National College of Ireland

Bounded Memory Coreference Resolution Using SpanBERT on Litbank Dataset

MSc Research Project

MSc Data Analytics

Mandeep Kaur Taneja

Student ID: 21123837

School of Computing

National College of Ireland

Supervisor: Christian Horn

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Mandeep Kaur Taneja		
Student ID:	21123837		
Programme:	MSc Data Analytics	**Year:**	2022-2023
Module:	Research Project		
Supervisor:	Christian Horn		
Submission Due Date:	15/12/2022		
Project Title:	Bounded Memory Coreference Resolution Using SpanBERT on Litbank Dataset		
Word Count:	5882 **Page Count** 19		

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Mandeep Kaur Taneja
Date:	15/12/2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Bounded Memory Coreference Resolution Using SpanBERT on Litbank Dataset

Mandeep Kaur Taneja

21123827

Abstract

One of the key undertakings of Natural Language Processing (NLP) is Coreference Resolution (CR) that attempts to distinguish and determine various references to an item in a record. It attempts to find all semantic articulations - "mentions" that allude to a single "entity". CR is a fundamental stage in numerous semantic benchmarks, similar to address replying, regular language derivation, and named element distinguishing. These semantic seat markings have shown huge upgrades with the present-day transformer-based BERT models. With the effective use of present-day Bidirectional Encoder Representations for Transformers model, these semantic benchmarking have shown tremendous upgrades in their overall efficiency and accuracy. SpanBERT is an expansion of the BERT model that predicts ranges of text all the more accurately, especially better at separating related but distinguishable elements (e.g., President and CEO). In spite of the fact that BERT and SpanBERT models perform wonderfully in short sentences, they have very large runtime, memory, and computational asset prerequisites in preparing and modelling, when performed on lengthy records as they require keeping every token/entity in memory all the time. To solve the issue of huge computational resource requirements, this paper proposes a technique of storing only a limited number of tokens at a given instance of time (bounded memory architecture) and effectively "forgetting" a previously tracked entity whenever a new token is introduced in a heuristic manner. In most of the CR research, a classical dataset was used called OntoNotes. However, this dataset was created in 2012 and lacked quality annotations for our present-day usage. Hence, this paper has performed analysis on a newer Litbank dataset which is a collection of 100 classic Literature novels and can be categorised as a long text document. This dataset is annotated and maintained using Automatic Content Extraction guidelines, hence making it a better choice than OntoNotes dataset.

1 Introduction

One of the most important problems in Natural Language Processing (NLP). is "coreference resolution" (CR). CR is the linguistic phenomenon in which words or phrases that appear in a series of sentences are resolved to the entity to which they refer. The references can link to a

person, place, item, or any other noun. This aspect of linguistic analysis aids in the study of language usage. It also helps in the modern subject of NLP, which functions as a foundation for numerous computer models that analyse speech. In order to understand the phenomenon in depth, we must first comprehend the meaning of discourse. In the context of NLP, discourse is a series of sentences that follow one another. Obviously, there will be entities being discussed and probable referrals to those entities in the debate. The term "mention" is used to refer to these references.

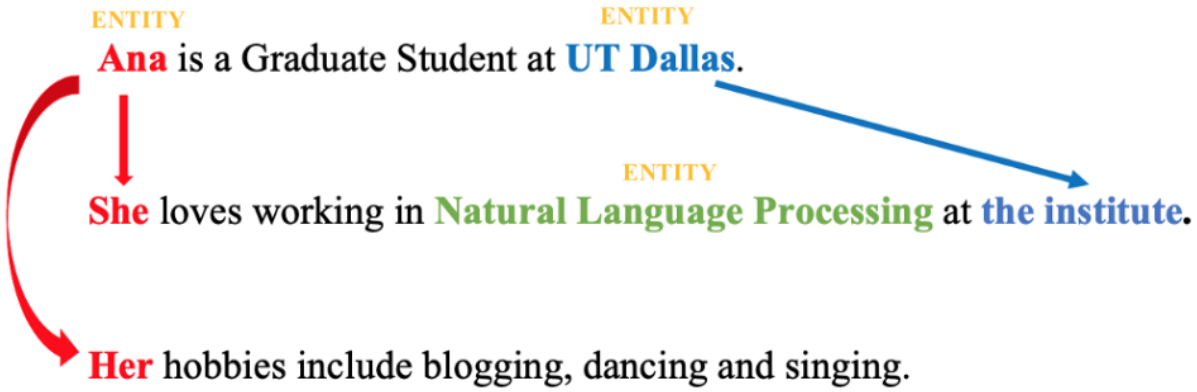

Fig 1: Example for an entity and its mentions (Team, n.d.)

