A Multi-task Learning Model for Gold-two-mention Co-reference Resolution

Ruicheng Liu[1], Guanyi Chen[2], Rui Mao[1], Erik Cambria[1]
[1] *Nanyang Technological University, Singapore*
[2] *Utrecht University, Netherlands*
ruicheng001@e.ntu.edu.sg, g.chen@uu.nl, {rui.mao, cambria}@ntu.edu.sg

***Abstract*—The task of resolving repeated objects in natural languages is known as co-reference resolution. It is an important part of modern natural language processing and semantic cognition as these implicit relationships are particularly difficult in natural language understanding in downstream tasks. Mention identification and mention linking are the two sub-tasks in the general co-reference resolution research community. Gold-two-mention style co-reference resolution is a special type of co-reference resolution that focuses on linking the ambiguous pronoun to one of the two candidate antecedents. In this paper, we proposed a joint learning model that learns mention identification and mention linking tasks together, because we find that the learning of mention identification can provide supportive dependent information for the learning of mention linking. As far as we know, we propose the first model that introduces a multi-task learning framework to the gold-two-mention co-reference resolution task. We find that our proposed model outperforms state-of-the-art baselines and a single-task learning model on three gold-two-mention co-reference resolution datasets. By comparing the errors made by either the single-task learning model or the multi-task learning model, our error analysis also yields interesting findings about in which way our multi-task learning model makes fewer resolution errors.**

***Index Terms*—Co-reference Resolution, Natural Language Processing, Linguistics, Deep Learning**

I. Introduction

To achieve coherence within natural language understanding, it is necessary to have a firm grasp of the argumentation structure and information flow. Co-reference resolution (CR) is among these parsing attempts and refers to the process of resolving any spans in a context that point to the same physical object or event. CR can be broken into two sub-tasks, namely mention identification, i.e., recognizing potential mentions in the previous discourse, and mention linking, i.e., linking the mentions that are co-referential.

Recently, there has been a line of work focusing on "hard" cases of CR [1], the resolution of which often requires either reasoning from the discourse or external lexical and commonsense knowledge. Given examples:

(1) a. My cat only eats canned food because **it** is very picky.
b. My cat only eats canned food because **it** is very tasty.

The pronoun 'it' refers to different antecedents, while only one token is different in these two sentences (i.e., 'picky' and 'tasty').

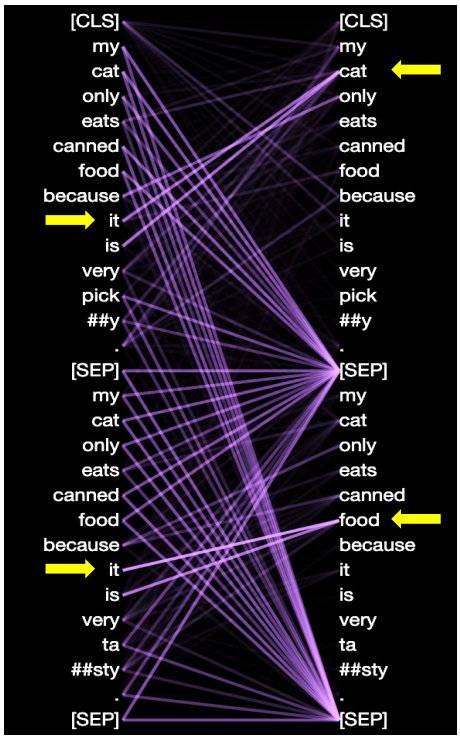


Fig. 1: Transformer attention weight visualization. Brighter color denotes higher weights (more attention).

Building on this, Levesque et al. introduced the Winograd Scheme Challenge (WSC) in which there are hundreds of such minimal pairs and which is designed to assess CR models' ability to link mentions in these pairs. Following this scheme, extensions have been done by enlarging the dataset [2] or by looking at other kinds of "hard" cases that link to gender bias [3], [4]. Since WSC-like CR tasks always provide the gold-standard position of each mention and always ask models to choose between two antecedents, tasks as such are often called Gold-two-mention CR (henceforth, GTM-CR).

In order to equip CR models with world knowledge, GTM-CR models often rely on pre-trained language models (PLMs). For example, Joshi et al. [5] and Attree [6] used BERT [7], while, most recently, Kocijan et al. [8] used a BERT that was re-trained on Wikicrem, a very large Wikipedia dataset where mentions are masked.

Here, we argue that, though GTM-CR models do not need to identify mentions, the ability to distinguish mentions from their contexts is still helpful for a neural network model, as it might enable the model to acquire more positional information that is related to CR. This could include information about the grammatical role, syntactic parallelism (i.e., whether the target referent is in the same syntactic position as its antecedent), and many others, which have been proved to be essential for CR [9], [10]. As a matter of fact, jointly modeling identification and linking has been deployed in the classic CR task. For instance, Daume and Marcu [11] organized the two tasks into a single search space and tackled CR as a searching problem. Lee et al. [12] modeled mention identification and linking as a single end-to-end task. These embody the superiority of joint modeling mention identification and linking. However, the searching-based paradigm [11] and single-task learning (STL) paradigm [12] cannot share useful dependency information that is learned from different tasks and training objectives.

