Advances in Artificial Intelligence and Machine Learning; Research 3 (2) 1122-1134 Received 10-04-2023; Accepted 11-06-2023; Published 18-06-2023

Linguistically-Inspired Neural Coreference Resolution

Xuanyue Yang xuanyuey@alumni.cmu.edu
Google Inc, 1600 Amphitheatre Parkway,
United States

Wenting Ye wye2@alumni.cmu.edu
ByteDance Ltd, 5800 Bristol
Pkwy Culver City, CA 90230
United States

Luke Breitfeller mbreitfe@andrew.cmu.edu
Carnegie Mellon University,
5555 Forbes Ave
United States

Tianwei Yue tyue@andrew.cmu.edu
Carnegie Mellon University,
5555 Forbes Ave
United States

Wenping Wang wenpingw@alumni.cmu.edu
Carnegie Mellon University,
5555 Forbes Ave
United States

Corresponding Author: Wenting Ye

Abstract

The field of coreference resolution has witnessed significant advancements since the introduction of deep learning-based models. In this paper, we replicate the state-of-the-art coreference resolution model and perform a thorough error analysis. We identify a potential limitation of the current approach in terms of its treatment of grammatical constructions within sentences. Furthermore, the model struggles to leverage contextual information across sentences, resulting in suboptimal accuracy when resolving mentions that span multiple sentences. Motivated by these observations, we propose an approach that integrates linguistic information throughout the entire architecture. Our innovative contributions include multitask learning with part-of-speech (POS) tagging, supervision of intermediate scores, and self-attention mechanisms that operate across sentences. By incorporating these linguistic-inspired modules, we not only achieve a modest improvement in the F1 score on CoNLL 2012 dataset, but we also perform qualitative analysis to ascertain whether our model invisibly surpasses the baseline performance. Our findings demonstrate that our model successfully learns linguistic signals that are absent in the original baseline. We posit that these enhance-

Citation: Xuanyue Yang, et al. Linguistically-Inspired Neural Coreference Resolution. Advances in Artificial Intelligence and Machine Learning. 2023;3(2):66.

ments may have gone undetected due to annotation errors, but they nonetheless lead to a more accurate understanding of coreference resolution.

Keywords: Coreference resolution, Part of speech tagging, Linguistic awareness, Natural language processing

1. INTRODUCTION

Coreference resolution aims to determine when two or more expressions within a text refer to the same entity. The state-of-the-art in coreference resolution comes from a neural model proposed by Lee et al. [1]. The specific task at hand utilizes annotated documents from the CoNLL 2012 database [2] to identify potential entity mentions and compute a coreference likelihood for each mention pair.

Our initial efforts revolved around error analysis, believing that the identification of prevalent errors within the Lee et al. [1] baseline could inspire improvements. We observed two primary categories of errors: 1) errors in entity recognition that led to subsequent errors, and 2) errors during the coreference resolution step. For both types, a significant proportion of these errors seemed to originate from the model overlooking crucial linguistic cues, hence affirming our hypothesis that linguistic awareness could enhance the state-of-the-art.

Table 1: CoNLL data and error analysis

Speaker 1: This is Xu Li. Thank **you everyone** for watching. Coming up is the Focus today program hosted by Wang Shilin. Good-bye, **dear viewers**. *Gold:* [you, everyone] *Baseline:* [everyone, dear viewers]

Examining the first type of error, we found instances where the model's chosen entities deviated from the gold-standard human annotations.[1] For instance, TABLE 1 demonstrates a case where the phrase "thank you everyone" was interpreted by the CoNLL human annotators as including an antecedent "you" and anaphora "everyone".[2] The model, meanwhile, groups "you everyone" as a single mention which naturally impacts accuracy on any coreference pairs which would include either of the original mentions. From these findings, we concluded that a heightened understanding of the interplay between grammatical constructions and entities could improve mention recognition.

Investigating the second type of error, we found the model overemphasized the similarity between entities, often disregarding broader context. For instance, in TABLE 2, two different speakers referred to "a friend," with the context indicating different entities, but the model considered the mentions coreferent. What is notable about this example is there are many traceable factors that led human eyes to see the mentions are not coreferent–that there are two different speakers talking, that the second mention uses the indefinite article "a" as though introducing new information, and the

[1] We also encountered cases where human annotations differed from our team's annotations, hinting that the CoNLL annotations may not accurately represent the ground truth.