In NLP, reference is a linguistic process in which a word in a sentence or discourse can refer to another word or entity. The process of resolving these references is referred to as Reference Resolution. Two examples of Reference Resolution are "She" and "Her" referring to the entity "Ana" and "the institute" referring to the entity "UT Dallas" in the above example(Team, n.d.). In particular, Coreference Resolution is the process of resolving pronouns to determine which entities they refer to. Additionally, it is a form of Reference Resolution. The resolved entities could be a person, location, organization, or event. There are typically two types of references: Exaphor and Endophor. Endophor refers to a term that describes an entity that emerges in the discourse. Exaphor, on the other hand, refers to an entity that is not present in the discourse. Understanding these linguistic features of the coreference relation enables us to do coreference resolution with the least amount of mistake possible.

Recently BERT pre-training models have displayed superb performance gains that mask individual words (Joshi et al., 2019). But there are numerous instances where NLP cannot be effectively implemented when deductions involve multiple spread of text. For example, "***Which team won the IPL 2019?"*** This question can be correctly answered if the machine knows that IPL refers to the Indian Premier League and "Mumbai Indians" is a premier league cricket team. Such mentions provide a major challenge to BERT coreference models that perform token level self-learning rather than span level and as a result, span level self-supervised learning, SpanBERT, outperforms them (Joshi et al., 2020).

Pragmatically speaking, most of the entities in real-life text documents have a small spread (distance between first and last mention of the entity) (Keller, 2010). Thus, not all entities need to be kept in memory all the time. This can considerably reduce the computational power

required while maintaining a similar level of accuracy. This prototype is trained in such a way in which it resolves the problem of limited storage by effectively "forgetting" a previously tracked entity when a new mention comes into picture via heuristic approach (Toshniwal et al., 2020). Even this state-of-the-art Learning based Bounded Memory Coreference Resolution requires 16 GB RAM to run efficiently which is not possible in many day-to-day computer systems. Hence, we have optimized this model by fine-tuning the hyperparameters and are still able to get good F1 scores (Dev set: 71.9% and Test set: 69.2%) on the benchmark CoNLL 2012 values with the 12GB RAM of Google Colab.

1.1 Terminology

Coreference resolution is the technique of matching phrases in a text to the relevant entity to which they link. This is considered a significant obstacle in many NLP topics, and it has been so for a long time. Conventional coreference resolution techniques face one of their key hurdles in the absence of a better solution for unclear pronoun resolution (Sukthanker et al., 2020) . Resolution of coreferences is the process of relating all mentions in a document with a shared reference entity to that entity. This referent entity may be the postcedent or the antecedent, and the mention to which it refers is typically a cataphora or anaphora pronoun. There are numerous types of references, which are described below.

Anaphora and Cataphora

Anaphora occurs in a sentence after the word it refers to, whereas cataphora occurs after the word it refers to. For anaphora the word it refers to is known as antecedent and for cataphora the word it refers to is known as postcedent (Karabiben, 2021).

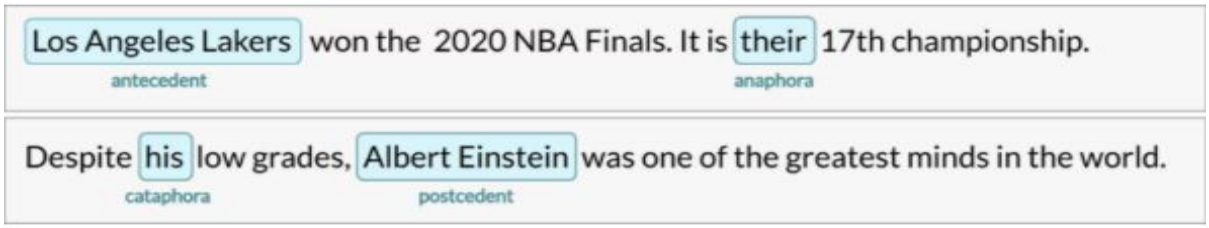

Fig 2 : Difference between cataphora and anaphora ("Paweł Mielniczuk," n.d.)

Here in the sentence above, anaphora refers to the pronoun their, which comes after the noun it refers to, i.e., Los Angeles Lakers.

Split Antecedent

In some sentences, an anaphora refers to multiple entities.
Example: Arun and Dhruv are friends, they studied together. In this example anaphora, they are referring to multiple entities: Arun and Dhruv.

Ambiguous Pronouns

The pronoun substitutes the noun in coreference, but it must always refer explicitly and precisely to the noun it supplants, the antecedent. If the pronoun's antecedent in a text is ambiguous, the text itself becomes unclear (Mohan and Nair, 2019). Ambiguous pronoun occurs when a text contains multiple antecedents, making it difficult for the system to differentiate between the antecedents that refer to the right pronoun in the text. This problem increases with increase in document length. With the increase in the number of mentions in a document, the resolution of coreference becomes more difficult.
Example: Sophie was taken by Anna to her restaurant. Given that she was a foodie Here, she appears to refer to Sophia, but it is unclear whether the word is referring to Anna or Sophia.