Taking the setting of GTM-CR into consideration, we follow a different paradigm: We propose to train a GTM-CR model by jointly identifying mentions and linking them by using Multi-Task Learning (MTL). Concretely, we develop an MTL model to learn the mention identification and linking tasks together with a shared encoder and task-specific towers upon the shared encoder. The shared encoder aims at learning the general dependency relationship of an input sequence. The task-specific towers aim at learning the dependencies of mention identification and linking, respectively. The sharing encoder and task-specific towers are Transformer [13]-based.

Given the fact that Transformer encodes features via a multi-head attention mechanism and a feed-forward layer, the multi-head attention can learn different dependency features from different tasks. The visualization of the attention in our model in Fig. 1 further demonstrates the need of mention identification: learning mention identification likely allows the attention head of a referent ('it') to attend to its dependent antecedent ('cat') in Example (1-a), while the attention head of 'it' attends to 'food' in Example (1-b). This is a strong signal for the learning of mention linking, e.g., linking 'it' with 'cat' for Example (1-a), and linking 'it' with 'food' for Example (1-b). Simply learning mention linking may miss such a helpful signal.

Additionally, to better balance the training of the two tasks of MTL, we propose a dynamic weight-balancing mechanism. During training, it dynamically adapts the weight for each task with respect to the ratio of the loss of the two tasks.

We examine our MTL model on three GTM-CR datasets: GAP [3], DPR [2], and Winogender [4]. We find that our model outperforms state-of-the-art (SOTA) baselines for GAP in terms of both F1 score and Bias (NB: GAP and Winogender were designed to assess bias in CR models). It also exceeds the SOTA baseline for Winogender and achieves a comparable result as the SOTA baseline for DPR even though the fine-tuning dataset we used is 99.95% smaller than that of the SOTA model. We further compare our MTL model with its STL alternative and observe that the MTL model exceeded the STL model on all these three datasets.

Finally, we conduct an error analysis, yielding interesting findings. For example, MTL dramatically helps CR systems to be less likely to make false predictions for feminine pronouns.

The contribution of this paper has three-fold:

- We propose an MTL learning paradigm for GTM-CR. The new learning paradigm exceeds previous learning paradigms on three datasets;
- We introduce a dynamic weight balancing mechanism which allows our multi-task learning co-reference resolver balance between mention identification and mention linking dynamically;
- We conduct a series of insight analyses for investigating the effects of dynamic weight balancing and multi-task learning for CTM-CR and analyzing what kind of error our model is more capable of overcoming.

II. Related Work

CR is considered as one of the most difficult tasks in natural language understanding. It is important for downstream natural language processing activities such as entity linking [14], named entity recognition [15], and sentiment analysis [16]. It also has strong connections in referring expression generations [17], [18], [19], [20]. In this section, we review CR models that use PLMs and the GTM-CR Models and introduce the motivations of this work.

A. Co-reference Resolution with Pre-trained Language Models

Approaches for CR can be break into four categories [21]: feature-based [22], recurrent neural network-based [23], knowledge-based [24] and Transformer-based [5], [25] approaches. Despite the fact that there are no absolute boundaries between their timelines, we can roughly conclude that over the years, the research interest has shifted from feature-based traditional machine learning models to deep learning models that rely on multilayer perception or recurrent neural networks, and then to Transformer-based large scale PLMs. According to the most recent survey by Liu et al. [21], the SOTA models on 18 CR datasets respectively are all Transformer-based models. This shows the strength of large-scale PLM-based models. Due to page limits, we would like to refer readers to [21] for the comprehensive list of models under these four categories, and their performances under different datasets.

The work of Joshi et al. [5] is the first one that incorporates PLMs into CR models. It is based on a coarse-to-fine co-reference model presented by Lee et al. [26], coined c2f-coref. The LSTM-based encoder in c2f-coref was fully replaced by BERT in [5]. BERT representation of the beginning word piece, ending word piece, as well as the attended form of the whole mention were concatenated to represent the mention. The representations are fed for CR following the same paradigm of c2f-coref. Later on, SpanBERT [25] was proposed and used in CR, which leads to better representations of mentions (better than roughly concatenate presentations of a number of tokens in the mention).

B. GTM-CR Models

Along with the introduction of several GTM-CR datasets, including GAP [3], DPR [2], and Winogender [4] (see more details in Section IV-A), a bank of models that focus on the GTM-CR task has been introduced.

The current SOTA GTM-CR model on the GAP dataset was presented by Attree et al. [6]. It includes two main components: the pronoun BERT module and the evidence pooling module. The pronoun BERT module extracted the last layer embedding for the pronoun from the BERT model. The evidence pooling module combined the clustering information from four other CR models: AllenNLP [27], NeuralCoref, Parallelism+URL [3] and e2e-coref [12]. The evidence pooling would encode the information from all these models via the self-attention mechanism and generate an evidence vector. Finally, the evidence vector is concatenated with the BERT embedding of the pronoun and goes through the linear and softmax layers to get the classification result.

The current SOTA GTM-CR model on DPR and Winogender datasets is [8] that re-trains BERT with WikiCREM. They first collected a large-scale unsupervised corpus generated from English Wikipedia, namely, the Wikipedia CoREferences Masked (WIKICREM) dataset, and re-trained BERT on it. On WIKICREM, they designed a training task targeting pronoun resolution as its downstream task. More specifically, they gave BERT a sentence with an antecedent or a pronoun masked out, together with two candidates. The model is then to predict which of the candidate is more proper. For CR, they trained two versions of BERT_WIKICREM: BERT_WIKICREM_DPR which was fine-tuned as a CR model on DPR and BERT_WIKICREM_ALL which was fine-tuned on both GAP and DPR. BERT_WIKICREM_ALL performed the best on DPR, achieving an accuracy of 84.8% while BERT_WIKICREM_DPR performed the best on Winogender, achieving an accuracy of 82.1%.