[2] This becomes more clear when a comma is added–"thank you, everyone" is understood to mean the same as the sentence without the comma, but more clearly demonstrates that "you" and "everyone" are separate entity mentions.

Table 2: CoNLL data and error analysis

Speaker 1: ... how the two of you found out the news on the day of the accident? *Speaker 2*: It happened that I was going to have lunch with **a friend**, um, at noon ... *Speaker 1*: And you, Yang Yang? *Speaker 3*: **A friend** happened to call me. *Gold:* [] *Baseline:* [a friend, a friend]

greater context that "friend" is a common word in the English and likely to be utilized for multiple distinct entities. The current SOTA ignores all these. Identifying this type of error showed that considering the entire document's context is crucial for coreference resolution.

In this paper, we re-implement and enhance the state-of-the-art neural coreference resolution model proposed by Lee et al. [1], focusing on improving the existing neural architecture to better capture the linguistic signals that exist within the input data. Our main contributions are as follows:

1. We introduce part-of-speech (POS) tagging as a multitask objective within the LSTM's hidden layer to provide the model with grammatical signals.
2. We supervise the intermediate coreference scores to improve convergence speed and overall accuracy.
3. We enhance the model's understanding of context across sentences by taking the full text as input and incorporating self-attention layers for inter-sentence relationships.

2. RELATED WORK

Deep learning based model has been widely adopted in natural language processing and its application [3–5] due to its outstanding performance. The majority of Lee et al. [1]'s coreference model was originally codified in Lee et al. [6]. Lee et al. [6] establishes the use of the bidirectional LSTM for mention identification and the trained coreference matrix (discussed in more detail in Section 3). Lee et al. [1] saw improvements over this prior model by adding neural layers to process second-order coreference information and added coarse-to-fine neural layers.

Many existing coreference approaches do integrate specific linguistic knowledge, either as direct features or through neural architectures designed to model linguistic processes. Moosavi and Strube [7] observed improvements in their model when they integrated certain types of linguistic features, and Ng [8]'s model also utilized linguistic data. Meanwhile, Tai et al. [9] found improvements by changing their bi-directional LSTM to a tree-structured LSTM and Dhingra et al. [10] utilized directed acyclic graphs to model external information. In both cases, these architectural changes allowed them to better model the linguistic structure of language. Though we did not implement these exact modules, this approach to feature selection and model architecture heavily influenced our own process.

We also examine some recent research advancements in the field. To address the challenges associated with extracting event components, Zeng et al. [11] proposed a novel approach that incorporates event-specific paraphrases and semantic embeddings to capture paraphrase relations and mitigate error propagation. Building upon span-based models commonly used in information extraction tasks, Lu and Ng [12] extended these models to tackle event coreference resolution. Subsequently, the same authors in Lu and Ng [13] emphasized the importance of knowledge-rich joint event coreference models, attributing the observed improvements to their approach. In another study, Lu and Ng [14] introduced a multi-task model that addresses event coreference through five distinct tasks. Additionally, Lu et al. [15] developed an end-to-end pipeline that jointly models event detection and event coreference resolution tasks, incorporating a type-enhanced event coreference mechanism. These approaches exhibit notable strengths and demonstrate superior performance on benchmark datasets compared to baseline models. However, it is important to note that these studies did not explicitly address the linguistic awareness of models, such as handling grammatical constructions and effectively leveraging contextual information for the event coreference task.

We also drew upon works outside the task of exact event coreference, and even examined older papers based more in linguistic theory than quantitative modeling results. For instance, the theory of grammatical centering Grosz et al. [16] helped us conceptualize the property of coreference we were hoping to capture by including linguistic data, and the majority of our direct inspiration came from a work on semantic role labeling rather than coreference.

Our chief inspiration in this work is a semantic role labeling model developed by Strubell et al. [17]. Though the paper did not focus on the task of event coreference, the motivations behind many of their innovations were derived from the desire to fully utilize existing linguistic information. For that reason, we believed that adapting their techniques to the task of coreference resolution might yield promising new results. For example, Strubell et al. [17] utilized multi-task learning using POS tags as a means of indirectly teaching the semantic role model about the grammatical structure of its inputs. It also incorporated self-attention learning in intermediate scoring steps, another technique we apply in this paper.

