NLP Project - Coreference Resolution in Hindi and English

Vamshi Krishna Bonagiri
2020114011
IIIT Hyderabad

Harshit Gupta
2020114017
IIIT Hyderabad

Vanshpreet S Kohli
2020114014
IIIT Hyderabad

Abstract

Coreference resolution is the task of finding mentions which refer to the same real-world entity. Although coreference resolution is a very important and common task in NLP, most of the previous work in it has been towards hand crafted features. Along with that, most of the work in coreference resolution has been in English, with little to none work for Hindi. In this work, we explore neural-network based methods on English and Hindi datasets. We also try to leverage BERT to produce contextual word embeddings which boost the performance of the models. We also create a baseline model for neural coreference resolution in Hindi.

1. Introduction

In linguistics, coreference occurs when two or more expressions refer to the same entity; they have the same referent - i.e they co-refer. Coreference resolution is the task of finding all expressions that co-refer in a text. It is an important step for a lot of higher level NLP tasks that involve natural language understanding, such as document summarization, question answering, and information extraction. Coreference resolution is non-trivial, and hence its automation is a deep NLP problem.

Although coreference resolution is a core NLP problem, most of the previous work dedicated to it has been focused on hand crafted complex features related to syntactic, semantic and discourse level information. Infact, a lot of them are completely rule based. Like the recent methods [1], we try to use deep learning methods to solve coreference resolution. There are almost no coreference resolution methods using neural networks when it comes to the Hindi Language, so we tried to extend our work in English to a Hindi dataset.

We used Mention Pair models for the task as they offer a very simple solution, as the task now reduces to looking at two mentions and answering the yes/no question "are they coreferent?". The mention pair model essentially consists of a word embedding layer, out of which mentions are paired and sent to a binary classifier which answers the yer or no question. We create a baseline by using pretrained Glove embeddings to represent mentions, along with a naive way to represent their contexts. We improve this model by using BERT embeddings, which produce contextual information which is very essential for coreference resolution. The fact that BERT uses attention is another major advantage for coreference resolution. We also try to make the problem a ternary classification one to specifically run on the GAP dataset which has gender ambiguous pronouns.

We tried to extend the same ideas to resolving coreference in Hindi. We used BERT embeddings followed by the binary classifier which seemed to work really well. We then tried to improve this model by adding an LSTM layer to capture more context, which did not work very well as it was very prone to overfit.

2. Related Work

For reading up on previous work in the field, we referred to Sukthanker et al's summary [2] of works in coreference resolution.

Clark (2015) has done pioneering work [1] in coreference resolution using deep learning that automatically learns dense vector representations for mention pairs for English and Chinese. He built them using the word embeddings in the mention and surrounding context, which will maintain the semantic similarity. Despite using a few hand-engineered features, he trained an incremental coreference system that can utilize entity-level information. His mention pair model acted as an inspiration for our feature representations, and we updated it for free word order languages. In free word order languages, despite changing the order of words in a sentence the overall meaning of the sentence will not change.

We refer to Tenney et al's work [3] to understand the importance of contextual word representations and how they can help us out with our task. This served as a very important reference since our model's performed really well with just the addition of contextual word vectors from BERT.

For Coreference Resolution, we refer to Mandar Joshi, Omer Levy, Daniel S. Weld, [4] and Luke Zettlemoyer's work [5] on BERT. They examine and contrast the findings and analyses of pretrained BERT Models on the GAP and OntoNotes datasets. To model long-range relationships more successfully, BERT employs pretraining on passage-level sequences (in combination with a bidirectional masked language modelling aim).

Radhika's work in Anaphora resolution for South Asian Languages [6] has shown amazing results for a low researched language such as Telugu, which inspired us to go for a similar approach but for Hindi. It is to be noted that they obtained good results due to the dataset being very specific and them using hand-crafted features, which we tried to avoid as we only used deep learning approaches.