C. Motivations of This Work

Previous SOTA GTM-CR models [11], [12] learned the task in a single-task learning fashion. However, the multi-task learning community [28], [29], [30] believes that learning related but different tasks can achieve complementary strengths for the learning of each task. Besides, we found that learning mention identification likely yields supportive dependent information for the learning of mention linking (see Fig. 1). Thus, we were motivated to introduce multi-task learning into the GTM-CR task. Since this was a new learning paradigm for GTM-CR, we were also inspired to explore the gains and losses of using multi-task learning, compared to single-task learning.

III. Method

In this section, we present our MTL-based GTM-CR model, namely Coref-MTL, as well as a dynamic weight balancing algorithm that allocates a dynamic weight for the learning of each task loss to support the MTL.

A. Joint Model for Mention Identification and Mention Linking

Fig. 2 represents the structure of the proposed model that collaboratively optimizes the mention identification and mention linking tasks. The input sentence is first processed by the PLM after tokenization, the outputs of which are then passed into two task-specific towers. In what follows, we introduce details of each task-specific tower.

1) Mention Identification: Given an input representation encoded by a PLM, the mention identification module decides whether each token in the input belongs to a mention or not. A mention is an antecedent or a referent. A mention can contain more than one token. The mention identification (MI) task-specific tower passes the representations from PLM through l layers of Transformer encoders. Formally, suppose $X_0 \in \mathbb{R}^{s\times e}$ is the contextualized representation of the input sentence, where s represents the sequence length and e represents the embedding size. Then, the representation ($X_i^{MI} \in \mathbb{R}^{s\times e}$) of the sentence after passing through the ith Transformer encoder $\text{TransEnc}_i^{MI}(\cdot)$ ($i \in \{1, ..., l\}$) is given by:

$$X_i^{MI} = \text{TransEnc}_i^{MI}\left(X_{i-1}^{MI}\right). \tag{1}$$

Subsequently, the hidden states of the last Transformer layer (X_l^{MI}) are linearly transformed and passed to a softmax layer to generate a set of probability distributions over whether a token belongs to an entity mention span or not:

$$P^{MI} = \text{softmax}\left(W_1^T X_l^{MI} + b_1\right), \tag{2}$$

where P^{MI} ($P^{MI} \in \mathbb{R}^{2\times s}$) denotes the set of the probability distribution over the output space for all input tokens in the mention identification task. W_1 ($W_1 \in \mathbb{R}^{e\times 2}$) and b_1 ($b_1 \in \mathbb{R}^2$) are parameters learned parameters. We use Cross Entropy loss L_{MI} for the mention identification task:

$$L_{MI} = CrossEntropy\left(\hat{Y}^{MI}, Y^{MI}\right), \tag{3}$$

where $\hat{Y}^{MI}$ is the predicted labels of the mention identification task, based on P^{MI}. Y^{MI} denotes the ground-truth labels.

2) Mention Linking: Similar to the mention identification task, for the mention linking task, the embedding of the tokenized sentence is first routed through numerous stacked Transformer layers as depicted in Fig. 2 to provide the final vectorized representation required for the mention linking task. In the mention linking (ML) task-specific tower, we used k stacked Transformer encoders. Then, the representation ($X_j^{ML} \in \mathbb{R}^{s\times e}$) of the sentence after passing through the jth Transformer encoder $\text{TransEnc}_j^{ML}(\cdot)$ ($j \in \{1, ..., k\}$) is given by:

$$X_j^{ML} = \text{TransEnc}_j^{ML}\left(X_{j-1}^{ML}\right). \tag{4}$$

The vector representations of mentions are extracted by a mask from the last Transformer layer of the mention linking task-specific tower. The mask is a list of digits of 0 or 1 with each element matching the corresponding token in the input sentence (see Fig. 2 (b)). The mask is used to do the element-wise products with the vector representations of tokens.

