several studies have demonstrated that multi-task training in neural networks can be further improved by integrating logical constraints to enforce a coherent output across different tasks (Li et al., 2019; Wang et al., 2020; Lu and Ng,

2021). However, for a complex task like bridging resolution, it is non-trivial to choose auxiliary tasks and model the relationships between these tasks in deep neural networks. In this work, we (1) jointly train three tasks (i.e., bridging resolution, coreference resolution, and IS classification); (2) design five soft cross-task consistency constraints to guide the training process; and (3) integrate prior knowledge about bridging into our MTL model.

3 Model

In this section, we present our constrained MTL framework for bridging resolution. Inspired by Yu and Poesio's (2020) span-based model for bridging resolution, which employs an *unconstrained* MTL framework that jointly learns bridging and coreference, our model takes as input a document D represented as a sequence of word tokens and gold mentions M, from which we create span representations. Our model simultaneously learns three tasks, namely IS classification, bridging, and coreference, as defined below.

The IS classification task aims to assign each span i an IS y_{is} taken from an IS inventory. The model predicts the IS of i to be $y_{is}^* = \arg\max_{y_{is}} s_{is}(i, y_{is})$, where s_{is} is a function suggesting i's likelihood of having y_{is} as its IS.

The bridging resolution task involves determining an antecedent for each bridging anaphor. Formally, it assigns span i an antecedent y_b, where $y_b \in \mathcal{Y}(i) = \{1, ..., i-1, \epsilon\}$. In other words, the value of each y_b is the id of its antecedent, which can be one of the preceding spans or a dummy antecedent ϵ (if the mention underlying i is not a bridging anaphor) in the associated document. We define the following scoring function:

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

where $s_a(i,j)$ is a pairwise bridging score computed over i and a preceding span j. The model predicts the antecedent of i to be $y_b^* = \arg\max_{y_b \in \mathcal{Y}(i)} s_b(i, y_b)$.

The entity coreference resolution task involves determining an antecedent for each identity anaphor. Formally, it aims to assign span i an antecedent y_c based on a scoring function s_c that can be defined in an analogous manner as the s_b function in the bridging resolution task.

3.1 Model Structure

Figure 2 shows the structure of our constrained MTL framework. Below we describe the details.

Span Representation Layer Following Yu and Poesio (2020), we use BERT embeddings as the input to a bidirectional LSTM (Hochreiter and Schmidhuber, 1997) to encode tokens and their contexts. Then, we set $\mathbf{g}_i$, the representation of span i, to $[\mathbf{h}_{start(i)}; \mathbf{h}_{end(i)}; \mathbf{h}_{head(i)}; \mathbf{f}_i]$, where $\mathbf{h}_{start(i)}$ and $\mathbf{h}_{end(i)}$ are the hidden vectors of the start and end tokens of i, $\mathbf{h}_{head(i)}$ is an attention-based head vector and $\mathbf{f}_i$ is a span width feature embedding.[1]

IS Prediction Layer For each span i, we pass its representation $\mathbf{g}_i$ to FFNN_{is}, a standard feed-forward neural network. FFNN_{is} outputs a vector $\mathbf{oi}_i$ of dimension of S, where S is the number of possible IS labels. Specifically:

$$\mathbf{oi}_i = \text{FFNN}_{is}(\mathbf{g}_i) \quad (2)$$

$$s_{is}(i, y_{is}) = \mathbf{oi}_i(y_{is}) \quad (3)$$

where $\mathbf{oi}_i(y_{is})$, the y_{is}-th element of $\mathbf{oi}_i$, is a score that indicates i's likelihood of belonging to IS y_{is}. This score is then used to compute s_{is}.

Bridging Prediction Layer To predict bridging links, we define the pairwise score between span i and span j as follows:

$$s_a(i,j) = \text{FFNN}_b([\mathbf{g}_i; \mathbf{g}_j; \mathbf{g}_i \circ \mathbf{g}_j; \mathbf{u}_{ij}]) \quad (4)$$

where $\circ$ denotes element-wise multiplication, $\mathbf{g}_i \circ \mathbf{g}_j$ encodes the similarity between span i and span j, $\mathbf{u}_{ij}$ is a feature embedding encoding the distance between two spans[1], and FFNN_b is the FFNN used in the bridging prediction layer. This pairwise score is then used to compute s_b (see Equation (1)).

Coreference Prediction Layer The coreference prediction layer is defined in the same way as the bridging prediction layer, with the coreference pairwise score $s_c(i,j)$ between two spans i and j computed by another FFNN, FFNN_c. Note that the first few layers of FFNN_c and FFNN_b are shared.

