Anaphora Resolution from Social Media Text

Vijay Kumari , Shaz Furniturewala, Gautam Bhambhani and Yashvardhan Sharma

Department of Computer Science and Information Systems, Birla Institute of Technology and Science Pilani, Pilani, Rajasthan

Abstract
Anaphora resolution for social media texts is essential yet difficult task for text understanding. An important characteristic of anaphora is that it creates a connection between the antecedent and the anaphor buried in the anaphoric sentence. This paper outlines the methods used to locate anaphora and their antecedents in a particular text. The text is a social media tweet for the SocAnaRes-IL 2022 challenge that was part of FIRE 2022. The proposed model uses a Neural Co-reference Network for the anaphora resolution.

Keywords
Co-reference Resolution, Anaphora Resolution, Natural Language Processing, Neural Co-reference,

1. Introduction

The process of locating all expressions in a text which connects the same object is known as coreference resolution. It is crucial for many Natural Language Processing activities, including information extraction, question-answering, and document summarization, which all require understanding natural language. With the non stopping uprise of numerous social media platforms like twitter, facebook etc. which are used for mass communication, there is an exponential increase in the data (mostly textual in nature) generated and shared by its users. Processing and extracting information from such is a new challenge for modern world NLP.

This work focuses on the task of FIRE 2022 for Anaphora Resolution of the Facebook post and tweets [1]. These tweets are generally very short, 71-100 characters in length, hence they lack sufficient context to determine an antecedent of an anaphor without the aid of background or world knowledge. Another challenge is in at least 20% of these text, the antecedent is not mentioned in the current tweet, it is either in posts which was already said a day before or it is understood with world knowledge.

The proposed model is developed using the Neural Co-reference Network [2] for the "Anaphora Resolution from Social Media Text (SocAnaRes-IL)" task to resolve the anaphora for english language. We generated a tokenized list of each tweet with new indices after eliminating tokens that were unnecessary and added word tokens to instances where tokenization had been done incorrectly. As a result, the data were more suitable for the proposed model. The

Forum for Information Retrieval Evaluation 2022, December 9-13, 2022, India

p20190065@pilani.bits-pilani.ac.in (V. K.); f20200025@pilani.bits-pilani.ac.in (S. Furniturewala); f20190101@pilani.bits-pilani.ac.in (G. Bhambhani); yash@pilani.bits-pilani.ac.in (Y. Sharma)
https://www.bits-pilani.ac.in/pilani/yash/profile (Y. Sharma)

developed model is able to provide the correct resolution for 73 anaphors by attempting on 285 on anaphors.

2. Related Work

The tasks of co-reference resolution [3], textual entailment [4], learning textual similarity [5], and discourse relation sense classification [6] have all demonstrated the huge success of neural techniques. The shell noun dataset and a part of ARRAU that has pronominal abstract anaphora in any form are used to train a neural mention-ranking model for the resolution of unrestricted abstract anaphora [7]. On the shell noun dataset, the model achieves state-of-the-art results for the unrestricted abstract anaphora resolution task. For pronominal anaphors, the model is still inadequate. The findings imply that frameworks for pronominal anaphors and nominal anaphors should be learnt separately. Using syntactic information, the model can choose candidates that make sense, but it can't tell the difference between candidates with the same type of syntactic information. On the other hand, if the model doesn't get syntactic information, it learns deeper features that help it choose the right antecedent without limiting the number of possibilities. Therefore, the model must be compelled to initially choose suitable candidates before continuing to learn features to separate them using a bigger training dataset to increase performance. A model can be developed that selects candidates from the broader context as well as sentences that contain the antecedent [7].

The pairwise model has been used in the majority of earlier work on bridge anaphora resolution, which make the assumption that the gold mention information is available [8, 9]. Without knowing any gold mention information, the antecedent for a given anaphor can be determined by bridging anaphora resolution as question-answering based on context [10]. In order to produce a significant amount of "quasi-bridging" training data, the developed question-answering architecture makes use of transfer learning and a cutting-edge technique.