Coreferring Noun Phrases

An anaphoric situation in which the second noun (2) in a phrase refers to a previous descriptive form of an expression (1).
Example: Many villagers (1) are ailing. These people (2) hardly get over COVID-19

Presuppositions/Bound Variables

Presupposition is a type of resolution that is categorized as a coreference (or any other "reference"). Due to the fact that a pronoun (2) is not strictly referential, it cannot be replaced by the quantified expression (1). The pronoun is, after all, a variable whose value is determined by its antecedent.
Example: Every country (1) is dealing with COVID in its (2) own way

Misleading Pronominal Reference

When there is no direct connection between the pronoun and other mentions in the text, but the pronoun is still present, certain situations can be misleading.
• **Cleft** A cleft statement is a complicated expression that has a simpler, less deceiving replacement. This is an instance in which the pronoun it is redundant, we can simply construct a statement that conveys the same meaning without using the pronoun.
• **Pleonastic** "it": This type of references are so widespread in English; it deserves special attention. It is required to complete a grammatical expression, although it does not refer to any other term in the phrase.

1.2 Research question

- *Can we implement a coreference resolution SpanBERT model using Google Colab (limited computational resources)?*

  With 32GB RAM, conventional versions like the BERT and SpanBERT perform well. Additionally, they work well when used with large documents (more than 10 sentences). Coreference resolution on large documents, however, has a significant

processing burden when using cutting-edge methods like BERT and Span-BERT. Large tech corporations or highly funded researchers may incur this overhead (M. Joshi & Zettlemoyer, 2019), but tackling a challenge like this with little funding is difficult. The approach of bounded memory for things is suggested in this research plan as a way to circumvent the issue of high processing power and achieve comparable coreference resolution outcomes.

1.3 Research objective

- We intend to build an optimal coreference resolution model using Google Colab with limited computational resources.
- We will prove that not all the mentions need to be stored in the memory at all times.

2 Related Work

One of the most critical tasks for natural language understanding (NLU) is teaching the word model correctly. Nevertheless, because of the compound nature of word usage and its variance in different scenarios makes creating an efficient representative model that covers a wide variety of linguistic factors is a difficult task. Peters et al. does a great work in building an extensive word representation model to solve the aforementioned issue. But the model introduced by them is fairly limited when it has to work with a long succession of texts (Peters et al., 2018).

In 2019, Devlin et al. extends the work of Peters et al. by implementing a newer representation method of the English language using a revolutionary machine learning model which they name - BERT (Bidirectional Encoder Representations for Transformers). This new model was trained on the huge Wikipedia corpus and it proved to be a success as it solved 12 complex classical NLP problems (Devlin et al., 2019). The biggest success factor for this model was its ability to work with large documents and texts of sentences. Furthermore, during this period Lee et al. introduced a highly optimised estimation model for higher-order interpretation of sentences which they call - c2f-coref machine learning model. (Lee et al., 2019). The approach used in this model was to use antecedent spread of words in a ranking-span architecture and repetitively define the span boundaries. Keeping in mind that the accuracy of the model should not be lowered, it aggressively pruned the existing state-of-art approach.

Extending this c2f-coref model (Devlin et al., 2019) the full long short term memory encoding (LSTM model) was fundamentally supplanted by the Bidirectional Encoder Representations for Transformers (BERT model) by a Facebook funded research team lead by Mandar Joshi (Joshi et al., 2019) and subsequently by Xie Wu and his team(Wu et al. in 2020). The non-overlapping portions in the independent variant each serve as a separate instance of BERT. The overlap variation divides the content into overlapping chunks to add context to the model beyond the 512 token limits. However, models proposed by both Mandar Joshi and Xie Wu became infeasible in the long run when it comes to dealing with long text documents as they

required huge memory and computation resources granted only by heavy end supercomputers. The base model (Joshi et al., 2020) was further improved in their next work by Joshi et al. in 2020 in which SpanBERT was introduced and used as the underlying encoder instead (Joshi et al., 2019).

The model introduced in the paper of Xia et al. uses a contextualized encoder, SpanBERT to encrypt an entire segment (Heinzerling et al., 2017). By expanding an incremental clustering approach to include contextualized encoders and neural components, they model coreference resolution under a fixed memory constraint. For any sentence, each mention span is proposed and scored against explicit entity representations derived from the prior document context by their end-to-end algorithm. Before being removed by the memory cell and being completely unaccounted for, these spread of words are used to update the representations of its entities and mentions throughout the document. (Heinzerling et al., 2017) Heinzerling furthermore enhanced the work of Xia and co by making major enhancements in the model and making it compete with the current state of the art accuracy. Curren and Webster made hyper tuning enhancements in the Xia model to keep the mention-entity paradigm up to date. (Webster and Curran, 2014). Rehman and company in 2011 and Clark et al. in 2014, unlike previous attempts, proposed an idea to keep in account only the implicit entities rather than storing the meta data of the corresponding mentions as well. Only entity cluster representations are stored by the new entity-mention paradigm, which is updated as coreference predictions are made. Less memory is required in this approach than those that additionally store mention representations. In real world scenarios, the entity spread (first and last mention of an entity) is pretty small. Hence, maintaining its footprint through the entirety of the model building process does not make sense. So, maintaining only a limited number of entities at a given point of time potentially solves the problem of huge computational resources and saves a lot of memory.