3. Model

In this section, we will first briefly review the model proposed in Lee et al. [1], then our techniques including 1) POS tag multi-task 2) supervision on intermediate scores, and 3) self-attention across sentences are described.

3.1 Task Definition

Following Lee et al. [6], we formulate the coreference as a set of decisions for every possible text span. Given the document, the coreference can be formulated as grouping mentions into different clusters, which each stands for one specific entity.

Let T be the length of document, then there will be $N = \frac{(T-1)T}{2}$ different text spans. The task is to assign each span i an antecedent y_i. The set of all possible assignments for y_i is $\mathcal{Y} = \{\epsilon, 1, ..., i-1\}$,

representing a dummy antecedent ϵ and all preceding spans. The dummy antecedent ϵ covers two possible scenarios: 1) the span is not an entity mention and thus has no antecedent or 2) the span is the first mention of the entity in the document. Using these decisions, we can also represent the coreference results as a clustering problem.

3.2 Model Structure

The basic structure of our model is that it first extracts features for all text spans in the document, then calculates the conditional probability distribution of the span over all possible assignments. The likelihood for each text span could be expressed as follows:

$$\begin{aligned} P(y_1, ..., y_N) &= \prod_{i=1}^{N} P(y_i|D) \\ &= \prod_{i=1}^{N} \frac{e^{f(i,y_i)}}{\sum_{y' \in \mathcal{Y}(i)} e^{f(i,y')}} \end{aligned} \tag{1}$$

where function f calculates the coreference score for each pair of text spans. However, since there are $O(T^2)$ possible text spans, if we calculate the coreference score for each pair of text spans, we need to call f $O(T^4)$ times. To alleviate this, Lee et al. [6] prunes the candidate spans greedily during both training and evaluation. Based on Lee et al. [6], Lee et al. [1] achieves better performance using two techniques: 1) higher-order inference to get better span representation, and 2) coarse-to-fine to achieve better scalability. The overview of the baseline model is shown in FIGURE 1.

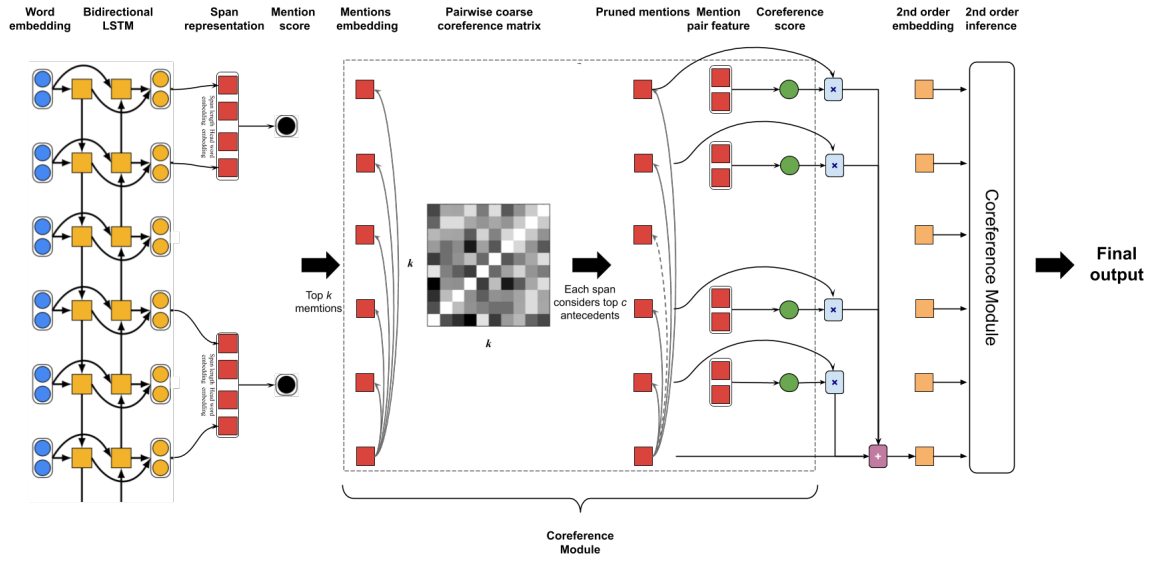

Figure 1: The baseline model overview. For illustration, we only draw the coreference score for the text span in the bottom.

