Comparative Exploration of Coreference Resolution with Transformer Models

Aditya Landge*, Advisor: Cecilia O. Alm, Ph.D.
Rochester Institute of Technology
*al2960@rit.edu

1. Introduction

Coreference resolution is the task of automatically determining the chain of expressions that refer to the same entity or *antecedent.*

Example 1: Coreference chain
“Barack Obama nominated **Hillary Rodham Clinton** as his **secretary of state** on Monday. He chose **her** because **she** had foreign affairs experience as a former **First Lady**.”
Chain: secretary of state, her, she, First Lady

Example 2: QA with pronoun-antecedent: *What is too {big,small}?*
2a) The trophy would not fit in the suitcase, because it was too **big**.
2b) The trophy would not fit in the suitcase, because it was too **small**.

Transformers and BERT:

- Google developed pretraining technique, with attention mechanism
- Learns language syntax and lexical semantics (words that co-occur).
- Can be *fine-tuned* for tasks such as MT, NER, QA, etc.
- **This project explored the use of BERT variants, focusing on Transformer models for neural coreference resolution.**

2. Goals

- Understand the coreference resolution problem which involves identifying a candidate entity given a reference from a corpus. The task can also highlight biases in a corpus and trained models.
- Understand the Transformer architecture, and especially BERT [2].
- Adapt BERT and ALBERT [4] variants and compare the performance.

3a. Datasets

ParCorFull [5] is a German-English parallel corpus (so-called bitext) with full coreference annotation, given 3 sources: **News**, **TED Talks**, and Discourse-Oriented Statistical Machine Translation (**DiscoMT**).

English	German
She wants to, you know, find Obi Wan Kenobi. He's her only hope.	Sie will Obi Wan Kenobi finden. Er ist ihre einzige Hoffnung.

Example coreference chains/clusters found by SpanBERT* [3]:

[((0, 6), 'Victoria Chen , CFO of Megabucks Banking'), ((7, 8), 'her'), ((14, 16), 'the 38 - year - old'), ((25, 26), 'she')] “Victoria Chen, CFO of Megabucks Banking, saw her pay jump to $2.3 million, as the 38-year-old became the company's president. It is widely known that she came to Megabucks from rival Lotsabucks.“

*SpanBERT is a redesign of BERT architecture where continuous tokens are masked instead of random tokens.

3b. Datasets

GAP [6] is a gender-balanced QA dataset sampled from Wikipedia (n ≈ 5000). Each row has: (1) a sentence, (2) an ambiguous pronoun (3) two candidate entities the could refer to the pronoun.