3.2 Incorporating Consistency Constraints

As noted before, we propose to guide the learning process by incorporating consistency constraints on the three tasks involved in our model. Below we design five cross-task consistency constraints and show how they can be incorporated into our model in a *soft* manner.

[1]This feature embedding is originally proposed by Clark and Manning (2016). See their paper for details.

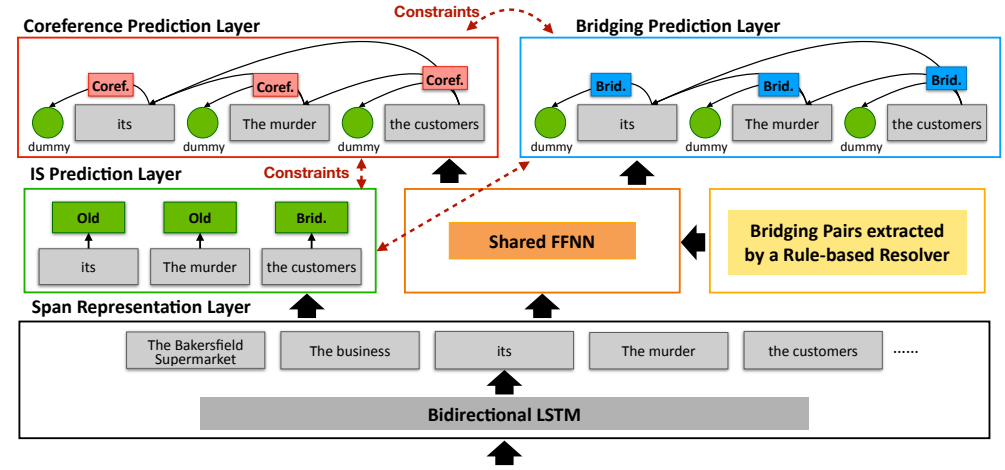

Figure 2: Model structure of the constrained MTL framework for bridging resolution.

Constraint P1: If a span i has BRIDGING as its IS value, then its bridging antecedent must not be the dummy antecedent.

To enforce **P1** in a soft manner in our model, we define a penalty function p_1, which imposes a penalty on span i if it violates the constraint, as shown below:

$$p_1(i) = \begin{cases} 0 & \arg\max_{y_{is} \in \mathcal{Y}} s_{\text{is}}(i, y_{is}) \neq \text{brid} \\ s_{\text{is}}(i, \text{brid}) - \max_{y_{is} \in \mathcal{Y} \setminus \{\text{brid}\}} s_{\text{is}}(i, y_{is}) & \text{otherwise} \end{cases} \quad (5)$$

where $\mathcal{Y}$ is the set of possible IS labels. Intuitively, p_1 estimates the minimum amount that needs to be adjusted so that span i's IS type is not BRIDGING. In particular, p_1 returns 0 (i.e., no penalty) if i's IS type is not BRIDGING.

We incorporate p_1 into the model as a penalty term in s_b (Equation (1)). Specifically, we redefine $s_b(i, j)$ when $j = \epsilon$, as shown below:

$$s_b(i, \epsilon) = s_b(i, \epsilon) - \gamma_1 p_1(i) \quad (6)$$

where γ_1 is a positive constant that controls the hardness of the constraint. The smaller γ_1 is, the softer the constraint is. Intuitively, if **P1** is violated, $s_b(i, \epsilon)$ will be lowered by the penalty term, and the dummy antecedent will less likely be selected as the antecedent of i.

Constraint P2: If a span i has OLD as its IS value, then its coreference antecedent must not be the dummy antecedent.

The penalty function p_2 used to enforce **P2** is formulated in the same way as **P1**.

Constraint P3: If the IS task predicts a span i as non-BRIDGING, then its antecedent selected in the bridging task must be the dummy antecedent.

Similar to **P1**, we define a penalty function p_3 to enforce **P3**:

$$p_3(i) = \begin{cases} 0 & \arg\max_{y \in \mathcal{Y}} s_{\text{is}}(i, y) = \text{brid} \\ \max_{y \in \mathcal{Y} \setminus \{\text{brid}\}} s_{\text{is}}(i, y) - s_{\text{is}}(i, \text{brid}) & \text{otherwise} \end{cases} \quad (7)$$

We employ p_3 to update s_b as follows:

$$s_b(i, j) = s_b(i, j) - \gamma_3 p_3(i) \quad (8)$$

where γ_3, like γ_1, is the hardness coefficient. This penalty is applied only when **P3** is violated. Specifically, if IS task predicts a span i as non-BRIDGING but its antecedent selected in the bridging task is not the dummy antecedent, then the penalty term will lower the s_b score for each of i's non-dummy antecedents, which in turn makes it more likely for the dummy antecedent to be selected as the antecedent of i.