The past work of Liu (Liu and Perez, 2017) introduced an idea of using a heuristic approach of learned bounded memory architecture which potentially showed great results in short text and it only worked on single token rather than spread of words. They implemented this idea of a newly annotated coreference dataset - Litbank which is a collection of 100 classical English literature and can be termed as a long text document. The learned bounded memory works by “forgetting” an entity/token which was previously being tracked when a new entity arrives hence solving the problem of limited memory. These ground-breaking research paves way to our research in the current day.

3 Research Methodology

3.1 Coreference Resolution

Human and machine communication is commonly referred to as "natural language processing" (NLP). NLP is one of the most complex subfields of artificial intelligence because of the ambiguity and exceptions of human language, which make it difficult for computers to adapt.

Eliminating unclear terms that require context to be properly understood is one step towards making it easier for them. Pronouns such as he, him and she can be replaced with the nouns they refer to are a good example of it. Coreference resolution (CR) is the process of identifying all semantic statements (called mentions) in a document that relate to the exact same entity. After locating and categorizing these occurrences, we replace these pronouns with their corresponding noun phrases (Mohan and Nair, 2019).

3.2 Data Understanding

Before the introduction of LitBank in 2019 by (Bamman et al., 2020), most of the research was conducted on the OntoNotes Dataset. OntoNotes contains a huge collection of multiple forms of text, including blogs from the web, news articles, communicative phone speech, news groups, telecasts, and TV talk shows, which are present in three languages: English, Arabic, and Chinese. Almost all present systems evaluate solely based on these OntoNotes (Hovy et al., 2006). However, it includes very specific areas that do not work well on new domains. We perform our experiments on the Litbank dataset by Bamman et al. (Bamman et al., 2020). which comprises 100 distinct fiction novels written in English from the year 1719 to the year 1922. All these novels are in the public domain in the United States. These writings contain a wide variety of linguistic styles as well as current annotations for entities such as places, people, foreign affairs, and automobiles using ACE-Style (Sims et al., 2019).

The style for annotating the Litbank data by defining the mentions and basis for coreference relationships is similar to the OntoNotes. The annotation criteria are as follows:

Singleton

Singleton mentions are the nouns that do not participate in coreference. OntoNotes do not consider these mentions, which result in complex coreference resolution in documents because of the extra step required to identify the potential entities for coreference, i.e., to separate singletons and phrases that are part of coreference. The Litbank dataset considers singleton mentions to be markable. The dataset contains 17.4 percent of singletons, whereas OntoNotes has 56 percent of singletons examined by (Recasens et al., 2013)

Quantified and negated noun phrases

OntoNotes does not annotate negated and quantified noun phrases which lead to unhandled coreference in some cases. However, such mentions are annotated in the Litbank dataset. For example -" [Not every person] can love [their] appearance".

Types of entities

A lot of unrestricted coreference is covered in OntoNotes; however, litbank markable entities are categorically annotated in a total of six entity categories:

- Person (PER)
- Organizations (ORG)
- Place (LOC)
- Facilities (FAC)
- Inter-political entities (IPE)

• Vehicles (VEH)

Type	Number	Frequency
PER	24,180	83.1
ORG	149	0.5
LOC	1,289	4.4
FAC	2,330	8.0
GPE	948	3.3
VEH	207	0.7

Table 1: Entity Type Count and Frequency

Span Maximum Length

As is the case with OntoNotes, the maximum span length is the extent of span in the sentence, as displayed in the following:

[The man who kicked the ball and drank beer] ran away.

Categories of Entities

Three types of noun phrases are included in the Litbank dataset.

• Proper names (PROP) - Thomas Edison, Sachin Tendulkar
• Pronouns (PRON) - you, she, her, thine, thou, them
• Common phrases (NOM) - the son, a gold chain

Category	Number	Frequency
PROP	3,550	12.2
PRON	15,816	54.3
NOM	9,737	33.5

Table 2: Entity Category Count and Frequency

Honorifics

As many coreference resolution models strip the honorific annotation of OntoNotes, like, ("[[Mr.] Messi]", and "[[Miss] Harsimran]]"), Litbank does treat make them as individual markable span, leaving them as "[Mr. Collins]" and "[Miss Havisham]".

3.3 Data Preparation

Entity Spread and Active Entities

For the mention span (x_i) in a given document G, Let START (x_i) and END (x_i) represent the start and end token positions. ENT (x_i) is the entity that mention span (x_i) refers to in G. With the help of these notations, we will explain the below concepts in detail.