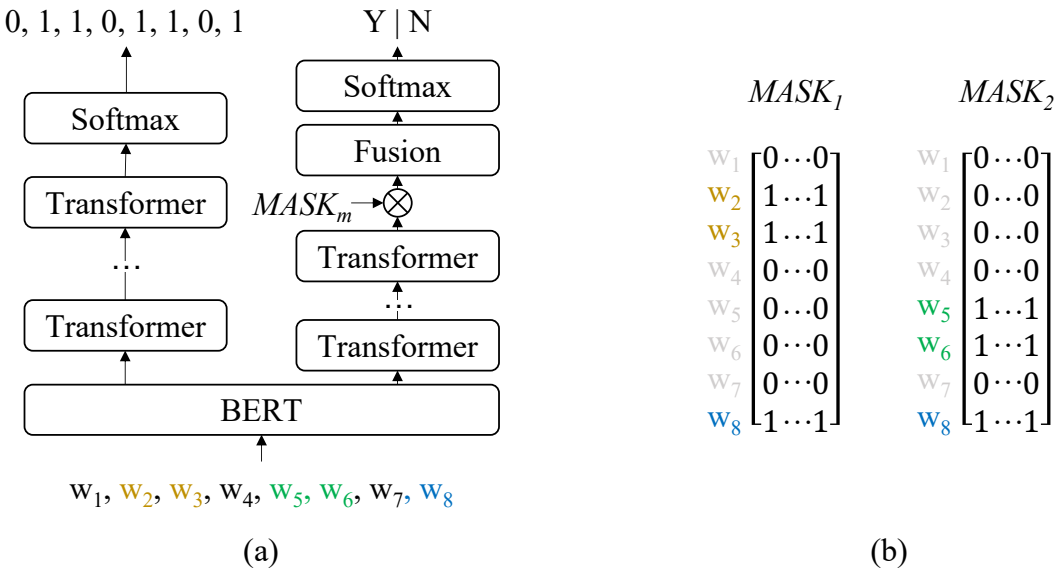


Fig. 2: (a) Model structure of multi-task co-reference resolution (Coref-MTL). The colored input tokens denote mentions and pronouns. $Mask_m$ is the mask for obtaining the representations of an antecedent-pronoun pair. $\otimes$ denotes element-wise product. (b) The shape of masks for obtaining the representations of pronoun w_8 and antecedent $\{w_2, w_3\}$ and the representations of pronoun w_8 and antecedent $\{w_5, w_6\}$.

For the irrelevant tokens (tokens out of mention spans in a sentence), their representations are wiped out by multiplying with the 0s in the mask. For candidate mentions and pronouns, their representations are retained by multiplying with the 1s in the mask. In practice, a sentence may have multiple masks, if there are more than two mention spans. Each mask retains the representations of an antecedent-pronoun pair. Formally, the representations ($V \in \mathbb{R}^{s \times e}$) of the input after masking out with a mask ($Mask_m$) for mention linking are given by

$$V = Mask_m \otimes X_l^{ML}, \tag{5}$$

where $\otimes$ represents element-wise products. In the end, we only retain two vectors. One represents a candidate mention. The other one represents the pronoun. These two vector representations are utilized to determine whether or not the represented candidate antecedent and the pronoun are co-referred.

The representations of the candidate mentions are the average of all unmasked constituent tokens. This is because an antecedent may contain more than one word, e.g., a multi-word expression. Let $\mathbf{v}_c \in \mathbb{R}^e$ denote the candidate antecedent representation vector and $\mathbf{v}_p \in \mathbb{R}^e$ denote the representation vector of the pronoun. These two vectors are fused via a fusion operation to obtain a fused vector $\mathbf{v}_f \in \mathbb{R}^e$. We examined various fusion methods in this study, including the concatenation of the two vectors, element-wise addition, element-wise products, and element-wise square of the difference. We found that element-wise products yielded the best performance. Formally, the new fusion vector ($\mathbf{v}_f \in \mathbb{R}^e$) is given by:

$$\mathbf{v}_f = \mathbf{v}_c \otimes \mathbf{v}_p \in \mathbb{R}^e. \tag{6}$$

After obtaining the fusion vector $\mathbf{v}_f$, it is processed through a softmax layer to produce the probability distribution over two classes: 1 (there is a co-reference relationship between the pronoun and the candidate mention) and 0 (there is no co-reference link between the two mentions)

$$p^{ML} = \text{softmax}\left(W_2^T \mathbf{v}_f + b_2\right) \in \mathbb{R}^2, \tag{7}$$

where p^{ML} ($p^{ML} \in \mathbb{R}^2$) denotes the probability distribution over the output space for the mention linking task. W_2 and b_2 ($b_2 \in \mathbb{R}^2$) are trainable parameters. The size of W_2 is $\mathbb{R}^{2e \times 2}$ if the fusion method is concatenation and $\mathbb{R}^{e \times 2}$ otherwise. Next, we use cross entropy loss to learn the mention linking task

$$L_{ML} = CrossEntropy(\hat{y}^{ML}, y^{ML}), \tag{8}$$

where $\hat{y}^{ML}$ denotes a mention linking predicted label, based on p^{ML}. y^{ML} denotes a ground-truth label of the mention linking task.

The final loss L_{Coref} is the weighted sum of the mention identification loss (L_{MI}) and the mention linking loss (L_{ML})

$$L_{Coref} = w_{MI} L_{MI} + w_{ML} L_{ML}, \tag{9}$$

where w_{MI} and w_{ML} are two hyperparameters that represent the loss weights of the two sub-tasks, respectively.

B. Dynamic Weight Balancing

When training mention identification and mention linking together for GTM-CR, it is important to decide what weight should be allocated to each task, because the difficulty of learning mention identification and mention linking tasks are different.

We adopted a dynamic weight-balancing method based on loss adaption. After each epoch, we record the losses of each task and compare that with the initial loss after the first epoch. For tasks that have more reduction in the loss, it will be weighted less in the next epoch of training. For tasks that have less reduction in the loss, they will have higher weights in the next iteration. Adjustment is based on the square of the relative percentage of all tasks and they are passed through a softmax function to make sure all the weights of different tasks sum to one. A detailed description of the proposed weight-balancing algorithm can be found in Algorithm 1.