The model first feeds the word embedding into a two-layer bi-directional LSTM, with context-dependent representation. For each text span, its feature vector is composed of its first and last

representation from the bi-LSTM, span length embedding, and head word representation. To get the head word representation, we use attention over all words in the text span to get weights, and sum over the word embedding using these weights.

After getting the span representation, we feed it into a multi-layer perceptron to calculate its mention score. To reduce the computational consumption, only the top M text spans are passed into our first-order coreference module. This module first calculates the coarse coreference score using a bilinear scoring function and only keeps the top K antecedents for each span. Coarse-to-fine pruning allows the model to consider long coreference links efficiently.

After coarse-to-fine pruning, we are left with MK pairs of mentions. We concatenate their feature vectors and feed them into a multi-layer perceptron to get the first-order coreference probability. Assuming g_i^n is the n-order representation for span i, we first compute its n-order coreference probability:

$$P_n(y_i) = \frac{e^{s(g_i^n, g_{y_i}^n)}}{\sum y \in \mathcal{Y}(i) e^{s(g_i^n, g_y^n)}} \tag{2}$$

where s is the coreference scoring function. The scoring function shares the same parameters across the model but is given different span representation. Having the antecedent probability $P_n(y_i)$, we then compute the $(n+1)$-order span representation by:

$$a_i^n = \sum_{y_i \in \mathcal{Y}(i)} P_n(y_i) \cdot g_{y_i}^n \tag{3}$$

$$f_i^n = \sigma(\mathbf{W}_f [g_i^n, a_i^n]) \tag{4}$$

$$g_i^{n+1} = f_i^n \circ g_i^n + (1 - f_i^n) \circ a_i^n \tag{5}$$

where a_i^n a_i^n is the expected antecedent representation for span i, and the learned gate vector f_i^n determines how much information should be extracted from its expected antecedent. After calculating the higher-order mention representation, we iteratively refine the coreference score.

3.3 POS Tag Multi-task

As discussed in the introduction, the baseline model places significant emphasis on the similarity of entity n-grams over broader context. We argue that guiding our model with linguistic features could lessen this emphasis and improve our overall performance. Following Strubell et al. [17], we use a hidden state in the bi-LSTM to predict the POS tag for each word, under the assumption that this use of multitask learning will allow the model to better learn grammatical structures and thus select better mention spans.

3.4 Supervision on Intermediate Scores

The baseline model from Lee et al. [1] only uses cross-entropy loss in the final output calculation step. However, most of the intermediate coreference scores (both coarse and fine) remain unsupervised. These intermediate scores not only have the same relation to the semantic meaning of the coreference pair as the final classification, but they also amount to the mention-wise attentions.

Similar to Strubell et al. [17], which uses dependency head to supervise the attention layers as a way to encode linguistic structure dependency, we find that the coreferent relations already provided by the dataset could naturally indicate dependencies across mentions. We thus reason that we can use the same training loss for these intermediate scores to observe better convergence speed and accuracy.

3.5 Self-attention across Sentences

In Lee et al. [6], the team argues that cross-sentence context was not helpful to the model. We note earlier that error analysis of the more successful Lee et al. [1] model suggests that a lack of broader context leads to a dependence on mention similarity and hinders the overall success of the model. It is especially counterintuitive to suggest that context across sentences provides no useful data to a coreference model, given that a significant proportion of anaphoric mentions occur in different sentences from their antecedent. Drawing from Strubell et al. [17], we propose two methods to solve this limitation.

First, we can encode information for full texts instead of individual sentences to pass to our LSTM (note that we use ELMo to generate our word embeddings). This would allow us to capture long-term dependencies without extra parameters. We can counter the increased memory load of this method by caching the full ELMo output before training.