Constraint P4: If a span i does not have OLD as its IS value, then its coreference antecedent must be the dummy antecedent.

The penalty function p_4 used to enforce **P4** is formulated in the same way as **P3**.

Constraint P5: If a span i has a non-dummy antecedent as its coreference antecedent, then its bridging antecedent must be the dummy antecedent.

The penalty function p_5 used to enforce **P5** is defined as follows:

$$p_5(i) = \begin{cases} 0 & \arg\max_{j \in \mathcal{Y}(i)} s_{\text{c}}(i, j) = \epsilon \\ \max_{j \in \mathcal{Y}(i) \setminus \{\epsilon\}} s_{\text{c}}(i, j) & \text{otherwise} \end{cases} \quad (9)$$

where $\mathcal{Y}(i)$ is the set of candidate antecedents of span i. We employ p_5 to update s_b as follows:

$$s_b(i,j) = s_b(i,j) - \gamma_5 p_5(i) \quad (10)$$

where γ_5 is the hardness coefficient.

3.3 Incorporating Prior Knowledge

Next, we incorporate the prior knowledge provided by rule-based resolvers into our model. Specifically, we employ the set of corpus-specific rules designed by Rösiger et al. (2018). Recall that the output of a rule-based bridging resolver is a set of links between a bridging anaphor and one of its antecedents. We incorporate these bridging links into our model by encoding them as a binary feature, $\mathbf{r}_{ij}$, whose value is 1 if and only if the rule-based resolver posits a bridging link between span i and span j. This feature will be used as an additional feature for FFNN_b and FFNN_c.

As noted by Rösiger et al. (2018), rule-based resolvers are precision- rather than recall-oriented. The reason is that these hand-crafted rules are designed to resolve specific (rather than all) categories of bridging anaphors. For instance, one rule is designed to resolve a building part (e.g., "the door") to the building of which it is a part (e.g., "the house"). Because of the low-recall nature of rule-based resolvers, the feature $\mathbf{r}_{ij}$, which we compute based on the rule-based outputs, could be perceived as not particularly useful by our model. Consequently, to encourage the model to seriously take into consideration the potentially useful information encoded in $\mathbf{r}_{ij}$, we design a rule loss (see Section 3.4), which imposes a penalty on the model during *training* if the antecedent selected by the model is a non-dummy antecedent that is neither a correct antecedent of i nor the one selected by the rules (as encoded in $\mathbf{r}_{ij}$).

3.4 Training

The loss function, $L(\Theta)$, consists of the losses of the three tasks and the rule loss as follows:

$$L(\Theta) = \sum_{i=1}^{d} (\lambda_b L_b + \lambda_c L_c + \lambda_{is} L_{is} + \lambda_r L_r) \quad (11)$$

where d is the number of training documents and the hyperparameters (i.e., the λ's), which determine the trade-off between the task losses, are tuned using grid search to maximize the average resolution F-scores on development data.

Task Losses We employ a max-margin loss for the bridging and coreference resolution tasks.

Defining the bridging loss is tricky since the antecedents for each bridging anaphor are evaluated in the form of coreference clusters. We adopt the entity coreference loss function originally defined by Wiseman et al. (2015). Specifically, let $\text{GOLD}_b(i)$ denote the set consisting of span i's bridging antecedent as well as the spans preceding i that are coreferent with the antecedent, and y_b^l be $\arg\max_{y \in \text{GOLD}_b(i)} s_b(i,y)$. In other words, y_b^l is the highest scoring (latent) antecedent of i according to s_b among all the antecedents of i.

The loss function for bridging is defined as:

$$L_b(\Theta) = \sum_{i=1}^{n} \max_{j \in \mathcal{Y}(i)} (\Delta_b(i,j)(1 + s_b(i,j) - s_b(i,y_b^l))) \quad (12)$$

where $\Delta_b(i,j)$ is a mistake-specific cost function that returns the cost associated with a particular type of error if an error exists and 0 otherwise (Durrett and Klein, 2013).[2] Intuitively, the loss function penalizes a span i if the predicted antecedent j has a higher score than the correct latent antecedent y_b^l.