We do not talk much about purely rule based approaches like the Hobbs Algorithm in our work since we only focused on deep learning based approaches. However, it is to be noted that Hobbs' work [7] is considered one of the very first and most appreciated methods for Anaphora Resolution due to its simple and intuitive nature.

3. Datasets Used

- **Hindi Coreference Annotated Data.** Dataset [8] from our very own FC Kohli Center on Intelligent Systems (KCIS), IIIT-H, India. We tried applying the concepts learned and our models on the dataset and tried creating our very own baseline for the same.
- **GAP Coreference Dataset.** GAP [9] is a gender-balanced dataset with 8,908 coreference-labeled pairs of (ambiguous pronoun, antecedent name) taken from Wikipedia and distributed by Google AI Language for testing coreference resolution in practical applications.

4. Models

4.1. Baseline

The first and simplest model we tried involved a binary classifier with pretrained Glove word embeddings as inputs. After obtaining the pretrained Glove embeddings for each word in the data, for each mention, their neighboring word embeddings are taken as context and concatenated with the mention embedding. Mention pairs are prepared accross the data, and their concatenated embeddings are sent to a neural network with three hidden layers of ReLU units ending with a sigmoid. The job of this model is to predict if the mention pair passed is corefferent or not, and is trained in the same way.

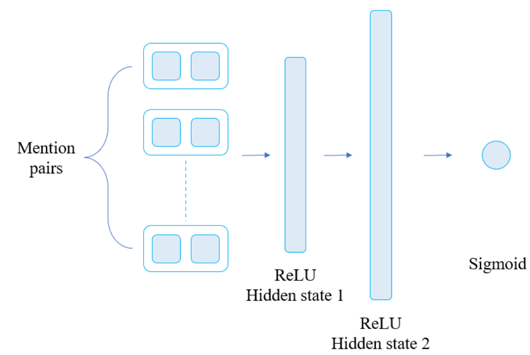

Figure 1. Binary Classifier with 3 ReLU layers

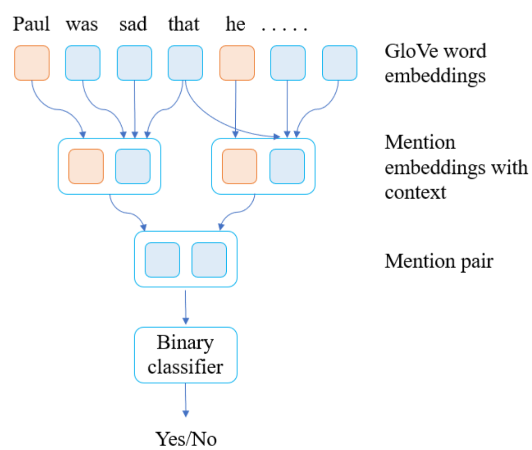

Figure 2. Baseline Model

Results:
Accuracy: 0.5753
Precision: 0.5250
Recall: 0.4377
Specificity: 0.6848
F1 score: 0.4774
Confusion Matrix:

$$\begin{bmatrix} 776 & 702 \\ 997 & 1525 \end{bmatrix}$$

We see that the model mostly takes random guesses, which is really bad. This is due to the oversimplified word representations.

4.2. Using Bert word embeddings

Since we learnt that contextual word embeddings can contribute a lot to coreference resolution from [3], we used BERT embeddings instead of Glove embeddings to represent the data. Since this generates contextual word vectors, we do not need to go through the trouble of creating additional contextual embeddings. Many features such as Gender or Number which are usually hand-crafted (like in Clark's work) will be encoded in the BERT embeddings which greatly improves the model over baseline. The model no longer takes random guesses and gives us a decent F1 score of 0.728.