3. Dataset

The training dataset provided by FIRE 2022 [1] was provided in 4 different languages, English, Hindi, Malyalam and Tamil respectively. Each of them contained about 100-700 text documents. Each document contained a tweet or a series of tweets. The structure of document was:

Each document was structured as a tab-delimited value file with 3 columns-

1. The first column consists of the words/tokens.
2. The second column consists of the Markables. (Here the markables are the anaphors and the antecedents).
3. The third column consists of the antecedent markable id.

This tab-delimited value file format was unsuitable for NLP tasks like entity recognition and vectorization. Since it was generated from tweet data it also contained unusable information like tweet id and in some cases had not been cleaned. There were rows in the documents that contained entire sentences with words separated by '\n' tokens. The data had to be pre-processed first into a more usable format and the structure of the table had to be disrupted to accommodate the denoising and new tokenization. We generated a tokenized list of each tweet with new indices due to the removal of unnecessary tokens and addition of word tokens in the samples where tokenization was incorrectly done. This made the data more suited for our proposed model.

4. Proposed Technique

We utilised a statistical model pretrained on the English language as part of the NeuralCoref network by huggingface. The first step of the process is to extract words or phrases that refer to real entities. These include proper nouns like names and objects or possessives like 'My brother'. The model then trains a set of features for each of these entities or 'mentions'. This is done by taking an initial set of word embeddings and training them on the OntoNotes Corpus, a large manually-annotated corpus used for coreference resolution tasks. This method of learning features helps with coreference resolution tasks by segregating word vectors along attributes like gender, which is helpful to the model when identifying antecedents for people's pronouns. However, it does not do a great job with coreference involving things and pronouns such as 'it' or demonstrative pronouns like 'this' and 'that'. To address these issues, the feature vector will have to take into account contextual information surrounding the entity word or phrase. This contextual information is added by averaging the feature vectors of words present around the entity word and adding some other features based on phrase length, word location, speaker information, etc. Once these vectors are ready the model train two separate neural networks to identify antecedents. The first is a classification network that scores each 'mention' to classify whether it is the first instance of an entity or whether it has an antecedent. The second neural networks generates a score for a 'mention' and every possible antecedent in the text to identify the most likely pair. The scores generated by both networks are non-probabilistic max-margin

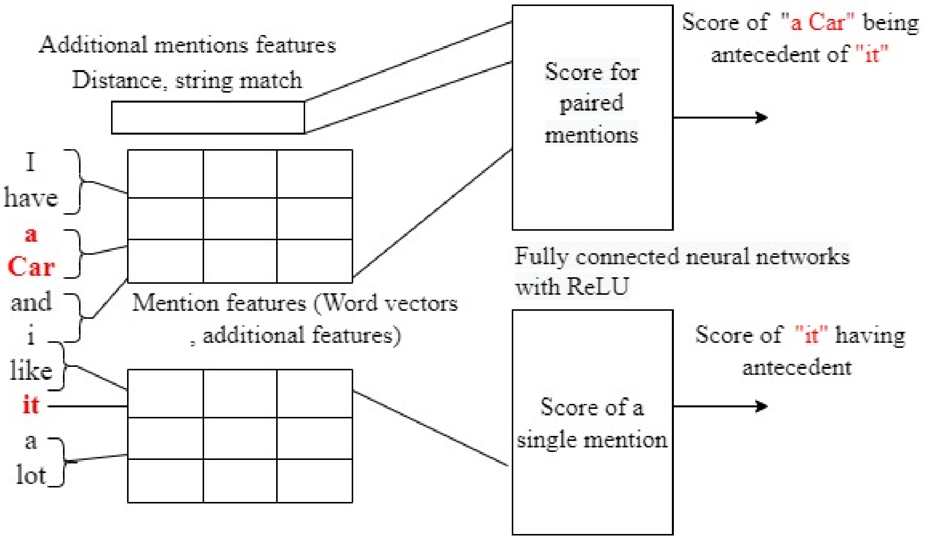
Additional mentions features
Distance, string match
I
have
a
Car
and
i
like
it
a
lot
Mention features (Word vectors , additional features)
Score for paired mentions
Score of "a Car" being antecedent of "it"
Fully connected neural networks with ReLU
Score of a single mention
Score of "it" having antecedent

Figure 1: Model for Co-reference Scoring

objective scores. Once both these models are trained, the representations for each mention are plugged into them and based on their scores the most probably antecedent is determined. Figure 1 shows the mechanisum for scoring the Co-references using neural network architecture [2].