Algorithm 1 Dynamic Weight Balancing based on Loss Changes

Given T tasks.
Initialize task weights to be $1/T$.
for each epoch e **do**
  for each batch b **do**
    get batch loss and add to epoch loss
    Update weighted loss $\ell_{(e,b)} = \sum_{i=1}^{T} \ell_{(e,b,t)} \times w_t$
    Update W with respect to $\ell_{(e,b)}$.
  end for
  Get the epoch loss on each task $\ell_e \in \mathbb{R}^T$.
  the first epoch loss as $\ell_0 \in \mathbb{R}^T$.
  for each task t **do**
    Set the task weight $w_t = \left(\frac{\ell_{(e,t)}}{\ell_{(0,t)}}\right)^2$.
  end for
  Unify the weights to make sure they sum to one

$$w_t = \frac{e^{w_t}}{\sum_{i=1}^{T} e^{w_i}}$$

end for

IV. Experiments

A. Datasets and Evaluation Protocols

In our experiment, we used three GTM-CR datasets: GAP, DPR, and Winogender. Table I records the statistics of each dataset.

a) GAP [3]: GAP is collected from Wikipedia to reflect the real-world challenges of pronoun co-reference resolution. It comprises 8908 pairs of pronoun and candidate mention that are split into 3 subsets: test (4000 pairs), development (4000 pairs), and validation (908 pairs). In each subset, there is the same number of examples with masculine pronouns (e.g., him, his) and feminine pronouns (e.g., she, her). We use F1 and Bias as the main evaluation measures on the GAP dataset. F1 scores are calculated on three kinds of examples: the group of examples with masculine pronouns, the group of examples with feminine pronouns, and the group of examples with all kinds of pronouns. The bias score is calculated as the ratio of feminine F1 over masculine F1.

TABLE I: Statistics of datasets, Val. stands for the validation set. P-M stands for the number of pronoun-mention pairs. As there are two mentions and one pronoun for each example, the number of P-M pairs is twice the total number of examples. Avglen stands for the average length of the examples in terms of words.

Dataset	Train	Val.	Test	Total	P-M	Avglen
GAP	4000	908	4000	8908	17816	71.57
DPR	1322	-	564	1886	3772	14.27
Winogender	-	-	720	720	1440	14.49

b) DPR [2]: The Definite Pronoun Resolution (DPR) corpus is a modified version of WSC minimal pairs (see Section I). These sentence pairs span a wide range of themes, from real occurrences to cinematic events to entirely fictitious circumstances, primarily representing pop culture as experienced by American children born in the early 1990s. DPR includes cases that do not need commonsense reasoning, as well as situations where the "special word" is a phrase. DPR contains 1322 training examples and 564 test examples. Totally, there are 1886 example sentences.

c) Winogender [4]: Winogender is a dataset for testing the gender biases in CTM-CR, using the WSC format. Each sentence has an occupational noun and a referring pronoun. The pronoun could be represented as "he", "she" or "they", respectively. The occupational nouns are usually gender-oriented. E.g., women are likely to be employed as secretaries. Given "the secretary asked the visitor to sign in so that he could update the guest log" [4], a co-reference resolution classifier may fail in connecting "he" to "secretary" if the classifier is gender-biased. This dataset means to examine how altering the gender of the pronoun impacts the accuracy of a model. Winogender contains 720 sentences in total. Winogender is only used as a test set.

B. Baselines

We used SOTA GTM-CR models on the three datasets as our baselines. ProBERT/GREP [6] is the current SOTA model for the GAP dataset. BERT fine-tuned on Wiki-CREM [8] (henceforth, BERT_WIKICREM) is the current SOTA model for the DPR and Winogender datasets. For BERT_WIKICREM, we tried two different versions, namely BERT_WIKICREM_ALL as well as BERT_WIKICREM_DPR. BERT_WIKICREM_ALL has the best performance on DPR dataset whereas BERT_WIKICREM_DPR has the best performance on Winogender dataset. Details of each model can be found in Section II.

C. Setups

NVIDIA RTX3080Ti GPU was used to train all models (16GB memory). The language model component was implemented using the *Transformers* library from huggingface [31], and initialization was accomplished using pre-trained checkpoints. Adam [32] optimizer was employed with $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1e^{-8}$, and a warm-up learning rate of $2e^{-6}$. The mention identification specific-tower has 2 Transformer encoder layers ($l = 2$), while the mention linking specific-tower has 4 Transformer encoder layers ($k = 4$).

TABLE II: Performance on the GAP test dataset, F1, and bias are used as the evaluation metrics.

Model	Overall	Masculine	Feminine	Bias
ProBERT[6]	89.70	90.80	88.60	98.00
GREP [6]	92.50	**94.00**	91.10	97.00
Coref-MTL (Ours)	**92.72**	92.65	**92.45**	**99.76**

All layers were regularized using a fixed dropout [33] rate of 0.2. For each Transformer encoder, embedding size e was set as 1024, and the number of heads was set to 16. The batch size is set as 32 while training. To evaluate the performance of each model, the model was trained on the DPR training set when evaluating DPR and Winogender and trained on the GAP training set when evaluating GAP, the best checkpoints were determined by the validation dataset if the validation set is available and by the test dataset if the validation set is not available. If the corresponding metrics (F1 or accuracy) are not improved after 60 consecutive epochs, the training will be stopped.

V. Results

In this section, we first compare our full model with the current SOTA models on the GAP, DPR, and Winogender datasets, followed by a hyperparameter analysis. Next, we added an ablation study to investigate how the model will perform after the DWB and MTL mechanisms were removed. Furthermore, we did an error analysis to discuss why our multi-task learning model could improve performance on GTM-CR datasets.