The task loss for coreference, L_c, is defined in the same way as the bridging loss, having an analogous mistake-driven cost function $\Delta_c(i,j)$.[3]

The task loss for the IS prediction task, L_{is}, is the weighted softmax cross entropy loss, where misclassified bridging mentions and non-bridging mentions are weighted according to a mistake-driven cost function $\Delta_{is}(i,j)$.[4]

The rule loss is motivated by the bridging loss. Specifically, the model will be penalized if there exists an incorrect non-dummy candidate antecedent whose s_b score is higher than the score of the antecedent chosen by the rules, as shown below:

$$L_r(\Theta) = \sum_{i \in N'} \max_{j \in \mathcal{Y}(i) \setminus \epsilon} (\Delta_r(i,j)(1 + s_b(i,j) - s_b(i,y_r))) \quad (13)$$

where N' is the set of candidate anaphors for which the rule-based system found a (non-dummy) an-

[2] In $\Delta_b(i,j)$, there are three error types: (1) false link (incorrectly resolved anaphoric mentions); (2) false new (anaphoric mentions misclassified as non-anaphoric); and (3) wrong link (non-anaphoric mentions misclassified as anaphoric). We use hyperparameters α_{b1}, α_{b2}, and α_{b3} to determine their trade-offs.

[3] In $\Delta_c(i,j)$, the error types are the same as those in $\Delta_b(i,j)$. We use hyperparameters α_{c1}, α_{c2}, and α_{c3} to determine their trade-offs.

[4] In $\Delta_{is}(i,j)$, there are two error types: (1) false new (bridging mentions misclassified as non-bridging); and (2) false bridging (non-bridging mentions misclassified as bridging). We use hyperparameters α_{is1} and α_{is2} to determine their trade-offs.

tecedent, y_r is the antecedent selected by the rules, and $\Delta_r(i, j)$ is an indicator function that returns 0 if j is the correct antecedent and 1 otherwise.

3.5 Pre-Training

As mentioned in the introduction, we pre-train the coreference module in our MTL framework on the English portion of OntoNotes 5.0[5], excluding those documents that appear in ISNotes or BASHI. To do so, we pre-train the full model shown in Figure 2, setting λ_b to 1 and the remaining λ's to 0 in the loss function so that only the network weights associated with the coreference module will be updated. Note that we follow Yu and Poesio (2020) and use the softmax cross entropy loss rather than the max-margin loss for L_b during pre-training, the reason being that this could simplify pre-training by obviating the need to tune the hyperparameters associated with the mistake-specific cost functions.

4 Evaluation

4.1 Experimental Setup

4.1.1 Corpora

We use three English corpora that are arguably the most widely used corpora for bridging evaluation, namely ISNotes (composed of 50 WSJ articles in OntoNotes) (Markert et al., 2012) , BASHI (The Bridging Anaphors Hand-annotated Inventory, composed of another 50 WSJ articles in OntoNotes) (Rösiger, 2018), and ARRAU (composed of articles from four domains, RST, GNOME, PEAR, and TRAINS) (Poesio and Artstein, 2008; Uryupina et al., 2020). Following previous work, we report results only on RST, the most comprehensively annotated segment of ARRAU. Table 1 shows the statistics on these corpora.

For ARRAU RST, we use the standard train-test split. For ISNotes and BASHI, we divide the documents in each corpus into 10 folds (8 folds for training, 1 fold for development, and 1 fold for testing) and report 10-fold cross-validation results.

4.1.2 Evaluation Setting

Following previous work (Hou et al., 2014; Roesiger et al., 2018), we report results for *full bridging resolution* based on gold mentions. In this setting, a system is given as input both a document and its the *gold* mentions. The goal is to identify bridging anaphors from the gold mentions and resolve them to their antecedents, which are also chosen from the gold mentions.

[5] https://catalog.ldc.upenn.edu/LDC2013T19

Corpora	Docs	Tokens	Mentions	Anaphors
ISNotes	50	40,292	11,272	663
BASHI	50	57,709	18,561	459
ARRAU RST	413	228,901	72,013	3,777

Table 1: Statistics on different corpora.

There is a caveat in this evaluation setting, however. In ISNotes and BASHI, some bridging antecedents correspond to *events* (see Example (4) in Table 5), and previous studies differ in terms of how event antecedents should be handled. The reason is that while these event antecedents are annotated, they are *not* annotated as gold mentions. When reporting results on resolving gold mentions, some previous work (e.g., Hou et al. (2014), Hou et al. (2018)) chose not to include these event antecedents in the list of candidate antecedents and others (e.g., Roesiger et al. (2018), Yu and Poesio (2020)) did. Obviously, the setting in which gold event antecedents are not included in training/evaluation is harsher because it implies that anaphors with event antecedents will always be resolved incorrectly. We believe that including gold event antecedents during evaluation does not represent a realistic setting, and will only report results using the "harsh" setting in this paper.