Entity Spread (ES)

It is the number of tokens between the beginning and ending mentions of an entity. Entity e's entity spread ES(e) is calculated by:

$$\mathrm{ES}(e) = [\min_{\mathrm{ENT}(x)=e} \mathrm{START}(x), \max_{\mathrm{ENT}(x)=e} \mathrm{END}(x)]$$

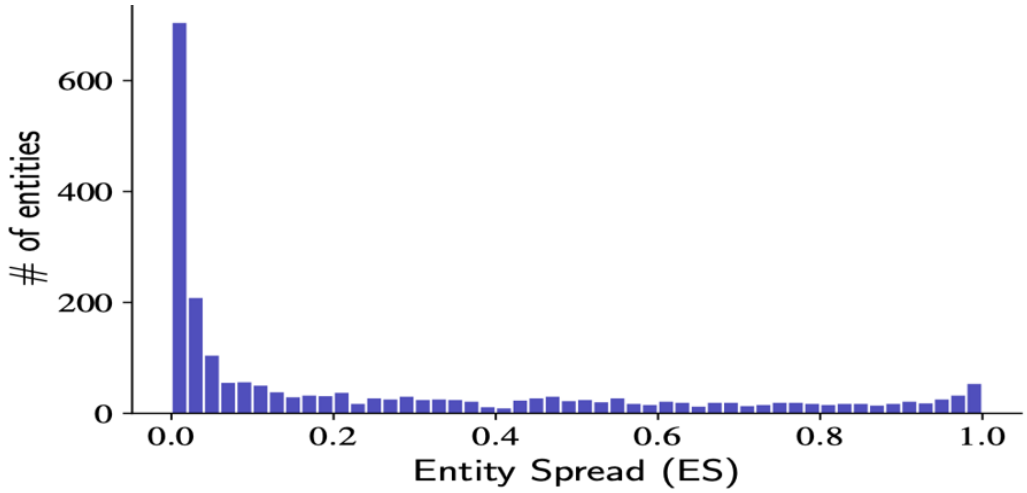

Fig 3: Entity Spread as fraction of Litbank Dataset

Figure 2 is a histogram of length of entity spread of non-singleton clusters divided by the length of the document.

Active Entity (AE)

AE(k) refers to the number of distinct entities whose distribution includes the token k.

Maximum Active Entity

It is simply the calculation of the maximum entities that are active at any particular token position k in the document G.

Table 1 shows the maximum entity count for the LitBank dataset. The given values prove the objective of the research that not all entities are required to be stored in the memory.

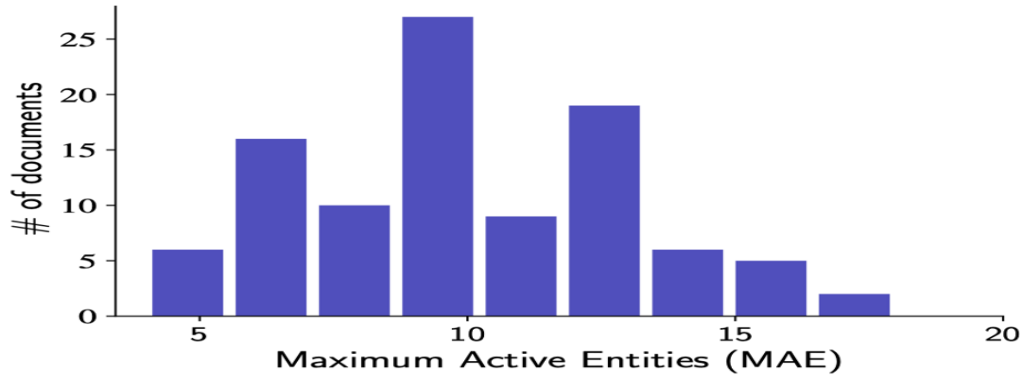

Fig 4: Maximum Active Entities for Litbank Dataset

Fig 3 displays the bar graph of the maximum number of active entities against the number of documents in the Litbank dataset.

3.4 Model Building

Previous research related to coreference resolution resulted in the discovery of Bidirectional Encoder Representations from Transformers (BERT), which has displayed promising results and focuses on masking individual tokens to predict. But new research has found that a better score is given for coreference resolution when a prediction is based on two or more spans of text. SpanBert, an extension of BERT, focuses on masking the span of text instead of a single token. Both these models require heavy load machinery and memory, which is not possible with Google colab or daily-use computers. In this research, we propose a bounded memory model that almost achieves coreference resolution. The model works well with limited memory by replacing an entity in the memory that has already been tracked with an entity that has not been tracked yet.

4 Design Specification

Fig. 4 illustrates the inner workings of the BERT model. To begin, a sequence of tokens is passed through embedding to convert them into vectors, which are then processed in a neural network. It produces a sequence of vectors of size A that correspond to the same index as the input vector. To predict the masked words, it requires-

- An additional layer of classification after the encoder output.
- To multiply the output by the matrix of embedding and then transform it to the vocabulary dimension.
- To calculate the probability of all the vocabulary tokens using SoftMax.

