Anaphora Resolution Analysis using simplified CODI-CRAC 2021 Dataset

Anuja Jyothi Vijayakumar
Student ID: 190211413
Project Supervisor: Dr.Massimo Poesio
MSc Big Data Science, QMUL

Abstract—

CODI-CRAC 2021 Shared-Task involves the three tracks: anaphoric identity, bridging references, and discourse deixis/abstract anaphora. This paper focuses on exploring the resolution of anaphoric identity. In this work we tried to implement the BiLSTM LSTM and BiLSTM GRU models and explore how the prediction happens on the datasets around different domains using different activation functions. The process of resolving repeated usage of an entity in a dialogue or document (in varying form, e.g., pronoun) is Anaphora Resolution. It is mainly used to interpret the human language both in written and spoken linguistics. Numerous techniques and methodologies have been emerged while researching anaphora resolutions. However, the issue regarding the interpretation prevails longer than we imagine. The accuracy of automatic anaphora resolution is relatively slow. To overcome all these deficiency in anaphora resolution, we are going to implement the bidirectional LSTM and bidirectional GRU and tried to study the precision and the performance evaluated for the anaphora resolution. Being bidirectional enable to handle the sequence classification problem effectively without compromising its accuracy of the model. This model has forward and backward LSTM/GRU. The existing issues commonly occurs because the model only look at one side either forward or backward direction in a sentence. Knowing antecedent and precedent words in a sentence is important. Thus, the proposed model process through the embedding layer, Forward LSTM/GRU, Backward LSTM/GRU and concatenate and flatten and effectively tackled the shortcomings in the Anaphora resolution.

Keywords— Long short-term memory (LSTM), Gated Recurrent Unit (GRU), Anaphora resolution (AR), Bidirectional LSTM (BiLSTM), Coreference Resolution (CR), Natural Language processing (NLP), Bidirectional GRU (BiGRU)

I. INTRODUCTION

The natural language processing community has long acknowledged the importance of coreference resolution for entity/event recognition tasks. The issue of a referring pronoun or a noun in a sentence is dealt with by Anaphora Resolution. Pronoun usage in a sentence is simple for humans to figure out. Anaphora resolution was used to prevent this behaviour from entering the system. Language representation models are used to resolve anaphora and co-reference resolutions in several studies. In linguistics, anaphora is the utilization of an expression whose understanding relies on another sentence in setting. In a smaller sense, anaphora relies explicitly on a precursor articulation. Hence, it is diverged from cataphora, which are mentioned before their referents. “The anaphoric (alluding) term is called an anaphor” [1]. There are various anaphora resolution techniques like Rule Based, Corpus Based, Discourse Based, Knowledge-poor and Hybrid approaches [2]

The hinderances in the progress of anaphoric resolution involves the size of the corpora, low consistency, evaluation complexity and the knowledge bottleneck [3]

The essential objective of our annotation exertion is to explore how the models can be generalized and to what extend when working with the datasets of different domains. It will be an interesting research study. We run a series of experiments to test the attainability, and afterward analyze the evaluation metrics to calculate the precision, recall and F-Score. There have been new activation functions recently introduced. Those activation functions are also compared here.

II. RELATED WORK

The Chinese Natural Language Interface for Navigation in Mobile GIS was investigated by Feng, [4]. On the intelligent level of mobile GIS, both "Mobile GIS" and "Voice innovation" have worked together. Due to the subject of scientific research, natural language instructions can be converted to GIS orders quickly. This research examines natural language words with geographic information systems (GIS) in order to better comprehend the technique for mobile voice GIS. Utilizing a machine learning technique made a change between natural language and GIS orders.

Sontakke [5], proposed techniques to access the database without knowledge of SQL. This advancement was suggested to contribute to Natural Language Interface to Database (NLIDB). Here the user need not have to know programming language in order to access the database. They can use their spoken language (in this case, Hindi) to query the database and the results are provided in the same language.

Poesio [6] investigated the framework's design using two calculations for depictions and pronoun resolution that were performed in the current adaptation of the framework. This study uses a tool called GUITAR (General Tool for Anaphora Resolution). They briefly look at the framework's design and implementation, as well as some preliminary assessment outcomes.