4.1.3 Evaluation Metrics

Following Yu and Poesio (2020), we report results for bridging recognition and resolution in terms of precision (P), recall (R), and F-score (F). For recognition, recall is the fraction of gold bridging anaphors that are correctly identified, whereas precision is the fraction of bridging anaphors identified by the system that is correct. For resolution, recall and precision are defined in a similar fashion. In addition, we report IS classification results in terms of accuracy and coreference results in terms of CoNLL score (Pradhan et al., 2014), which is the unweighted average of the F-scores provided by three metrics, MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), and $CEAF_e$ (Luo, 2005).

4.1.4 Implementation Details

To train the neural models in our experiments, we use ADAM (Kingma and Ba, 2014) as the optimizer and set all model parameters that originated in Yu and Poesio's (2020) model to the same values as those reported in their paper. Each model is trained for up to 150 epochs in ISNotes and BASHI

Model	Bridging						IS Classification Accuracy	Coreference Resolution CoNLL
	Recognition			Resolution				
	P	R	F	P	R	F		
ISNotes								
Roesiger et al. (2018)	46.8	17.7	25.6	32.0	12.1	17.5	-	-
Y&P-MTL	51.8	27.2	36.7 (±1.6)	25.3	12.5	17.4 (±1.3)	-	62.6
Hybrid	44.8	35.5	39.6 (±0.2)	24.7	19.6	21.9 (±1.6)	-	62.6
MM-MTL	45.5	41.6	43.4 (±0.8)	21.1	19.3	20.2 (±0.7)	-	64.5
Full model	54.1	48.0	**50.9** (±0.2)	27.6	24.5	**26.0** (±0.0)	78.0	76.3
BASHI								
Roesiger et al. (2018)	33.5	22.9	27.2	17.3	11.8	14.0	-	-
Y&P-MTL	35.7	15.2	21.3 (±1.5)	19.3	8.2	11.5 (±0.8)	-	57.2
Hybrid	32.4	32.3	32.3 (±0.7)	16.3	16.3	16.0 (±0.4)	-	57.2
MM-MTL	37.9	27.7	32.0 (±0.3)	15.6	11.4	13.2 (±0.6)	-	57.0
Full model	40.7	35.3	**37.5** (±0.7)	20.1	17.5	**18.6** (±0.1)	85.3	72.6
ARRAU RST								
Roesiger et al. (2018)	18.3	33.9	23.7	11.7	21.7	15.2	-	-
Y&P-MTL	27.6	23.1	25.2 (±0.3)	20.5	17.2	18.7 (±0.1)	-	55.9
Hybrid	16.8	43.2	24.2 (±0.1)	11.3	29.1	16.3 (±0.1)	-	55.9
Full model	26.1	45.6	**33.2** (±1.2)	17.1	29.8	**21.7** (±0.0)	84.5	61.2

Table 2: Results of different resolvers on bridging resolution and related tasks. Each result is the average of two runs. For each model with a learning component, the recognition and resolution F-scores are accompanied with the corresponding standard deviation scores (in parentheses).

and up to 200 epochs in ARRAU, with early stopping based on the development set.

For our model, we pre-train the coreference model for 15 epochs, and the remaining parameters are chosen jointly using grid search to maximize resolution F-score on development data. Specifically, the weights associated with each task and the rule in the loss function (i.e., the λ_i's) are searched out of $\{0.1, 0.5, 1, 5, 10, 20, 30\}$. The weights associated with the mistake-driven cost functions (i.e., the Δ_i's) are searched out of $\{0.1, 0.5, 1, 5, 10, 15, 20\}$. The hardness coefficients of the consistency constraints (i.e., the γ_i's) are searched out of $\{0.05, 0.1, 0.5, 1, 5, 10, 20, 30\}$.[6]

4.2 Baseline Systems

We employ three baselines. The first one is Rösiger et al.'s (2018) rule-based approach, which consists of rules that are built on top of Hou et al. (2014).[7] The second one, Y&P-MTL, is Yu and Poesio's (2020) MTL system.[8] The third one is the Hybrid rule-based and learning-based system proposed by Kobayashi and Ng (2021) in which the rules are first applied and then Y&P-MTL is used to resolve the remaining bridging anaphors.