Our second method of capturing cross-sentence contexts is to use attention in the final layer of the LSTM to obtain the representation for the previous sentences. Let $\mathbf{h}_{s,l,i}$ be the l-th layer hidden state of the i-th word of the s-th sentence. We adopt the Scaled Dot-Product Attention:

$$\mathbf{a} = (\mathrm{softmax}(\mathbf{q}^\mathsf{T}\mathbf{K}))^\mathsf{T} \longrightarrow \text{supervised} \tag{6}$$

$$\mathbf{c} = \frac{1}{\sqrt{d_h}}\mathbf{V}\mathbf{a} \tag{7}$$

$$\tag{8}$$

where d_h is the hidden size, and each query $\mathbf{q}$, key $\mathbf{k}$, and value $\mathbf{v}$ can be calculated by:

$$\mathbf{q}_{s,i} = \mathrm{MLP}_\mathbf{q}(\mathbf{h}_{s,l,i\pm1}) \tag{9}$$

$$\mathbf{k}_{s-1,j} = \mathrm{MLP}_\mathbf{k}(\mathbf{h}_{s-1,l-1,j}) \tag{10}$$

$$\mathbf{v}_{s-1,j} = \mathrm{MLP}_\mathbf{v}(\mathbf{h}_{s-1,l-1,j}) \tag{11}$$

We use $\mathbf{c}$ along with the previous sentence embedding to get an summarizing vector. Then, we concatenate this vector with the second LSTM layer hidden state, and feed into third LSTM layer.

4. Experiments

4.1 Dataset

The work done by [1] utilizes the CoNLL 2012 dataset [2]. This dataset is comprised of 3494 transcripts from 6 domains (newswire, magazine articles, broadcast news, broadcast conversations,

web data, and conversational speech data) and were hand-annotated for a variety of linguistic information. Following Lee et al. [1], we use entity mention and anaphoric coreference in the English corpus. The English corpus has three partitions: train (1.3M words, 2802 parts), development (160K words, 343 parts), and test (170K words, 348 parts).

4.2 Metrics

Lee et al. use the MUC [18], B3 [19], and CEAFe [20], three popular metrics used for coreference resolution. The F1 score for each of the three is shown in a table and averaged to create their final F1 score. We do the same in evaluating our model.

Lee et al. report achieving a test set F1 score of 73.0 [1]. When we re-ran the code they released using the best configuration, however, we were just able to achieve a score roughly at 72.4. We are uncertain what led to this discrepancy.

4.3 Settings

Following Lee et al. [1], we split the whole document into sentences, and use 3 layer 1024-dim ELMo Peters et al. [21] as our word embedding. We also try to encode the whole document to capture the long dependency, but we don't observe any improvement in the experiments. After that, A three-layer LSTM is used as encoder. After scoring all mentions, we only consider top $0.4L$ mentions. In coarse inference, we only keep top 50 antecedents for each remaining span. The model is optimized using Adam with 5e-4 initial learning rate and decay 0.999 for each 100 iterations. The code is publicly available[3].

4.4 Results

Due to hardware limitations, we are not able to train the model with the GloVe and Character CNN embedding vectors, since for each training step this document level model disallow the use of `nn.DataParallel` in PyTorch and we could just run the model in one GPU. The results is shown in TABLE 3. Our implementation without GloVe and CharCNN achieves an average F1 score of 72.4. For comparison, we remove the GloVe and CharCNN from Lee et al. [1] and re-train it, it only achieves 72.3 F1 score. By adding POS multi-task loss and self attention across sentences, we improve the F1 score by 0.1. We could further improve the F1 score to 72.6 by supervising the intermediate scores. We have not yet specifically performed any hyper-parameter tuning, so better performance is expected.

4.5 Analysis

A note we made regarding the CoNLL 2012 annotated data is that the annotations provided did not always match our own analysis of what did and did not count as entity mentions. In some cases,

[3] https://github.com/YangXuanyue/pytorch-e2e-coref

Table 3: Results on the test set on the English CoNLL-2012 shared task.