Steinberger [7] proposed and discussed the use of anaphoric data in latent semantic analysis (LSA). Another aspect of our own anaphora goal framework, GUITAR, which integrates formal person, place, or object goals, is anaphoric information. The consistency of the outline supplied by our summarizer is checked using anaphoric information.

In 2012, Multilingual coreference resolution in OntoNotes was done as a joined shared task [3]. It involved languages such as English, Chinese and Arabic on the large-scale corpus of OntoNotes. The results of the participating systems were discussed there along with the learning frameworks and markable identifications.

A gender balanced labeled corpus was released to address the issues of gender bias in the coreference resolution where the systems were favoring masculine entities. [8]

Research targeted at increasing anaphora resolution systems' performance by obtaining the commonsense knowledge required to handle more complex examples of anaphora, such as bridging references. The focus was mainly on the problem of obtaining information about part-of relations. [9]

III. Anaphora Resolution

Before AR is an intra-etymological phrasing, which implies that it alludes to re-tackling references utilized inside the content with an equivalent sense (i.e., alluding to a similar element). Likewise, these elements are typically present in the content and, subsequently, the need of world information is negligible. Coreference (CR) has an extensive degree and used as an extra-phonetic wording in most phrases. Co-referential terms could have unique syntactic construction and capacity (e.g., sex and grammatical form) but then, by definition, they could allude to a similar extra phonetic element. Here, substance could be a solitary item in a world or a gathering of articles, which together structure another single element. CR treats elements in a manner more as if how we get talk, i.e., by regarding every element as a special element continuously [10].

The process of understanding the antecedent and anaphora in a sentence in any language is important to implement the anaphora resolution to the proposed model. For instance: *"Rose works for a law firm where she is working on a criminal case"* In this sentence, 'she' is referring to name 'Rose'. This formation known as anaphor. In addition, the name 'Rose' indicate as an antecedent. To determine the anaphor and antecedent in a sentence is identified as an Anaphora Resolution [11] [12].

AR classified into two categories. They are intra-sentential and Inter-sentential. In the intra-sentential, the antecedent and the anaphor lies in a same sentence. However, the Inter-sentential has antecedent and anaphor placed in different phrases [2].

There are different types of references in the language used to resolve the anaphor and antecedent. Therefore, the AR model has to address all kinds of reference and type used in the concerned language. The following list some of the types of anaphora commonly found.

A. Zero Anaphora:

Fillmore [14] is the one who initially notices about the Zero anaphora. It is nothing but usage of more than two antecedents in a sentence. They are 'invisible' anaphors. They are understood even when omitted.

B. Demonstratives:

Dixon [15] figured out that in a sentence, the antecedent word refers to an anaphor that has been already used in earlier sentence. Refers to complex propositions often.

C. Presuppositions:

In this type, the sentence has singular indefinite pronoun. It may create a bit complexity on identifying its antecedent work in a particular sentence [16].

D. Discontinuous Sets:

Clause splitting in a sentence is the main cause of split anaphor. More than one antecedent refers to the anaphor [17].

E. Pronominal Anaphora:

The most common words in a sentence creates ambiguity while referring the relation or unique entity of the antecedent. It has four different types one anaphora, indefinite pronominal, definite pronominal and adjective pronominal.

F. Cataphora:

In a sentence, initially anaphor appears and then it followed by antecedent, and this formation referred as cataphora.

IV. Existing Models

There are four main models that are popular these recent years. They include the following:

A. Rule-Based Models

A naïve algorithm on coreference resolution was introduced.

B. Mention-Pair Models

A binary classifier was introduced. It finds the probability of the mentions and compare the probability and assign the probability to be connected if they appear in a sentence. Thus, the negative examples will have values close to 0 and the positive to 1.

C. Mention-Ranking Models

A ranking is given to each pair of antecedent and coreference candidate. Only one of the pairs is chosen at the end of the model execution. Dealing with singletons is the difficulty with this model and therefore a dummy NA mention is used to leave the singletons alone.

D. Clustering-Based Models

A clustering algorithm is used here. Starting with singleton clusters, the mentions are merged with pair of clusters representing same entity.