4.3 Results and Discussion

Results are shown in Table 2. A few points about the baseline results deserve mention. First, in terms of bridging recognition and resolution performance, the best baselines are Hybrid for both ISNotes and BASHI and Y&P-MTL for ARRAU RST. Hence, these two baselines can be viewed as the prior state of the art. Second, while Rösiger et al.'s rule-based model never achieves the best results on any of the three datasets, it is not always the worst performer: Y&P-MTL is the worst baseline on BASHI in terms of resolution.[9] Third, Hybrid fails to improve the performance of Y&P-MTL in ARRAU RST, meaning that the rules fail to provide additional benefits to Hybrid. This could be attributed to the fact that the rules in ARRAU RST have much lower recognition and resolution precision scores than those in ISNotes and BASHI (Roesiger et al., 2018).

While Y&P-MTL uses undersampling (to reduce the number of negative examples used to train the bridging module) and a likelihood loss, we additionally experiment with a max-margin loss (see Section 3.4) without undersampling in our model. To see how these two changes impact performance, we create another model, MM-MTL, which is simply a max-margin version of Y&P-MTL without

[6] See Appendix A for the final hyperparameters chosen for the full model.

[7] We used the publicly available implementation of these rule-based systems from https://github.com/InaRoesiger/BridgingSystem.

[8] We used their publicly available implementation from https://github.com/juntaoy/dali-bridging.

[9] The baseline results in Table 2 are lower than those reported in the original papers because (1) we report results using the "harsh" setting (see Section 4.1.2); (2) Roesiger et al. (2018) and Kobayashi and Ng (2021) postprocess the system output with gold coreference information, and (3) Yu and Poesio (2020) and Kobayashi and Ng (2021) use additional labeled data for model training.

undersampling. Results on the development set are mixed: while MM-MTL outperforms Y&P-MTL on ISNotes and BASHI, the reverse is true on ARRAU RST. Consequently, we use the max-margin loss without undersampling when training our model on ISNotes and BASHI, but fall back on the likelihood loss with undersampling for ARRAU RST. To better understand the impact of using a max-margin loss with undersampling, we show in Table 2 the test results of MM-MTL. As we can see, MM-MTL outperforms Y&P-MTL by 6.7–10.7% points in F-score for bridging recognition and 1.7–2.8% points in F-score for bridging resolution in ISNotes and BASHI.

The last row of each section of Table 2 shows the results of our full model, which outperforms the best baseline by 5.2–11.3% points in F-score for bridging recognition and 2.6–4.1% points in F-score for bridging resolution. Hence, the full model establishes new state-of-the-art results on these three datasets. For bookkeeping purposes, we also report the scores for each component of our model in terms of IS classification accuracy and coreference CoNLL score.

4.4 Model Ablations

To evaluate the contribution of the different components in our full model, we show in Tables 3 and 4 ablation results on ISNotes, which we obtain by removing one component at a time from the model and retraining it. Note that for coreference we show the anaphor recognition results as they are affected by the consistency constraints.

Consistency constraints. Ablating the consistency constraints means removing all the penalty terms from s_b and s_c. The resulting system resembles a typical multi-task learning setup, where the different tasks only interact via a shared representation. As we can see in Table 3, bridging resolution F-score drops by 1.7% points, coreference recognition F-score drops by 0.5% points, and IS bridging recognition F-score drops by 1.2% points. These results suggest the effectiveness of using consistency constraints in a multi-task setup.

Soft→Hard. Next, we replace soft constraints with hard constraints. Comparing with the results in row 2, bridging resolution F-score drops by 1.2% points. This indicates that having hard constraints is worse than having *no* constraints at all.

Rule loss and feature. Bridging resolution F-score drops by 1.1% points when ablating only the rule loss and by 2.4% points when ablating both the rule loss and the rule feature. These results suggest that the rule feature is useful and that the rule loss enhances the effectiveness of the rule feature.

		Bridging		IS		Coref.
		Recog.	Resol.	Brid.	Old	Recog.
1	Full	**50.9**	**26.0**	48.3	86.7	88.4
2	– Constraints	46.1	24.3	47.1	86.7	87.9
3	Soft→Hard	47.7	23.1	46.4	87.1	88.8
4	– Rule loss	49.8	24.9	47.6	86.6	88.3
5	– Rule loss+feat.	48.1	23.6	46.9	86.5	88.1
6	– Pre-training	49.6	20.2	49.1	86.6	84.5
7	– Coref. task	47.4	22.6	46.3	88.6	-
8	– IS task	44.2	22.7	-	-	87.8

Table 3: Ablation results of the full model.