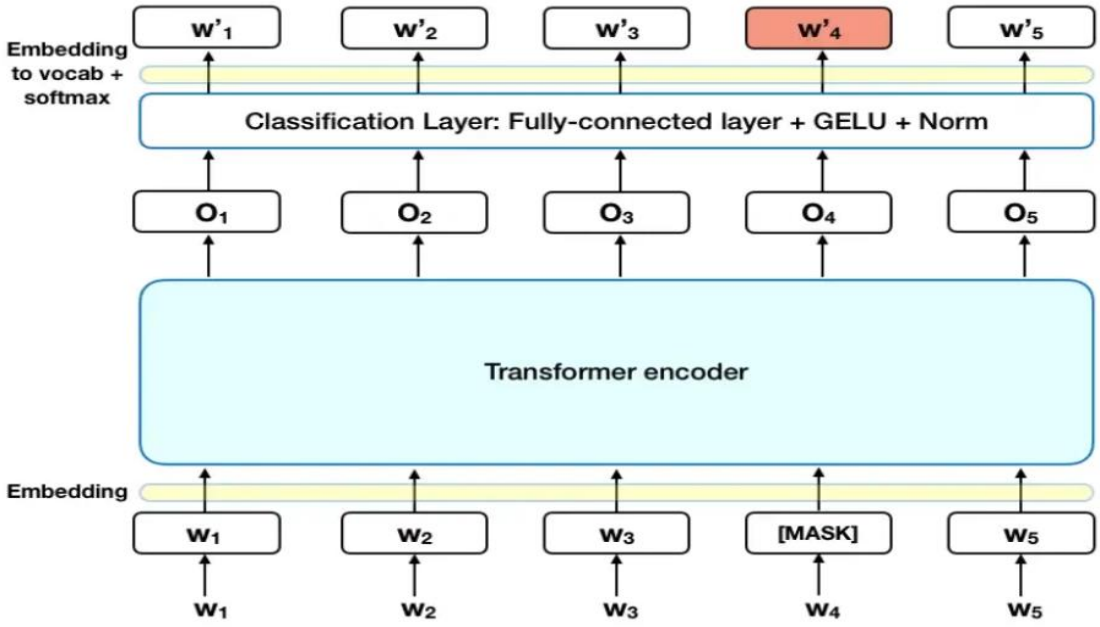

Fig 5: BERT Architecture (Horev, 2018)

4.1 Bidirectional Encoder Representations from Transformers (BERT)

BERT is a free open-source model of machine learning developed by researchers in Google. It became quite popular in the community of Machine Learning as it provides promising results in an assortment of NLP tasks which include Question and Answering, sentimental analysis, Multi-Genre Natural language inference (MNLI). BERT utilizes Transformers, an attention mechanism which dynamically calculates the weightings between related words on the basis of their connection. It is made up of two different parts: an encoder that takes the input text and reads it and a decoder that makes a prediction for the task. Traditionally, models used to read text sequentially in one direction only—from left to right or right to left. But with the Transformers, BERT can read text in both directions simultaneously, which is why it is called a bidirectional model.

Masked Language Model (MLM)

Masking is a phenomenon in which certain parts of sentences are concealed and then fed to the model to predict those gaps. In this model, 15% of the words are masked using the [MASK] token in each phrase before feeding it to BERT. The hidden words are then predicted by the model with the help of the context provided by other words in the sentence.

Next Sentence Prediction (NSP)

Next-sentence prediction (NSP) aims to establish a continuing link between sentences, whereas MLM trains the relationship between tokens. It includes providing BERT with two sentences, S1 and S2, and asking it to predict if S2 is the subsequent sentence of S1. In the training phase, 50 percent of the total inputs consist of pairs where the S2 is the next sentence in the original text, whereas the other 50 percent are random phrases. The presumption is that the random phrase is unrelated to the first.

4.2 BERT vs SpanBERT

SpanBERT is an extended form of BERT. It enhances the performance by predicting spans of text instead of single tokens. Its difference from BERT is explained with the following fields.

Masking scheme

The masking scheme of BERT is to mask words in a sentence at random, whereas SpanBERT masks spans of connecting text randomly.

Training Objective

Additional distinction involves the training objectives of both models. BERT training focused on two criteria (two loss functions). MLM and NSP, which we have already explained above. However, in SpanBERT, we use only the Span Boundary Objective (SBO) to train the model, which eventually leads to the loss function. In SBO, rather than considering the representation of the masked token, SBO takes into account the representations of the neighbouring tokens.