V. Bidirectional LSTM

A Bidirectional LSTM, or BiLSTM, is a sequence-processing model that comprises of two LSTMs: one taking the contribution to a forward course, and the other a regressive way. BiLSTM viably increment the measure of data accessible to the organization, further developing the setting accessible to the calculation (for example realizing what words promptly follow and go before a word in a sentence). In Bidirectional Recurrent neural network (RNN) [18], Recurrent neural networks recollect the succession of the data and use data examples to give the expectation. RNN utilizes criticism loops, which makes it not quite the same as other neural networks. Those loops assist RNN with handling the arrangement of the data. This loop permits the data sharing to various hubs and forecasts as indicated by the assembled data. This interaction can be called memory.

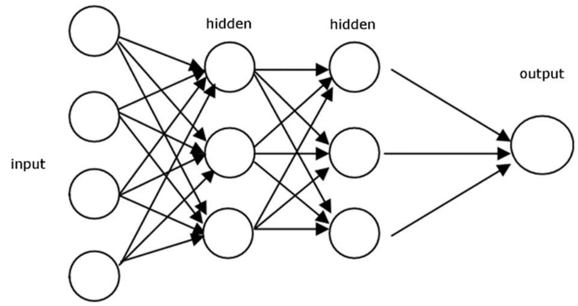

Fig 1: Recurrent Neural Network

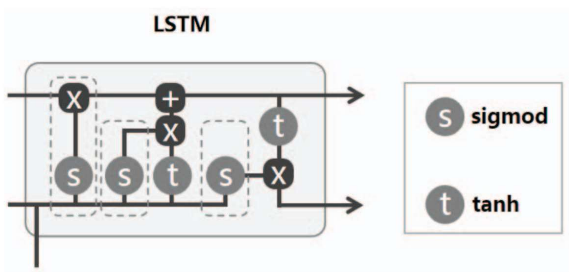

Fig 2: Internal Structure of LSTM

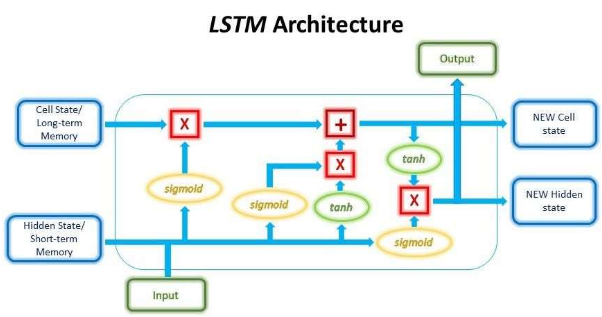

Fig 3: Inner workings of the LSTM cell

RNN and the loops make the networks that permit RNN to share data, and furthermore, the loop structure permits the neural network to take the succession of input data. RNN changes an independent variable over to a dependent variable for its next layer.

Bidirectional long-short term memory (bi-LSTM) is the way toward making any neural network have the grouping data in the two ways in reverse (future to past) or forward (past to future). In bidirectional, our feedback streams in two ways, making a bi-LSTM not quite the same as the customary LSTM. With the customary LSTM, we can make input stream one way, either in reverse or in forward. Nonetheless, in bi-directional, we can make the info stream in the two ways to save the future and the previous data.

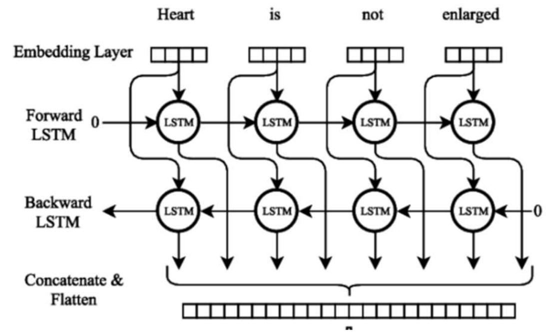

Fig 4: Bidirectional LSTM Architecture

The proposed model process through the embedding layer, Forward LSTM, Backward LSTM, concatenate, and flatten the input to generate the solution. Initially the dataset uses keras and bidirectional LSTM for classification of data. Then convert the classified data into sequence matrix to make it mounted into the neural network. Generate the model for BiLSTM layer and pre-processed data processed in the layer. Subsequently, train our model with epochs and plot the model performance.