	MUC			B^3			CEAF			
Model	**Prec.**	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**	**Avg. F1**
Clark and Manning [22]	76.1	69.4	72.6	65.6	56.0	60.4	59.4	53.0	56.0	63.0
Wiseman et al. [23]	77.5	69.8	73.4	66.8	57.0	61.5	62.1	53.9	57.7	64.2
Clark and Manning [24]	79.2	70.4	74.6	69.9	58.0	63.4	63.5	55.5	59.2	65.7
Lee et al. [1]	81.4	79.5	80.4	72.2	69.5	70.8	68.2	67.1	67.6	73.0
Lee et al. [1] w/o GloVe, CharCNN	81.0	79.2	80.0	71.3	69.1	70.2	68.3	65.1	66.7	72.3
Our Reimplementation w/o GloVe, CharCNN	80.9	79.0	80.0	71.2	69.3	70.2	68.0	66.0	67.0	72.4
+ POS	81.7	78.5	80.1	72.5	68.2	70.3	69.1	64.9	66.9	72.4
+ POS, SelfAtt	81.2	79.0	80.0	71.7	68.8	70.2	69.2	65.0	67.0	72.5
+ Multi-layer Super.	81.0	79.3	80.1	71.8	69.1	70.4	68.0	66.2	67.1	72.6

the issue arose from underlying ambiguities with the task, but in others there are clear grammatical agents in the sentence that are not tagged as mentions, or outright incorrect coreference mentions. A consequence of this is some changes to the model may show little improvement in the F1 score, but may create coreference classifications that actually do make more sense to a human annotator. It is, however, difficult to capture these sorts of changes without a large scale re-annotation. We do believe we have seen improvements in our model's output that make more sense than the coreference data from the gold standard annotations, but to quantify this would require a large-scale re-annotation of the CoNLL data, which we do not have the time or resources to implement. Instead, we showcase some of the qualitative improvements we have seen through specific examples, especially those which are not reflected in the final F1 score.

Returning to the example from the introduction (see TABLE 4 for more detailed look), we noted that the Lee et al. [1]'s model failed to differentiate between two mentions of "a friend" even though the mentions clearly referred to different people. We include the table again below, this time with full coreference clusters and our own model's predictions.

As we can see in the table, our model correctly discerns that the second mention of "a friend" refers to a different entity, suggesting that the linguistic features added in our model do at least mitigate the over-reliance of Lee et al. [1]'s model on similarity of mentions. In this case, our output does align with the human annotations, and thus would improve our model's F1 score.

There are cases, however, where clear entities are ignored by both the Lee et al. [1] model and the human annotations. Here, we would argue that the improvements to our model lessen our F1 score, but bring us closer to a truly correct understanding of the data. For example, in TABLE 5, our model is the only one to recognize "one of the..." as an entity mention, despite the fact that "one of the..." is the grammatical subject of sentence and is explicitly noted to be coreferent with "Professor Zhou Hanhua". Though our model also fails to identify these mentions as coreferent, the fact that it can identify both mentions puts it a step closer to true correctness than the Lee et al. model. It is noteworthy, though, that this output is a clear result of the integration of POS tags in the mention

Table 4: Comparison analysis between baseline model and our proposed model

Speaker 1: ... how the two of you found out the news on the day of the accident? *Speaker 2*: It happened that I was going to have lunch with **a friend**, um, at noon. And then, **the friend** first sent me an SMS, uh-huh. Saying **he** would come pick me up to go together .... *Speaker 1*: And you, Yang Yang? *Speaker 3*: **A friend** happened to call me. *Gold:* [the friend, a friend, he][a] *Baseline:* [a friend, the friend, he, a friend] *Our model:* [the friend, a friend, he]

[a] The second instance of "a friend" is not coreferent with anything, so does not appear in output.

identification step–because our model has linguistic information about the grammatical structure of the sentence, it can make better inferences about where in a noun phrase the head word is located, and whether it serves a grammatical role in the sentence.

Table 5: Comparison analysis between baseline model and our proposed model

Speaker 1: Well, we have invited **two honorable guests** to the studio today to follow this topic with us. **One of the two honorable guests in the studio** is **Professor Zhou Hanhua** from the Institute of Law of the Chinese Academy of Social Sciences. *Gold:* [two honorable guests, the two honorable guests] *Baseline:* [two honorable guests, the two honorable guests in the studio] *Our model:* [two honorable guests], [one of the two honorable guests in the studio]

Our final example (TABLE 6) looks at another scenario where the Lee et al. model and the human annotation fail to capture essential entities. In this example, the human annotation completely ignores any mentions of the entity "e-government" or "e-governments". Meanwhile, despite the extreme specificity of the word, the Lee et al. model here actually chose to differentiate these two mentions. This is possibly because the rarity of the word may prevent it from having detailed embedding information. Both the Lee et al. model and our model fail to capture many mentions of the "e-government" entity, but ours does successfully identify them as coreferent. Again, because the human annotation does not identify these as mentions, our model's correct identification of the coreference will hurt our F1 score.