A. Overall Results

a) Results on GAP: Table II charts the results of our model and baselines on GAP. As seen, our Coref-MTL outperforms all the baseline models in terms of overall F1 and Bias. This improvement mainly attributes to the ability to link feminine pronouns to the correct candidates (feminine F1 increased by 1.3% compared with GREP), therefore reducing the overall bias.

b) Results on DPR and Winogender: Table III presents the results of our models and the baselines on the DPR test set and Winogender dataset. As seen, on DPR, Coref-MTL's performance is slightly below the current SOTA model BERT_WIKICREM_ALL, but is higher than BERT_WIKICREM_DPR. The major reason for being lost to BERT_WIKICREM_ALL is because that BERT_WIKICREM_ALL was re-trained on multiple corpora (i.e., WIKICREM, GAP, and DPR) in a mention-aware way (which was designed specifically for CR) and, more importantly, one of them, i.e., WIKICREM, is remarkably large (approximately 2.4M samples). In contrast, our Coref-MTL is fine-tuned on only DPR, which, for reference, contains only 0.05% samples of WIKICREM. Additionally, Coref-MTL wins BERT_WIKICREM_DPR, which was re-trained on both WIKICREM and DPR.

TABLE III: Performance on the DPR and Winogender datasets.

Model	DPR acc.	Winogender acc.
BERT_WIKICREM_ALL [8]	**84.80**	76.70
BERT_WIKICREM_DPR [8]	80.00	82.10
Coref-MTL (Ours)	84.57	**83.06**

TABLE IV: Ablation study of removing dynamic weight balancing (DWB) and multi-task learning (MTL) respectively. Acc stands for accuracy. For GAP Bias, the closer to 100 the better.

Model	GAP F1	GAP Bias	DPR acc.	Winogender acc.
Coref-MTL	**92.72**	**99.76**	**84.57**	**83.06**
-DWB	92.35	99.74	84.31	80.35
-MTL	92.22	99.22	83.60	80.21

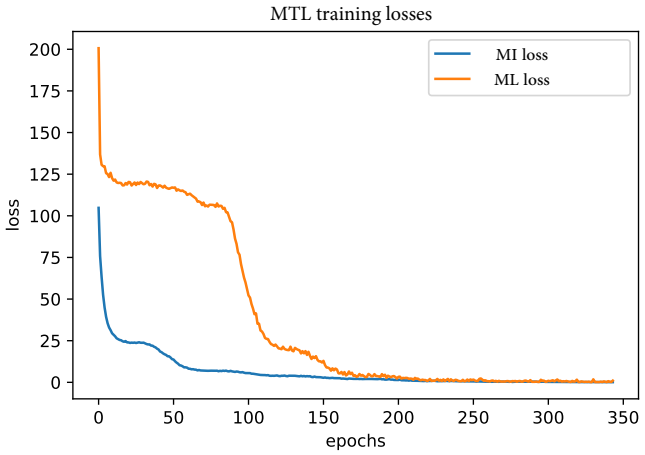
This suggests that modeling mention identification jointly helps to train a CR model with way less data, which is not only saving time but also is environmentally friendly (as we trained our model with much fewer computing resources). On Winogender, in line with the results on GAP, which also aims at gender bias in CR, our Coref-MT performs very well and defeats both baselines. Its success on GAP and Winogender embodies that better modeling the locations of antecedents (through modeling mention identification jointly) helps with eliminating spurious correlations while training a co-reference resolver. Additionally, by comparing the two baselines, in addition to reducing biases, MTL, again, saves time and energy.

B. Ablation Study

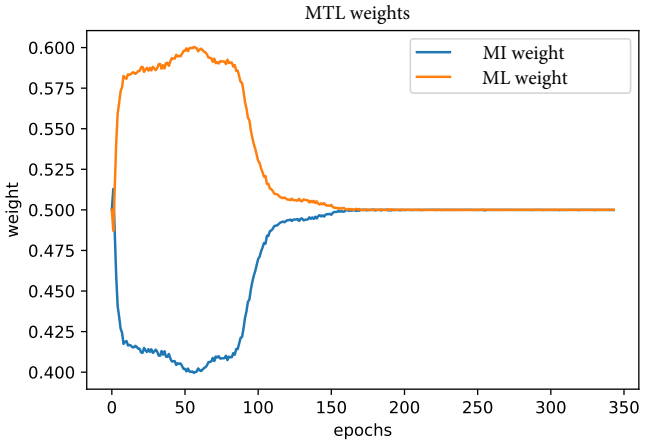
To analyze the contribution of each component of our model, we report on an ablation study by removing either DWB or both DWB and MTL (which results in a single-task learning (STL) model, henceforth Coref-STL).

Table IV records the results of the ablation study on the three datasets. Generally, on all three datasets, removing any components would scarify the performance. For example, on GAP, the overall F1 score drops by 0.37% after removing DWB, and further drops to 92.22% by further removing MTL. Similar trends also happen in the rest two datasets. These suggest that both MTL and a good strategy for balancing the tasks during training play vital roles for CR. We discuss more in what way DWB and MTL help CR in Section V-C.

Focusing the results on Winogender, we observe that the gain of using MTL and DWB is significantly higher than that of the other two datasets. Recall that Winogender is the only dataset that has no training set which models cannot fine-tune on. One explanation is that the capability of reducing biases of joint modeling mention identification and linking using MTL and DWB may come from its contribution to the generalizability of the model (by eliminating spurious correlations).