VI. Bidirectional GRU

The GRU is a better version of the RNN, which was presented to overcome the vanishing gradient problem. It indeed uses two gates: update and reset. In essence, there are two vectors that determine what data should be sent to the output. They are unique in that they can be trained to retain knowledge from the past without having to clean it away over time or delete information that is unrelated to the forecast. The update gate assists the model in determining how much historical data (from earlier time steps) should be passed on to the future. The reset gate is utilised by the model to determine how much of the previous information should be forgotten. Thus GRU can store and filter information using the two gates.

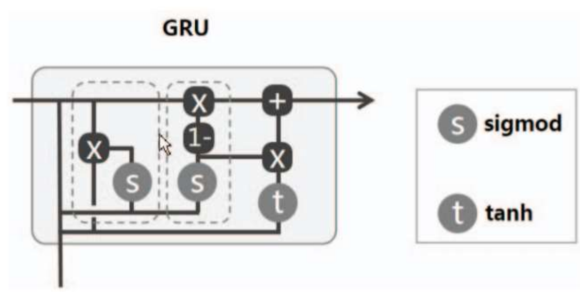

Fig 5: Internal Structure of GRU

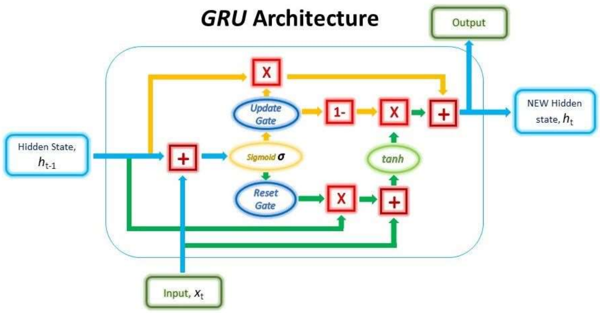

Fig 6: Inner workings of the GRU cell

By applying Bidirectional to the GRU we are making it two ways just like it's done for BiLSTM.

VII. GRU VS LSTM

A. Structural Differences

The number of gates in both the GRU and LSTM varies. GRU update gate is similar to LSTM Input and forget gates. Information added and information retained through Input gate and Forget gate are independent to each other in LSTM. In GRU update gate is responsible for what stays and what get added as new information. So its not independent in GRU.

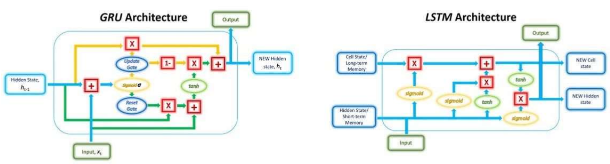

Fig 7: Comparison

B. Speed Differences

There are fewer number of weights and parameter updates in GRU, because of the lesser number of gates. GRU has two gates while LSTM has three gates.

C. Performance Evaluations

Both GRU and LSTM can be used interchangeably and can achieve similar results as they both come under RNN.

VIII. DATASET USED

A. AMI corpus

The AMI Meeting Corpus is a 100-hour multi-modal data set. Around two-thirds of the data was gathered through a scenario in which participants took on various responsibilities in a design team and completed a design assignment in one day.

B. LIGHT corpus

The LIGHT corpus from Parl.AI. This is a corpus of dialogues between participants to a text-based fantasy adventure game. It is crowdsourced.

C. Persuasion for Good corpus

The Persuasion for Good corpus contains online conversations in which one participant attempts to persuade the other to donate to a charity. It contains 1017 conversations

D. Switchboard dialogue act corpus

The Switchboard dialogue act corpus, a subset of the Switchboard corpus of telephone conversations between two participants on a range of topics annotated with dialogue acts. 440 speakers participate in 1,155 conversations.

E. TRAINS 1993 spoken dialogue corpus

TRAINS 1993 spoken dialogue corpus of task-oriented dialogues. It includes 98 dialogs, collected using 20 different tasks and 34 different speakers.

F. TRAINS-91 corpus

The smaller TRAINS-91 corpus collected as the pilot version of TRAINS-1993

G. RST Discourse Treebank

The documents in the RST Discourse Treebank, about 1/3 of the news articles in the WSJ portion of the Penn Treebank.

H. GNOME corpus

The documents from the GNOME corpus, which contains texts from museum labels, pharmaceutical leaflets, and tutorial dialogues.