An additional note is that our model consistency achieved much higher precision, but lower recall with the addition of POS tagging multitask. We can attribute this to better recognition of entity mentions such that our model is more likely to correctly identify mentions from the annotations, but will also identify mentions that are not present in the annotated data (either because the mentions are not valid, or because they are valid but missed by the human annotators).

Table 6: Comparison analysis between baseline model and our proposed model

Speaker 1: Yes, when we talked about **e-government** in the past, **it** seemed to be only done through the Internet. Right. It seemed that computers, ah, and the Internet, ah, were an expression of **e-government**, to let people find out information. Well, presently in fact, **e-governments** in various countries apply **this** so - called last multiple - exposure method, such as call centers, such as electronic IDs, and such as, these electronic signatures. *Gold:* [] *Baseline:* [e-government, it],[this, e-government] *Our model:* [e-government, e-government]

5. Future work

In future work, we would plan to explore further papers featuring linguistically-minded neural innovations to see what works with our current dataset. Currently, our specific aims include improving the accuracy of the annotated dataset, as we have established that the CoNLL 2012 data misses important mention data and makes F1 scores an unreliable measure of the model's success. This might involve re-annotating the current dataset or building some new dataset with more comprehensive mention labeling. We are also curious about the impacts of adding coreference loss in the self-attention step, adding dependency parsing as a multitask learning task and as a control for self-attention (inspired by Strubell et al. [17]), and utilizing coreference annotations during the self-attention step.

6. Conclusion

Our implementation shows improvement over the Lee et al. baseline model when we account for the lack of GloVe and CharCNN embeddings in our final implementation, with Lee et al.'s model having an F1 score of 72.3 and ours with an F1 of 72.6. But we also note that the ground truth annotations in the CoNLL 2012 dataset often leave out valid entity mentions, and thus a model which captures those valid entities may score worse in a strict measurement of F1 scores. We analyze our implementation output qualitatively (see Section 4.5 Analysis) and find instances where our model captures entity coreference data missing from the ground truth annotation. Though not conclusive, this qualitative analysis suggests that our additions to the state-of-the-art model may reflect more of an improvement than is directly apparent in the F1 scores.

In particular, we note that these examples seem to reflect a greater understanding of linguistic structures in our model than in the Lee et al. baseline, and we see many of the specific issues addressed in our error analysis of Lee et al. solved by our own implementation. Our most useful additions to the baseline model, according to the F1 score, are the addition of multi-layer supervision on intermediate coreference scores, a technique inspired by Strubell et al. [17]'s use of supervision based on dependency data in the task of semantic role labeling. Our deeper error analysis, meanwhile, suggests that the integration of POS tagging yields better mention identification that is not necessarily reflected in the final F1 score. We can see there remain many facets of coreference our model cannot adequately

capture. But quantitatively and qualitatively, we demonstrate improvement over the current state-of-the-art when we integrate techniques inspired by a linguistically-focused mindset, and we believe this mindset will yield more successes on this task in the future.

7. Acknowlegements

We would like to thank Prof. Graham Neubig for his insightful instruction, Chunting for suggesting using LEA metric and considering cross-sentence dependency, and the reviwers whose invaluable feedback and insights have greatly improved the quality and clarity of this paper.

References

[1] Lee K, Luheng H, Zettlemoyer L. Higher-Order Coreference Resolution With Coarse-To-Fine Inference. In: Proceedings of the 2018 conference of the north American Chapter of the Association for Computational Linguistics: Human Language Technologies (Short papers). Vol. 2. Louisiana: Association for Computational Linguistics. New Orleans. 2018: 687-692.

[2] Pradhan S, Moschitti A, Xue N, Uryupina O, Zhang Y. Conll-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in Ontonotes. In: Joint Conference on EMNLP and CoNLL-Shared Task. Association for Computational Linguistics; 2012:1-40.

[3] Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Comput. 1997;9:1735-1780.