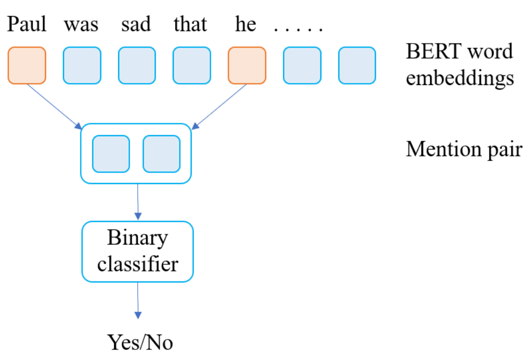

Figure 3. Using BERT word embeddings

Using English bert embeddings on GAP dataset, we get the following results:
Accuracy: 0.7274
Precision: 0.6927
Recall: 0.7056
Specificity: 0.7450
F1 score: 0.6991
Confusion Matrix:

$$\begin{bmatrix} 1251 & 555 \\ 522 & 1672 \end{bmatrix}$$

4.3. Using Bert word embeddings for GAP

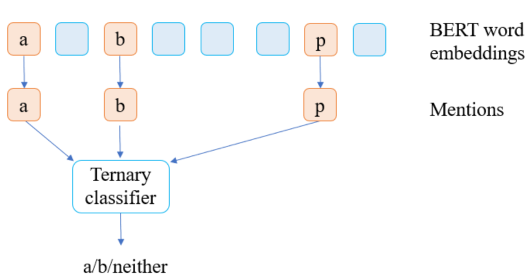

Figure 4. Using BERT word embeddings for the GAP dataset

This model was specifically made to work with the GAP dataset. Due to the structure of the GAP dataset, we experimented by changing the binary classifier to a ternary classifier. After generating BERT embeddings, we pass a, b, pronoun through the classifier which tells us if the pronoun is referring to a, b or neither of them. We see an improvement in the scores simply by changing the number of classes in the output, while the core idea of the model remains the same.

Results:
Accuracy: 0.830500
Precision: 0.830590
Recall: 0.830500
F1 score: 0.828175
Confusion Matrix:

$$\begin{bmatrix} 133 & 40 & 54 \\ 23 & 765 & 130 \\ 20 & 72 & 763 \end{bmatrix}$$

4.4. MuRIL + Hindi

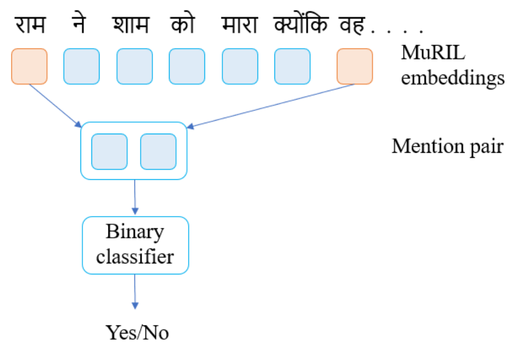

Figure 5. Using MuRIL word embeddings for Hindi

We use the MuRIL(Multilingual Representations for Indian Languages) [10] word embeddings instead of BERT as it has better word representations and slightly outperforms mBERT. Using the same method as English Bert and using MuRIL has shown decent results for the Hindi Dataset.

results:
Accuracy: 0.8558
Precision: 0.2949
Recall: 0.6754
Specificity: 0.8703
F1 score: 0.4106
Confusion Matrix:

$$\begin{bmatrix} 1165 & 2785 \\ 561 & 18689 \end{bmatrix}$$

4.5. MuRIL + Hindi + LSTM

We tried to expand on the previous model by passing the word embeddings through an LSTM layer to capture more context. Despite being computationally complex, we did not see an improvement in scores. We noticed that this was because the LSTM layer caused overfitting.