IX. PREPROCESSING

The above corpora were used. The data is preprocessed and the json lines were created in the following format:

{
"doc_key": <Doc Key>,
"sentences": <Segments>,
"speakers": <Speakers>, ## Optional
"clusters": <Gold Coreference Clusters>
}

The existing preprocessed dataset was used in this paper.

```
[{'doc_key': 'large_cat',
  'sentences':[['The', 'large', 'cat', 'yawned', '.'],
               ['He','was', 'very', 'hungry', 'as', 'he', 'had', 'not', 'eaten', 'since', 'breakfast','.'],
               ['An', 'unfortunate', 'rat', 'came', 'along', '.'],
               ['The', 'cat', 'gobbled', 'him', 'up', '.']],
  'clusters': [[[0, 2], [5, 5], [10, 10], [23, 24]], [[17, 19], [26, 26]]]
}]
```

Fig 8: Sample input data

X. PARAMETERS

The optimizer used is "adam" as it can handle sparse gradients on noisy problems effectively and it converges faster. The loss function used is "binary cross entropy" as it applies independent probabilities. The embedding dropout rate is maintained as 0.5. The size of the hidden layer is maintained as 50. The number of hidden layers used for feedforward neural network is 2. The hidden dropout rate is 0.2. The embedding size is 300. The maximum number of candidate antecedents for each candidate mentions is 250. The ratio of negative to positive examples is maintained as 2.

XI. EVALUATION METRICS

The commonly used metrics include:

i) MUC (link based) [19],
ii) B-CUBED (mention based) [20]
iii) CEAF (entity based) [21]
iv) BLANC [22]

The errors generated are considered by MUC (Message Understanding Conference). The golden annotation is compared to the response by the system while using this metric.

The items in the same cluster if they belong to the same category is defined by the B-CUBED precision.

Constrained Entity- Alignment F-Measure (CEAF) is measured with the constraint that at most a system/reference entity is aligned with one reference/system entity.

The standard evaluation method used in "CoNLL2012 shared task" is used here [3] which takes a weighted average of three metrics: MUC, B-CUBED, and CEAF.

XII. Activation Functions

The activation function in a neural network is responsible for converting the node's summed weighted input into the node's activation or output for that input. For the models and domains used in this paper, the following activation functions were investigated.

A. RELU

The derivative of the Rectifier linear unit (ReLU), a linear function, is either 0 or 1. When x is less than or equal to 0, the ReLU function returns 0, but when x is larger than 0, it returns x. The output of the function can be generalised as max (0, x).

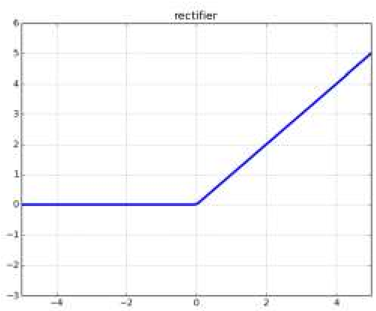

Fig 9: ReLU function

B. SWISH

In 2017, the Google Brain team announced Swish activation as an alternative to ReLU. [23]

$$y = x.sigmoid(x)$$

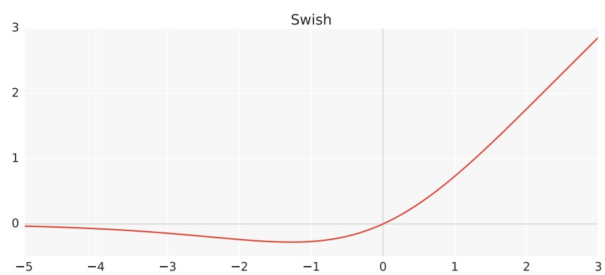

Fig 10: Graph of swish function

C. MISH

A Self Regularized Non-Monotonic Neural Activation Function. The function is a combination of popular activation functions, hyperbolic tangent and softplus. [24]

$$f(x) = x.tanh\,(softplus(x))$$

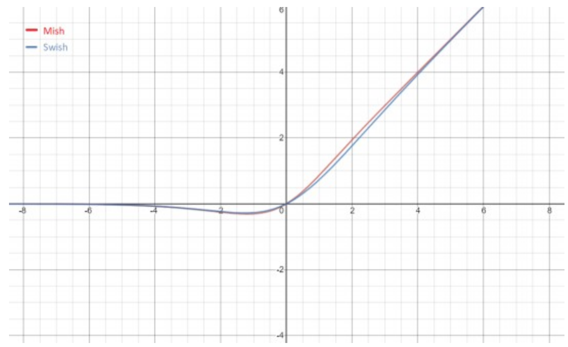

Fig 11: Mish vs Swish

"When compared to other commonly used functions like ReLU (Rectified Linear Unit) and Swish, the Mish function is utilised because of its inexpensive cost and numerous qualities such as its smooth and non-monotonic nature, unbounded above, bounded below property boosts its performance. Mish's characteristics are described in full below:

1. Unbounded above and below: Unbounded above is a desired attribute for any activation function because it prevents saturation, which slows down training. As a result, the learning process will be accelerated. The attribute of being bounded below aids in generating strong regularisation effects (fits the model properly). (With a range of [0.31,), Mish's attribute is identical to ReLU's and Swish's.)
2. Non-monotonic function: This trait aids in the preservation of small negative values, thus stabilising network gradient flow. Most commonly used activation functions, such as ReLU [f(x) = max(0, x)] and Leaky ReLU [f(x) = max(0, x), 1], fail to keep negative values since their differentiation is 0, and hence the majority of neurons are not updated.
3. Mish is a smooth function that is good at generalisation and effective optimization of outcomes since it has an infinite order of continuity. Between ReLU and Mish, there is a significant difference in the smoothness of the terrain of a randomly initialised neural network. Swish and Mish, on the other hand, have a landscape that is almost identical." [25]

XIII. Results

The Datasets AMI, Treebank, Light, Switchboard and Gnome are tested using the models, parameters, and activation function by combining different development and test sets of above-mentioned datasets along with the two additional datasets Trains 91 and Trains 93. The performance of each of the activation functions on these cross-domain models is shown below. The swish and mish seem to be performing better than the ReLU in most of the cases and wherever these performances seem to be higher the Bidirectional GRU is a better performing model. By all counts and by the obtained results it could inferred that the activation function is contributing to the model's better performance in a cross-domain scenario.