[4] Wang H, Liu X, Tao Y, Wenting Y, Jin Q, et al. Automatic Human-Like Mining and Constructing Reliable Genetic Association Database With Deep Reinforcement Learning. In BIOCOMPUTING 2019 : Proceedings of the Pacific symposium. 2018:112-123.

[5] Wenting Y, Yang H, Zhao S, Fang H, Shi X, et al. A Transformer-Based Substitute Recommendation Model Incorporating Weakly Supervised Customer Behavior Data. 2022. arXiv preprint: https://arxiv.org/pdf/2211.02533.pdf

[6] Lee K, He L, Lewis M, Zettlemoyer L. End-to-end neural coreference resolution. 2017. ArXiv preprint arXiv: https://arxiv.org/pdf/1707.07045.pdf

[7] Moosavi NS, Strube M. Using Linguistic Features to Improve the Generalization Capability of Neural Coreference Resolvers. In: Proceedings of the 2018 conference on empirical methods in natural language processing. 2018:193-203.

[8] Ng V. Advanced Machine Learning Models for Coreference Resolution. Anaphora resolution. 2016.

[9] Tai KS, Socher R, Manning CD. Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks. 2015. ArXiv preprint: https://arxiv.org/pdf/1503.00075.pdf

[10] Dhingra B, Yang Z, Cohen WW, Salakhutdinov R. Linguistic Knowledge as Memory for Recurrent Neural Networks.2017. ArXiv preprint: https://arxiv.org/pdf/1703.02620.pdf

[11] Zeng Y, Jin X, Guan S, Guo J, Cheng X. Event Coreference Resolution With Their Paraphrases and Argument-Aware Embeddings. In: Proceedings of the 28th international conference on computational linguistics. International Committee on Computational Linguistics; Barcelona, Spain. 2020:3084-3094.

[12] Lu J, Ng V. Span-Based Event Coreference Resolution. AAAI Conference on Artificial Intelligence. 2021;35:13489-13497.

[13] Lu J, Ng V. Conundrums in Event Coreference Resolution: Making Sense of the State of the Art. In: Proceedings of the 2021 conference on empirical methods in natural language processing. Association for Computational Linguistics. Online and Punta Cana, Dominican Republic. 2021. 1368-1380.

[14] Lu J, Ng V. 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; 2021:4504-4514.

[15] Lu Y, Lin H, Tang J, Han X, Sun L. End-To-End Neural Event Coreference Resolution. Artif Intell. 2022;303:103632

[16] Grosz BJ, Weinstein S, Joshi AK. Centering: A Framework for Modeling the Local Coherence of Discourse. Comp Linguist. 1995;21:203-225.

[17] Strubell E, Verga P, Andor D, Weiss D, McCallum A. Linguistically Informed Self-Attention for Semantic Role Labeling. 2018. ArXiv preprint : https://arxiv.org/pdf/1804.08199.pdf

[18] Vilain M, Burger J, Aberdeen J, Connolly D, Hirschman L. A model-theoretic coreference scoring scheme. In: Proceedings of the 6th conference on message understanding. Association for Computational Linguistics. 1995:45-52.

[19] Bagga A, Baldwin B. Algorithms for Scoring Coreference Chains. In The first international conference on language resources and evaluation workshop on linguistics coreference. 1998;1:563-566.

[20] Luo X. On Coreference Resolution Performance Metrics. In: Proceedings of the conference on human language technology and empirical methods in natural language processing. Association for Computational Linguistics; 2005:25-32.

[21] Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, et al. Deep Contextualized Word Representations. 2018. ArXiv preprint: https://arxiv.org/pdf/1802.05365.pdf

[22] Clark K, Manning CD. Entity-Centric Coreference Resolution With Model Stacking. In: Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint Conference on Natural Language Processing. China: Association for Computational Linguistics; 2015;1:1405-1415.

[23] Wiseman S, Rush AM, Shieber SM. Learning Global Features for Coreference Resolution. In: Proceedings of the 2016 conference of the north American Chapter of the Association for Computational Linguistics: Human Language Technologies. CA: Association for Computational Linguistics. San Diego. 2016:994-1004.

[24] Clark K, Manning CD. Deep Reinforcement Learning for Mention-Ranking Coreference Models. In: Proceedings of the 2016 conference on empirical methods in natural language processing. TX: Austin. Association for Computational Linguistics; 2016:2256-2262.