	Constraints	Bridging		IS		Coref.
		Recog.	Resol.	Brid.	Old	Recog.
1	Full	**50.9**	**26.0**	48.3	86.7	88.4
2	– P1	50.2	24.2	49.6	86.7	88.2
3	– P2	49.1	24.5	48.1	86.5	88.0
4	– P3	48.2	24.8	48.0	86.6	88.0
5	– P4	49.5	24.3	47.9	86.8	88.1
6	– P5	50.0	23.7	48.1	86.6	88.0

Table 4: Ablation results of the full model w.r.t individual soft constraints.

Pre-training. Next, we do not pre-train the coreference component in the multi-task framework. This causes bridging resolution F-score and coreference recognition F-score to drop abruptly by 5.8% points and 3.9% points respectively, suggesting the important role played by pre-training.

Coreference resolution and IS classification tasks. Next, we ablate one of the tasks in the multi-task framework. Bridging resolution F-score drops by 3.4% points when ablating coreference and by 3.3% points when ablating IS classification. These results suggest that both tasks contribute considerably to bridging resolution performance.

Individual soft constraints. Finally, we ablate one soft constraint at a time from the full model. Results are shown in Table 4. Bridging resolution F-score drops by 1.2–2.3% points, suggesting the positive contribution of each soft constraint.

While our discussion of these results has focused on bridging resolution, the same trends can be observed for bridging recognition for the most part. Overall, these results suggest that each component contributes positively to bridging resolution.

4.5 Error Analysis

Although our full model outperforms all previous models for bridging resolution, it is still far from perfect. To better understand what areas of improvement are required, we discuss some common

errors made by our full model in this subsection.

Bridging anaphora recognition errors. Recall errors in bridging anaphora recognition are the result of a system's failure in identifying bridging anaphors. We find that on the three datasets, the highest proportion of the recall errors (57% on ISNotes, 61% on ARRAU, and 82% on BASHI) is due to the fact that a large number of bridging anaphors are misclassified as *new* or *other*[10] mentions in the IS classification module, such as "**income**" in Example (1) in Table 5.

Precision errors in bridging anaphora recognition are the result of a system's misclassification of non-bridging mentions as bridging anaphors. Similar to the recall errors described above, most precision errors are *new* or *other* mentions being misclassified as bridging, which account for 50%, 74% and 82% of the precision errors in ISNotes, ARRAU, and BASHI, respectively. In Example (2), "**service**" is misclassified by both the bridging and IS components as a bridging anaphor.

In general, it seems that our system struggles to distinguish bridging anaphors from generic *new* mentions with simple syntactic structures, an observation that has also been reported in previous work (Hou, 2021; Kobayashi and Ng, 2021). Note that most of these *bridging* or *new* mentions are relational nouns (de Bruin and Scha, 1988). Normally, whether additional implicit arguments are required to interpret such relational nouns depends on the surrounding context. In Example (1), "*the industry*" is necessary to fully understand the meaning of "**income**"; while in Example (2), no additional implicit arguments are required to understand the meaning of "**service**".

Bridging anaphora resolution precision errors. Precision errors in bridging anaphora resolution appear when a system selects the wrong antecedent for a bridging anaphor. A major reason for this error is that our model largely fails to exploit contextual information. In Example (3), the model links the bridging anaphor "*a spokesman*" to the wrong antecedent "[the state]", which is reasonable if one does not look into the context. However, according to the context, the correct antecedent should be "*Gov. Deukmejian*", which requires a system to know that "Gov." is the abbreviation for "Governor" and that normally a governor will have a spokesman.

(1) In 1984, an attempt was made to crack down on *the industry* with tougher restrictions. Then, in 1988, a proposal to keep better track of **income** by selling prepaid cards for pachinko was fielded in parliament.
(2) The Bay Area Rapid Transit system, which runs subway trains beneath the bay, is braced for a doubling of its daily regular ridership to 300,000. BART has increased **service** to 24 hours a day in preparation for the onslaught.
(3) Both Mr.Brown, the state's most influential legislator, and *Gov. Deukmejian* favor a temporary sales tax increase – should more money be needed than [the state] can raise from existing sources and the federal government. According to **a spokesman**, the governor is also studying the possibility of raising state gasoline taxes.
(4) ... the drug still *lacks* federal approval for use in the youngest patients. As **a result**, many youngsters have been unable to obtain the drug ...

Table 5: Examples of the errors made by our full model.