4.3 Learned Bounded Memory Architecture (LB Memory)

Learned Bounded Memory architecture is a technique which helps in circumventing the problem of limited memory and resource allocation. It is a heuristic method of tracking the number of entities in memory at a given point of time. Let N be the maximum number of entities that could be stored in memory and M is the number of currently tracked number of entities in time *t*. Let's assume that a new entity *e* is to be tracked now. If N > M, it means that the memory is still not full and the entity *e* can be stored without any issue. However, if N <= M, then the method has to heuristically (Joshi et. al., 2020) decide whether to delete an existing entity to store the new one or not. This kind of architecture is called Learned Bounded Memory architecture.

The Learned Bounded Memory architecture (LB-MEM) tries to predict a learned feedforward neural network score f_r. Then calculates:

$$D = arg\ min([f_r(e_1), \dots, f_r(e_M), f_r(x_i)])$$

And performs the operation as shown in Figure 3

Value	Perform
1<d<M	Forgets prev entity and adds new in the memory
d=M+1	Ignore the entity due to full memory
d=M+2	Predicts the mention as invalid

Table 3: Operations on Learned Bounded Memory

5 Implementation

To implement the given research, we will first install all the required libraries some of which are SciPy, torch, transformers etc. The following steps are followed to implement this research project.

5.1 Dataset Preparation

LitBank is a new literary dataset annotated using ACE guidelines for coreference. As mentioned earlier, it contains excerpts from 100 classic English literature novels with a mean length of 2100 words. It is one of the best datasets to perform long document coreference resolution. Due to our limited computational resources, we did our evaluation on 25% of the LitBank data, which is divided into two categories: independent and overlap. For this research, we are focusing on overlap data. It is further cut into 10-fold-cross validations over 80/10/10 splits.

5.2 Model Building

Further enhancing the results of previous research (Joshi et. al.,2020), we build the SpanBERT model using a learned bounded memory approach which tackles the issue of limited computational resources. The model teaches itself in time how to heuristically process the

entities as they come sequentially whether or not to ignore them or to include them in the cluster. The method of adding is as followers -

- The entity is put into an existing cluster, if memory is not full.
- The entity is put into a new cluster.
- The entity is neglected due to lack of memory.
- The entity is ignored as invalid, if already been worked upon.

Implementing the LB Memory model is a twostep approach.

Encoding of Documents

The documents are divided into multiple chunks of lengths [128, 256, 384, 512] and each chunk is encoded independently. For our experiment, considering the computational limitations of our laptop, we have used 256 sequence length as opposed to 512 (Toshniwal et al., 2020).

Mention Modelling

This stage predicts which high-scoring mentions should be grouped together in a cluster. To implement this , we have picked a span width of 0.15 percent of the total document size.

5.3 Hyperparameters

We did not adjust the Feedforward neural network (FFNN) depth and size hyperparameters (Joshi et al., 2020) and kept them equal to Joshi et al. After performing many test runs, we decided to increase the dropout rate from 0.3 to 0.5 as we could see some gains during the experimentation. The current breakthrough learned bounded memory model (Toshniwal et al., 2020) keeps learning rate as $2 * 10^{-4}$ with a maximum of 30 epochs. This is not feasible in computer systems with limited memory resources. Hence, we increase the learning rate to $1.5 * 10^{-4}$ and decrease the number of epochs to a maximum of 10. Most importantly, we reduce the maximum sequence length parameter from 512 to 256 tokens which resolves the memory issue significantly.

6 Evaluation

To evaluate we utilized the standard CoNLL perl scripts for only the final evaluation, i.e. when training ends. The CoNLL evaluation is done when both the standard conll scripts and the ground truth conll data are accessible.

The table in Fig. 4 shows the results of our findings and comparisons. We achieved the best result of Dev F1 - 70.9% and Test F1 - 69.2% using 10 memory cells. When compared with the original BERT model on the Litbank dataset which attains Test F1 of 64.8%, we can clearly notice that our experiments have shown massive improvements.

The latest (Joshi et al., 2020) SpanBert model based on unbounded memory architecture with huge amounts of computational resources attains Dev F1 - 77.1% and Test F1 - 76.5%.

Obviously, their score will be much better than ours. Our most significant comparison is done with the current (Toshniwal et al., 2020) bounded memory architecture. We used 256 maximum sequence length of the document as opposed to their 512 maximum sequence length. We also choose the maximum of 10 memory cells rather than 30 memory cells. This was mainly because the current advance model cannot run on a simple day-to-day system and also not on Google Colab which offers at most 13 GB RAM. Despite our limited resources, we were able to achieve almost similar DEV F1 and TEST F1 results.

Memory Model	Memory Cells	Dev F1	Test F1
Original BERT Score (Bamann et al.)	-	-	64.8
U-MEM(Joshi et al.)	Unbounded	77.1	76.5
LB-MEM (Toshniwal et al.)	10	71.9	70.3
	20	75.0	74.7
	30	75.7	75.1
LB-MEM (Our Score)	5	65.2	63.1
	7	67.7	66.3
	10	70.9	69.2

Table 4: Final Results on LitBank dataset

6.1 Discussion

The research aimed to implement coreference resolution using Google colab. We were able to achieve the objective with an F1 score similar to that of the most advanced models, but we were unable to improve the model's accuracy. To further improve the F1 score, we can replace the heuristic approach of the learned bounded memory architecture with a model that accurately predicts non-active entities in the memory and replace it with untracked tokens.