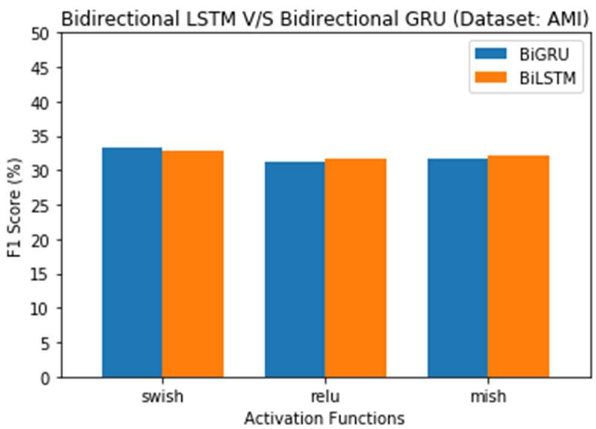

Fig 12: Test – AMI, Dev – AMI, Train - Treebank

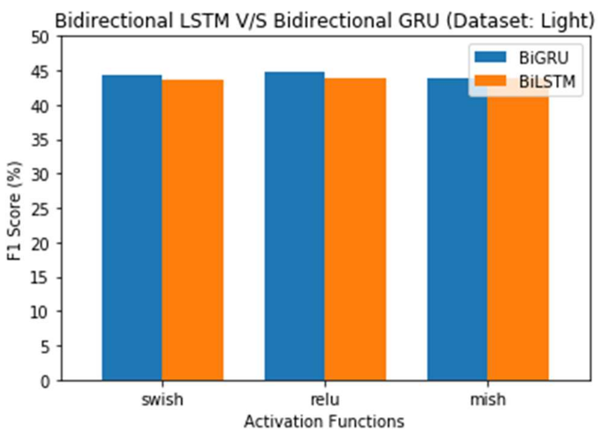

Fig 13: Test – Light, Dev – Light, Train – Train93

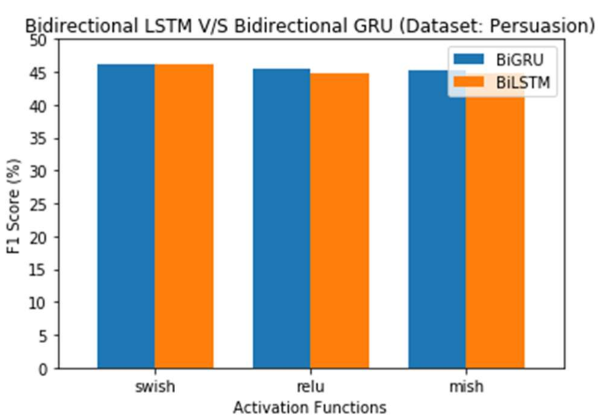

Fig 14: Test – Persuasion, Dev – Train91, Train – Train93

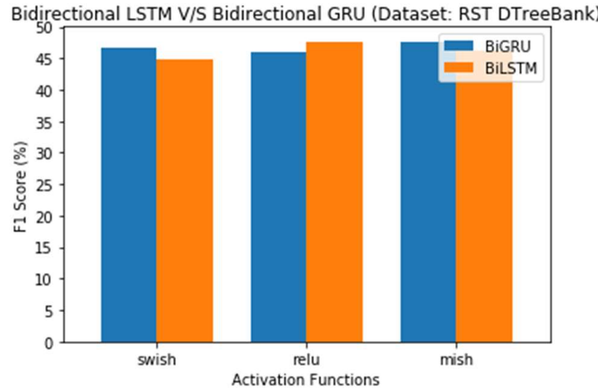

Fig 15: Test – Treebank, Dev – Treebank, Train - Treebank

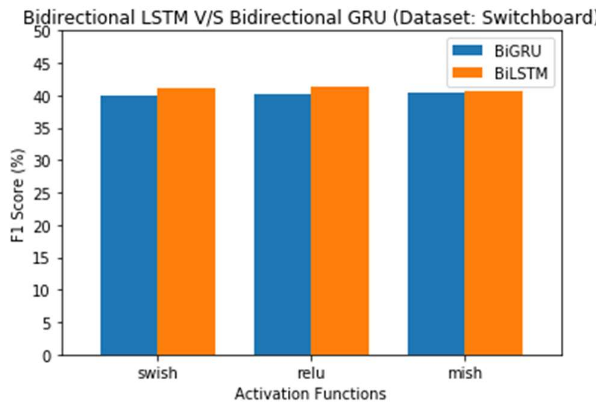

Fig 16: Test – Switchboard, Dev – Switchboard, Train - Gnome

Layer (type)	Output Shape	Param #	Connected to
input_1 (InputLayer)	[(None, None, None,	0	
lambda (Lambda)	(None, None, 300)	0	input_1[0][0]
dropout (Dropout)	(None, None, 300)	0	lambda[0][0]
bidirectional (Bidirectional)	(None, None, 100)	105600	dropout[0][0]
bidirectional_1 (Bidirectional)	(None, None, 100)	45600	bidirectional[0][0]
lambda_1 (Lambda)	(None, 100)	0	bidirectional_1[0][0]
dropout_1 (Dropout)	(None, 100)	0	lambda_1[0][0]
input_2 (InputLayer)	[(None, None, 4)]	0	
lambda_2 (Lambda)	(None, None, 4, 100)	0	dropout_1[0][0] input_2[0][0]
reshape (Reshape)	(None, None, 400)	0	lambda_2[0][0]
dense (Dense)	(None, None, 50)	20050	reshape[0][0]
dropout_2 (Dropout)	(None, None, 50)	0	dense[0][0]
dense_1 (Dense)	(None, None, 50)	2550	dropout_2[0][0]
dropout_3 (Dropout)	(None, None, 50)	0	dense_1[0][0]
dense_2 (Dense)	(None, None, 1)	51	dropout_3[0][0]
lambda_3 (Lambda)	(None, None)	0	dense_2[0][0]

Fig 17: Model

XIV. CONCLUSION

This paper tried to explore the datasets from different domains and how they perform when used in a mixed way of training and testing with the models of LSTM and GRU (Bidirectional for both). The new activation functions swish and mish are also tried upon, and the end results are compared. The use of the activation and the further fine tuning can be explored more, on different models and a generalized model can be made possible, cross domain, if further analysis takes place.

XV. ACKNOWLEDGEMENT

I am grateful to my Supervisor and Professor, Dr. Massimo Poesio, who have contributed extensively to this field, and I am glad I could also be a small part of it by working on this project. I would like to thank my parents and siblings for their love and support throughout my master's degree.