(a)



(b)

Fig. 3: (a) illustrates loss changes for Coref-MTL when training on the DPR training set. (b) illustrates the weight changes when training Coref-MTL. MI stands for mention identification task while ML stands for mention linking task

C. The Effects of MTL and Dynamic Weighting

To better understand the effects of MTL and the DWB mechanism, we print how the loss and weight of each task change across time in Fig. 3. When the training begins, the loss of both tasks drops dramatically, but since the mention identification loss starts at a relatively low value, mention identification loss approaches zero. Thus, subsequently, the weight for mention identification decreases while the weight for mention linking increases, making the training focus more on mention linking. From this moment, a good mention identification helps to train mention linking better and faster. Along with the reduction of mention linking loss, the loss of mention identification and mention linking increases and decreases, respectively. After the two weights converge to be balanced and steady, the two tasks are trained collaboratively and further improve the performance of mention linking.

TABLE V: Error analysis on Winogender dataset based on pronoun types, Male represents male pronouns (he, his, and him), Female represents female pronouns (she and her), Gender neural represents gender-neutral pronouns (they, them, and their)

Model	Masculine	Feminine	Gender-neutral	Total
Coref-MTL	88	75	81	244
Coref-STL	86	117	84	287

D. Error Analysis

To understand better the behavior of our Coref-MTL model, we analyze and compare the errors made by Coref-MTL and Coref-STL on the Winogender dataset. In what follows, we describe our observations.

First, we counted the errors that are related to masculine (i.e., 'he', 'his' and 'him'), feminine (i.e., 'she' and 'her'), and gender-natural (i.e., 'they', 'them' and 'their') pronouns made by each model. The counts are reported in Table V. Consistent with our evaluation experiments (see Section V-A) and the ablation study (see Section V-B), compared to Coref-STL, the contribution of MTL to the overall performance is that it makes much less error on feminine pronouns. The errors made by our Coref-MTL are almost uniformly distributed over the three types.

Second, it has been pointed out that the use of gender-neutral pronouns (e.g, they, for reducing gender bias) would cause agreement mismatch problems [34]. For example, in Winogender, there are many cases like the following:

(2) The clerk provided someone with paperwork to return to **them** upon completion.

in which, the plural pronoun 'them' refers to a definite plural noun phrase. Nonetheless, as we can see from Table V, it appears that both Coref-MTL and Coref-STL work fine in these mismatch cases. This embodies that neural CR models might not resolve co-references by learning semantic constraints.

Third, the data in GTM-CR dataset is a form, where each text is judged by a model twice: once for the link between the pronoun and the first candidate antecedent and once for the link with another antecedent. By looking into the distribution of the incorrect predictions of the two models, we found that Coref-STL is more inclined to predict incorrectly on both two judges than Coref-MTL. Additionally, we also observed that, among all incorrect predictions, Coref-STL is likely to predict that there is no co-reference relation while the incorrect predictions of Coref-MTL are rather uniformly distributed. This said, maybe, for some inputs, Coref-STL gets stuck and predicts randomly.

Last, regarding indefinite antecedents (i.e., 'someone'), Coref-MTL and Coref-STL have very similar behaviors, which suggests MTL does not help in this respect.

VI. Conclusion

GTM-CR is an essential type of CR. Existing models only focus on a mention-linking sub-task in GTM-CR exploiting the fact that mentions are provided. In this paper, we demonstrated that mention identification is still helpful for building a GTM co-reference resolver and proposed a multi-task learning model that jointly trains the mention identification and mention linking tasks. This is achieved by assuming the mentions are not known during mention identification and forcing the model to identify them. Meanwhile, the learning of two different but related tasks may share complementary dependent information. The weights of the two tasks are adjusted dynamically during the training process. This setting has achieved new SOTA performance on two GTM style datasets (GAP and Winogender) and comparative results on another dataset (DPR) without fine-tuning on additional large corpora. As future work, we plan to develop linguistics-inspired models [35], since CR is a field with a solid theoretical foundation in linguistics. Incorporating linguistic theories, such as donkey sentences [36], and pronoun-dropping [37], may enhance the explainability and effectiveness of neural network models.

References

[1] H. Levesque, E. Davis, and L. Morgenstern, "The winograd schema challenge," in *Thirteenth international conference on the principles of knowledge representation and reasoning*, 2012.

[2] A. Rahman and V. Ng, "Resolving complex cases of definite pronouns: The Winograd schema challenge," in *Proceedings of EMNLP-CoNLL*. Association for Computational Linguistics, 2012, pp. 777–789.

[3] K. Webster, M. Recasens, V. Axelrod, and J. Baldridge, "Mind the GAP: A balanced corpus of gendered ambiguous pronouns," *Transactions of the Association for Computational Linguistics*, pp. 605–617, 2018.

[4] R. Rudinger, J. Naradowsky, B. Leonard, and B. Van Durme, "Gender bias in coreference resolution," in *Proceedings of NAACL (Short Papers)*. Association for Computational Linguistics, 2018, pp. 8–14.

[5] M. Joshi, O. Levy, L. Zettlemoyer, and D. Weld, "BERT for coreference resolution: Baselines and analysis," in *Proceedings of EMNLP-IJCNLP*. Association for Computational Linguistics, 2019, pp. 5803–5808.

[6] S. Attree, "Gendered ambiguous pronouns shared task: Boosting model confidence by evidence pooling," in *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, 2019, pp. 134–146.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL*, 2019, pp. 4171–4186.

[8] V. Kocijan, O.-M. Camburu, A.-M. Cretu, Y. Yordanov, P. Blunsom, and T. Lukasiewicz, "WikiCREM: A large unsupervised corpus for coreference resolution," in *Proceedings of EMNLP-IJCNLP*. Association for Computational Linguistics, 2019, pp. 4303–4312.