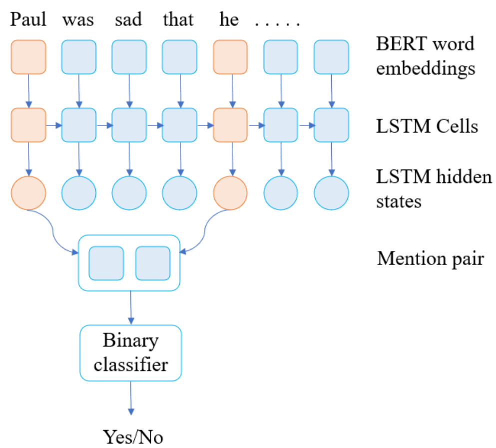

Figure 6. Model with LSTM utilised

Results:
Accuracy: 0.8949
Precision: 0.4438
Recall: 0.5809
Specificity: 0.9266
F1 score: 0.5031
Confusion Matrix:
$$\begin{bmatrix} 1002 & 1256 \\ 723 & 15852 \end{bmatrix}$$

5. Results

Model	Accuracy	Precision	Recall	F1 Score
Model 1	0.5753	0.5250	0.4377	0.4774
Model 2	0.7274	0.6927	0.7056	0.6991
Model 3	0.8900	0.4464	0.733	0.5549
Model 4	0.830500	0.830590	0.830500	0.828175
Model 5	0.8949	0.4438	0.5809	0.5031

Model 1: English Baseline
Model 2: English + BERT Embeddings
Model 3: Hindi + MURIL Embeddings
Model 4: English + BERT Model (GAP Dataset)
Model 5: Hindi + MURIL Embeddings + LSTM

We see that there is a huge jump in results (model 1 to 2) after using contextual word embeddings. The increase in scores from Model 2 to 4 is simply due to the structure of the data, we aimed to produce better results for the GAP dataset which is the reason for treating it as a multiclass problem. Finally, we see how MURIL + LSTM model lead to huge overfitting, which was the reason for the bad results in this model.

6. Challenges Faced

Some the major challenge we faced are:

- Coreference Resolution has very less number of deep learning approaches, which made it incredibly challenging to find resources which could help us out.
- Lots of experimentation had to be done due to the reason mentioned above, and it was very challenging to get to a working solution with very less resources.
- There were no pre-exisiting deep learning methods implemented on the Hindi Dataset. As a result of which, we had to implement everything from scratch.
- Using pretrained BERT models and ensuring that the RAM of the system doesn't seize was extremely important. We had to tune the siz of the batches, the frequency of back propagation and the size of the dataset.

7. Conclusion

We explore the task of Coreference Resolution using Mention Pair models with contextual word embeddings and Deep Learning Approaches. We also see how attention and contextual word embeddings can help in coreference resolution methods. We also propose a baseline model for neural Hindi Coreference Resolution task, which has an F1 score of 0.85. Future work can be done by using Mention Ranking and Clustering methods instead of Mention Pair models.

References

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

[2] Rhea Sukthanker, Soujanya Poria, E. Cambria, and Ramkumar Thirunavukarasu. Anaphora and coreference resolution: A review. *ArXiv*, abs/1805.11824, 2020. 1

[3] Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*, 2019. 2, 3

[4] Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. BERT for coreference resolution: Baselines and analysis. 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 5803–5808, Hong Kong, China, November 2019. Association for Computational Linguistics. 2

[5] Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. End-to-end neural coreference resolution. *CoRR*, abs/1707.07045, 2017. 2

[6] Vinay Annam, Nikhil Koditala, and Radhika Mamidi. Anaphora resolution in dialogue systems for south asian languages. *CoRR*, abs/1911.09994, 2019. 2

[7] J.R. Hobbs. Resolving pronoun references. *Lingua 44*, pages 311–338, 1978. 2

[8] Vandan Mujadia, Palash Gupta, and Dipti Misra Sharma. Coreference annotation scheme and relation types for Hindi. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 161–168, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). 2

[9] Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. Mind the gap: A balanced corpus of gendered ambiguou. In *Transactions of the ACL*, page to appear, 2018. 2

[10] Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha P. Talukdar. Muril: Multilingual representations for indian languages. *CoRR*, abs/2103.10730, 2021. 3