REFERENCES

[1.] Helene Schmolz, 2015. "*Anaphora Resolution and Text Retrieval*" (https://doi.org/10.1515/9783110416756)

[2.] Shivangi Shekhar, Umesh Kumar, 2018. "Review On The Techniques Of Anaphora Resolution" (10.21172/1.82.006)

[3.] Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, Yuchen Zhang 2012. "CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes". (https://aclanthology.org/W12-4501)

[4.] Jiangfan Feng, Nan Xu, 2013. "Using Chinese Natural Language Interfaces for Navigation in Mobile GIS." (https://ijcsi.org/papers/IJCSI-10-1-3-21-24.pdf)

[5.] Abhijeet R. Sontakke, Prof. Amit Pimpalka, 2014. "A Review Paper on Hindi Language Graphical User Interface to Relational Database using NLP".

[6.] Massimo Poesio, Mijail A. Kabadjov, 2004. "A General-Purpose, Off-the shelf Anaphora Resolution Module: Implementation and Preliminary Evaluation" (http://www.lrec-conf.org/proceedings/lrec2004/pdf/559.pdf)

[7.] Josef Steinberger, Massimo Poesio, Mijail A. Kabadjov, Karel Jezek, 2007. "Two uses of anaphora resolution in summarization. Information Processing & Management". (https://doi.org/10.1016/j.ipm.2007.01.010)

[8.] Kellie Webster, Marta Recasens, Vera Axelrod, Jason Baldridge, 2018. "Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns". (https://aclanthology.org/Q18-1042)

[9.] Massimo Poesio , Tomonori Ishikawa, Sabine Schulte im Walde, Renata Vieira R, 2002. "Acquiring Lexical Knowledge for Anaphora Resolution" (http://www.lrec-conf.org/proceedings/lrec2002/pdf/117.pdf)

[10.] Ruslan Mitkov , "Outstanding issues in anaphora resolution, in: International Conference on Intelligent Text Processing and Computational Linguistics".

[11.] Triveni Lal Pal, Kamlesh Dutta, Pardeep Singh, 2012. "Anaphora Resolution in Hindi: Issues and Challenges".

[12.] Ishani Sharma and Pradeep Kumar Singh, 2016. "A Survey on Anaphora Resolution".

[13.] Ashima and Rajni Mohana, "Anaphora Resolution in Hindi: Issues and Directions", Indian Journal of Science and Technology,Volume-9 (32),August 2016

[14.] C.J. Fillmore , Pragmatically controlled zero anaphora, in: Annual Meeting of the Berkeley Linguistics Society, 12, 1986, pp. 95–107

[15.] R.M. Dixon, Demonstratives: a cross-linguistic typology, Stud. Lang. Int. J. Sponsor. Foundat. "Foundations of Language" 27 (1) (2003) 61–112 .

[16.] R.A. Van der Sandt , Presupposition projection as anaphora resolution, J. Semant. 9 (4) (1992) 333–377 .

[17.] R. Mitkov , Outstanding issues in anaphora resolution, in: International Conference on Intelligent Text Processing and Computational Linguistics, Springer, 2001, pp. 110–125

[18.] Schuster, Mike, and Kuldip K. Paliwal. "Bidirectional recurrent neural networks." Signal Processing, IEEE Transactions on 45.11 (1997): 2673-2681.2

[19.] Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. "A model theoretic coreference scoring scheme". In Proceedings of the Sixth Message Undersatnding Conference (MUC6), pages 45–52.

[20.] Amit Bagga and Breck Baldwin. 1998. "Algorithms for Scoring Coreference Chains". In The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference, pages 563–566.

[21.] Xiaoqiang Luo. 2005. "On coreference resolution performance metrics". In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, pages 25–32, Vancouver, British Columbia, Canada, October.

[22.] Marta Recasens and Eduard Hovy. 2011. "Blanc: Implementing the rand index for coreference evaluation. Natural Language Engineering".

[23.] Prajit Ramachandran, Barret Zoph, Quoc V. Le, 2017. "Swish: A Self-Gated Activation Function"

[24.] Diganta Misra, 2019. "Mish: A self-regularized non-monotonic neural activation function"

[25.] https://medium.com/clique-org/mish-activation-function-in-yolov4-49d2f30c6bb3