5. Results and Evaluation

Table 1 shows the results obtained using the Neural Co-reference model [2] for anaphora resolution. The results can be improved as in many documents the 3rd Person Pronominals were not resolved correctly. For the Noun-Noun coreference resolution, the coreference chain was not completely identified. There were 100 documents altogether for the evaluation, with a total of 575 pronouns in the anaphora. The proposed system did not give output for 42 documents. The model correctly resolved 73 anaphors by attempting on 285 anaphors. The computations for precision and recall are as follows:

$$Precision(P) = 73/285 \, Recall(R) = 73/575 \tag{1}$$

Table 1
Results from the SocAnaRes-IL dataset utilising the neural co-reference network for English anaphora resolution

Precision	Recall	Correctly Resolved Anaphors	System Attempted Anaphors
25.61%	12.69 %	73	285

The Result from the Table 1 can be improved by applying several pre-processing procedures, such as removing unnecessary number strings and precise tokenization for '\n' recognition in tweets.

6. Conclusion and Future Work

In our three-step model, the first step of entity detection has minimal scope for improvement. However, the feature generation step could be further built upon. Neuralcoref uses contextual information surrounding the entity word to improve its representation vector. This includes averaging the vectors for the surrounding words as well as taking integer representations of factors like speaker and location. This can be further improved upon by using different methods to add context to the feature representation. With a larger dataset, the third step of classification neural networks could also be further improved by adding layers of complexity. With the current dataset size, using large and complex language models doesn't yield results but that is liable to change as the number of datapoints increase.

The current model is operating on a small dataset with requirement of several pre-processing steps like removal of unnecessary numerical strings and accurate tokenization due for detection of '\n' in the tweets. A thorough pre-processing of the dataset will automatically lead to a jump in accuracy of the model.

References

[1] Anaphora resolution from social media text in indian languages (socanares-il 2022), http://78.46.86.133/SocAnaRes-IL22/ (20212).

[2] Neuralcoref 4.0: Coreference resolution in spacy with neural networks, https://github.com/huggingface/neuralcoref (2021).

[3] K. Clark, C. D. Manning, Deep reinforcement learning for mention-ranking coreference models, arXiv preprint arXiv:1609.08667 (2016).

[4] S. R. Bowman, J. Gauthier, A. Rastogi, R. Gupta, C. D. Manning, C. Potts, A fast unified model for parsing and sentence understanding, arXiv preprint arXiv:1603.06021 (2016).

[5] J. Mueller, A. Thyagarajan, Siamese recurrent architectures for learning sentence similarity, in: Proceedings of the AAAI conference on artificial intelligence, volume 30, 2016.

[6] A. Rutherford, V. Demberg, N. Xue, A systematic study of neural discourse models for implicit discourse relation, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, 2017, pp. 281–291.

[7] A. Marasović, L. Born, J. Opitz, A. Frank, A mention-ranking model for abstract anaphora resolution, arXiv preprint arXiv:1706.02256 (2017).

[8] M. Poesio, R. Mehta, A. Maroudas, J. Hitzeman, Learning to resolve bridging references, in: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04), 2004, pp. 143–150.

[9] Y. Hou, A deterministic algorithm for bridging anaphora resolution, arXiv preprint arXiv:1811.05721 (2018).

[10] Y. Hou, Bridging anaphora resolution as question answering, arXiv preprint arXiv:2004.07898 (2020).