7 Conclusion

In our work, a memory model is presented that monitors a limited number of entities. Due to memory constraints, we only pre-processed, modelled, and experimented on 25 percent of the LitBank dataset. According to the findings we obtained, our more compact bounded memory model is still competitive when compared to the most sophisticated model. This model has a lot of potential for enhancing its outcomes by utilizing better GPUs, and increasing its memory structure architecture is also something that lies within the purview of future research and development.

References

Bamman, D., Lewke, O., Mansoor, A., 2020. An Annotated Dataset of Coreference in English Literature. https://doi.org/10.48550/arXiv.1912.01140

Clark, K., Manning, C.D., 2015. 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 (Volume 1: Long Papers). Presented at the ACL-IJCNLP 2015, Association for Computational Linguistics, Beijing, China, pp. 1405–1415. https://doi.org/10.3115/v1/P15-1136

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://doi.org/10.48550/arXiv.1810.04805

Heinzerling, B., Moosavi, N.S., Strube, M., 2017. Revisiting Selectional Preferences for Coreference Resolution. https://doi.org/10.48550/arXiv.1707.06456

Intuition of SpanBert, 2020. . GeeksforGeeks. URL https://www.geeksforgeeks.org/intuition-of-spanbert/ (accessed 12.13.22).

Joshi, M., Chen, D., Liu, Y., Weld, D.S., Zettlemoyer, L., Levy, O., 2020. SpanBERT: Improving Pre-training by Representing and Predicting Spans. https://doi.org/10.48550/arXiv.1907.10529

Joshi, M., Levy, O., Weld, D.S., Zettlemoyer, L., 2019. BERT for Coreference Resolution: Baselines and Analysis. https://doi.org/10.48550/arXiv.1908.09091

Karabiben, M., 2021. Three topics that will be the talk of the data town in 2021 [WWW Document]. Medium. URL https://towardsdatascience.com/three-topics-that-will-be-the-talk-of-the-data-town-in-2021-1bbe1067f85c (accessed 12.13.22).

Keller, F., 2010. Cognitively Plausible Models of Human Language Processing, in: Proceedings of the ACL 2010 Conference Short Papers. Presented at the ACL 2010, Association for Computational Linguistics, Uppsala, Sweden, pp. 60–67.

Liu, F., Perez, J., 2017. Gated end-to-end memory networks, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. pp. 1–10.

Mohan, M., Nair, J.J., 2019. Coreference Resolution in Ambiguous Pronouns Using BERT and SVM. 2019 9th Int. Symp. Embed. Comput. Syst. Des. ISED 1–5. https://doi.org/10.1109/ISED48680.2019.9096245

Paweł Mielniczuk [WWW Document], n.d. . Medium. URL https://medium.com/@p.mielniczuk (accessed 12.13.22).

Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L., 2018. Deep contextualized word representations. https://doi.org/10.48550/arXiv.1802.05365

Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., Zhang, Y., 2012. CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes, in: Joint Conference on EMNLP and CoNLL - Shared Task. Presented at the CoNLL 2012, Association for Computational Linguistics, Jeju Island, Korea, pp. 1–40.

Recasens, M., Hovy, E., Martí, M.A., 2010. A Typology of Near-Identity Relations for Coreference (NIDENT), in: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). Presented at the LREC 2010, European Language Resources Association (ELRA), Valletta, Malta.

Sukthanker, R., Poria, S., Cambria, E., Thirunavukarasu, R., 2020. Anaphora and coreference resolution: A review. Inf. Fusion 59, 139–162. https://doi.org/10.1016/j.inffus.2020.01.010

Team, T.A., n.d. Unraveling Coreference Resolution in NLP! – Towards AI. URL https://towardsai.net/p/nlp/c-r, https://towardsai.net/p/nlp/c-r (accessed 12.13.22).

Toshniwal, S., Wiseman, S., Ettinger, A., Livescu, K., Gimpel, K., 2020. Learning to Ignore: Long Document Coreference with Bounded Memory Neural Networks.

Webster, K., Curran, J.R., 2014. Limited memory incremental coreference resolution, in: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. Presented at the COLING 2014, Dublin City University and Association for Computational Linguistics, Dublin, Ireland, pp. 2129–2139.

Wu, W., Wang, F., Yuan, A., Wu, F., Li, J., 2020. Coreference Resolution as Query-based Span Prediction.

A. Rahman and V. Ng, "Narrowing the modeling gap: A cluster-ranking approach to coreference resolution". Journal of Artificial Intelligence Research, 2011, 40, pp.469-521.