In addition, on ISNotes, 6% of the bridging anaphors have a non-mention antecedent (see "**a result**" in Example (4)) and 12% of the bridging anaphors have antecedents that are more than five sentences away. Currently our system does not handle these difficult cases.

5 Conclusion

We proposed the first neural model for full bridging resolution that (1) exploits the connection between information status classification, entity coreference resolution, and bridging resolution in a multi-task learning framework, (2) employs soft cross-task consistency constraints to guide the learning process, (3) pre-trains the entity coreference model, and (4) integrates prior knowledge encoded in handcrafted bridging resolution rules into the learning framework. Our model outperformed several strong baselines and achieved state-of-the-art results on three evaluation datasets. Ablation results provided suggestive evidence that each component of our model contributed positively to bridging resolution performance.

Acknowledgments

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

[10] Unlike ISNotes and ARRAU, BASHI does not have IS annotations. We use heuristics to derive four IS types: *old*, *mediated/bridging*, *mediated/comparative* and *other*. A mention's IS is *other* if it is not annotated as *mediated* and is not coreferent with any previous mentions.

References

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

Aoife Cahill and Arndt Riester. 2012. Automatically acquiring fine-grained information status distinctions in German. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 232–236, Seoul, South Korea. Association for Computational Linguistics.

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

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. Association for Computational Linguistics.

Jos de Bruin and Remko Scha. 1988. The interpretation of relational nouns. In *26th Annual Meeting of the Association for Computational Linguistics*, pages 25–32, Buffalo, New York, USA. Association for Computational Linguistics.

Greg Durrett and Dan Klein. 2013. Easy victories and uphill battles in coreference resolution. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1971–1982, Seattle, Washington, USA. Association for Computational Linguistics.

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

Yufang Hou. 2018. 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. 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.

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

Yufang Hou, Katja Markert, and Michael Strube. 2013. 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. 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*, pages 2082–2093.

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

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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

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

Emmanuel Lassalle and Pascal Denis. 2011. Leveraging different meronym discovery methods for bridging resolution in French. In *Proceedings of the 8th International Conference on Anaphora Processing and Applications*, DAARC'11, page 35–46, Berlin, Heidelberg. Springer-Verlag.

Tao Li, Vivek Gupta, Maitrey Mehta, and Vivek Srikumar. 2019. A logic-driven framework for consistency of neural models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3924–3935, Hong Kong, China. Association for Computational Linguistics.

Jing Lu and Vincent Ng. 2021. Constrained multi-task learning for event coreference resolution. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4504–4514, Online. Association for Computational Linguistics.

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

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

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

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

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

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

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

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

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

Ina 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. Association for Computational Linguistics.

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

Ina Rösiger, Maximilian Köper, Kim Anh Nguyen, and Sabine Schulte im Walde. 2018. Integrating predictions from neural-network relation classifiers into coreference and bridging resolution. In *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 44–49, New Orleans, Louisiana. Association for Computational Linguistics.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

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

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

Haoyu Wang, Muhao Chen, Hongming Zhang, and Dan Roth. 2020. Joint constrained learning for event-event relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 696–706, Online. Association for Computational Linguistics.

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. Association for Computational Linguistics.

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

A Final Hyperparameters and Computing Environment

We conduct our experiments using a NVIDIA QUADRO RTX 6000. The estimated GPU hour per model in this paper is approximately 6 hours on average. Table 6 shows the final hyperparameters for our full model on the three datasets.

Parameter Source	Parameters	ISNotes	BASHI	ARRAU RST
Loss function	$\lambda_b, \lambda_c, \lambda_{is}, \lambda_r$	1.0, 1.0, 1.0, 0.5	1.0, 1.0, 1.0, 20.0	1.0, 1.0, 1.0, 20.0
Δ_b	$\alpha_{b1}, \alpha_{b2}, \alpha_{b3}$	0.1, 5.0, 5.0	0.1, 10.0, 10.0	0.1, 5.0, 5.0
Δ_c	$\alpha_{c1}, \alpha_{c2}, \alpha_{c3}$	1.0, 1.0, 1.0	1.0, 1.0, 1.0	1.0, 1.0, 1.0
Δ_{is}	$\alpha_{is1}, \alpha_{is2}$	10.0, 1.0	10.0, 1.0	10.0, 1.0
Constraints	$\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5$	0.5, 0.05, 0.5, 0.05, 1.0	1.0, 0.05, 0.5, 0.05, 1.0	0.5, 0.05, 0.5, 0.05, 1.0

Table 6: Final hyperparameters for the full model.