[9] R. J. Stevenson, R. A. Crawley, and D. Kleinman, "Thematic roles, focus and the representation of events," *Language and Cognitive Processes*, vol. 9, no. 4, pp. 519–548, 1994.

[10] C. G. Chambers and R. Smyth, "Structural parallelism and discourse coherence: A test of centering theory," *Journal of Memory and Language*, vol. 39, no. 4, pp. 593–608, 1998.

[11] H. Daumé III and D. Marcu, "A large-scale exploration of effective global features for a joint entity detection and tracking model," in *Proceedings of EMNLP*. Association for Computational Linguistics, 2005, pp. 97–104.

[12] K. Lee, L. He, M. Lewis, and L. Zettlemoyer, "End-to-end neural coreference resolution," in *Proceedings of EMNLP*. Association for Computational Linguistics, 2017, pp. 188–197.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[14] G. Kundu, A. Sil, R. Florian, and W. Hamza, "Neural cross-lingual coreference resolution and its application to entity linking," in *Proceedings of ACL*. Association for Computational Linguistics, 2018, pp. 395–400.

[15] Z. Dai, H. Fei, and P. Li, "Coreference aware representation learning for neural named entity recognition." in *IJCAI*, 2019, pp. 4946–4953.

[16] E. Cambria, S. Poria, A. Gelbukh, and M. Thelwall, "Sentiment analysis is a big suitcase," *IEEE Intelligent Systems*, vol. 32, no. 6, pp. 74–80, 2017.

[17] G. Chen, K. van Deemter, and C. Lin, "Modelling pro-drop with the rational speech acts model," in *Proceedings of INLG*. Tilburg University, The Netherlands: Association for Computational Linguistics, 2018, pp. 159–164.

[18] F. Same, G. Chen, and K. Van Deemter, "Non-neural models matter: a re-evaluation of neural referring expression generation systems," in *Proceedings of ACL*, 2022, pp. 5554–5567.

[19] G. Chen, F. Same, and K. van Deemter, "Neural referential form selection: Generalisability and interpretability," *Computer Speech & Language*, vol. 79, p. 101466, 2023.

[20] R. Sukthanker, S. Poria, E. Cambria, and R. Thirunavukarasu, "Anaphora and coreference resolution: A review," *Information Fusion*, vol. 59, pp. 139–162, 2020.

[21] R. Liu, R. Mao, A. T. Luu, and E. Cambria, "A brief survey on recent advances in coreference resolution," *Artificial Intelligence Review*, 2023.

[22] X. Ding and B. Liu, "Resolving object and attribute coreference in opinion mining," in *Proceedings of COLING*, 2010, pp. 268–276.

[23] K. Clark and C. D. Manning, "Entity-centric coreference resolution with model stacking," in *Proceedings of ACL-IJCNN*. Association for Computational Linguistics, 2015, pp. 1405–1415.

[24] A. Emami, A. Trischler, K. Suleman, and J. C. K. Cheung, "A generalized knowledge hunting framework for the Winograd schema challenge," in *Proceedings of NAACL: Student Research Workshop*. Association for Computational Linguistics, 2018, pp. 25–31.

[25] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy, "SpanBERT: Improving pre-training by representing and predicting spans," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 64–77, 2020.

[26] K. Lee, L. He, and L. Zettlemoyer, "Higher-order coreference resolution with coarse-to-fine inference," in *Proceedings of NAACL (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, 2018, pp. 687–692.

[27] M. Gardner, J. Grus, M. Neumann, O. Tafjord, P. Dasigi, N. F. Liu, M. Peters, M. Schmitz, and L. S. Zettlemoyer, "AllenNLP: A deep semantic natural language processing platform," in *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, 2017.

[28] R. Mao and X. Li, "Bridging towers of multi-task learning with a gating mechanism for aspect-based sentiment analysis and sequential metaphor identification," *Proceedings of AAAI*, vol. 35, no. 15, pp. 13 534–13 542, 2021.

[29] M. Ge, R. Mao, and E. Cambria, "Explainable metaphor identification inspired by conceptual metaphor theory," *Proceedings of AAAI*, vol. 36, no. 10, pp. 10 681–10 689, 2022.

[30] R. Mao, X. Li, M. Ge, and E. Cambria, "MetaPro: A computational metaphor processing model for text pre-processing," *Information Fusion*, vol. 86-87, pp. 30–43, 2022.

[31] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of EMNLP*. Association for Computational Linguistics, 2020, pp. 38–45.

[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014. [Online]. Available: https://arxiv.org/abs/1412.6980

[33] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014.

[34] M. Poesio, J. Yu, S. Paun, A. Aloraini, P. Lu, J. Haber, and D. Cokal, "Computational models of anaphora," *Annual Review of Linguistics*, vol. 9, pp. 561–587, 2023.

[35] R. Mao, C. Lin, and F. Guerin, "End-to-end sequential metaphor identification inspired by linguistic theories," in *Proceedings of ACL*. Association for Computational Linguistics, 2019, pp. 3888–3898.

[36] A. Brasoveanu, "Donkey pluralities: plural information states versus non-atomic individuals," *Linguistics and philosophy*, vol. 31, no. 2, pp. 129–209, 2008.

[37] H. Bussmann, K. Kazzazi, and G. Trauth, *Routledge dictionary of language and linguistics*. Routledge, 2006.