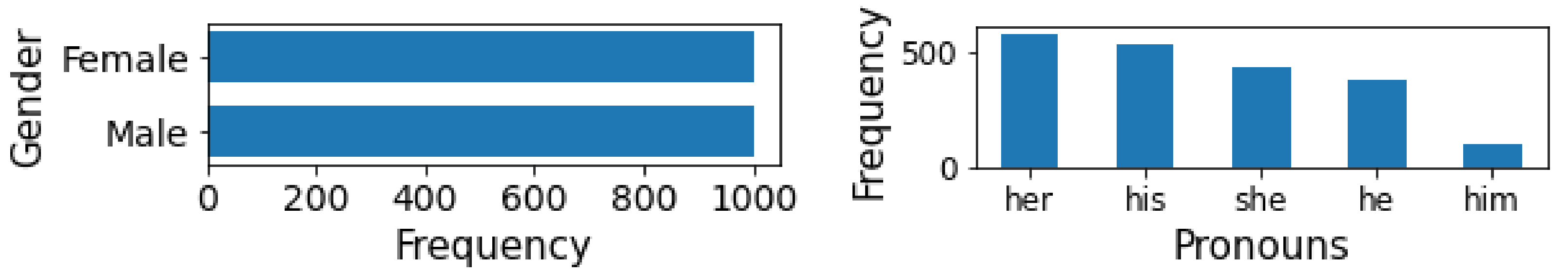
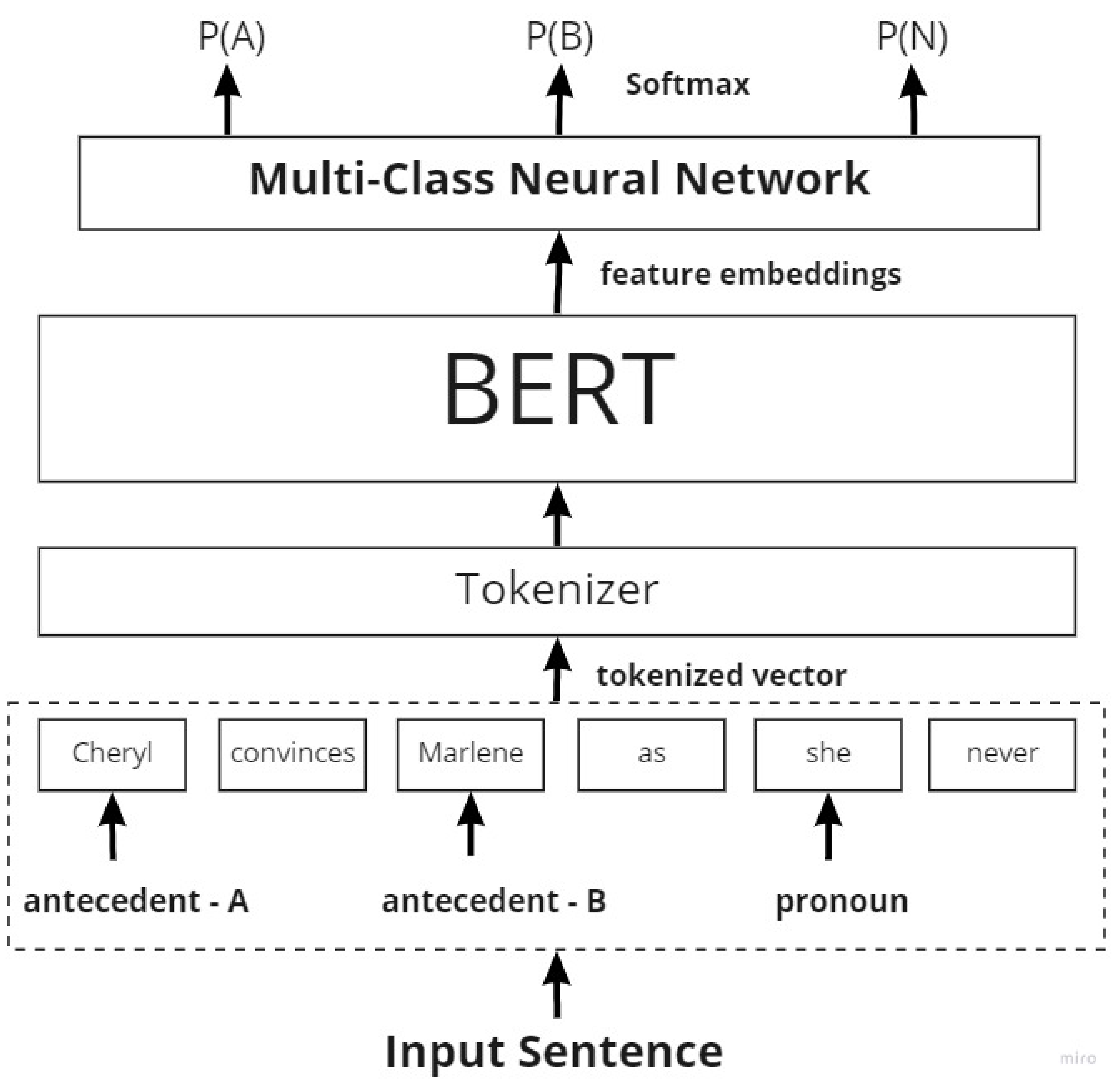


Fig. 1: Left - Gender-balanced data. Right - Object pronouns dominate.

Table 1: Example of a data instance from the GAP dataset.

Text	A	B	Pronoun
Phoebe played **Cheryl Cassidy**, **Pauline**'s friend and also a year 11 pupil in Simon's class. dumped **her** boyfriend ...	Cheryl Cassidy (True)	Pauline (False)	her

4. Model: BERT (base) + NN



P(A) P(B) P(N)

Softmax

Multi-Class Neural Network

feature embeddings

BERT

Tokenizer

tokenized vector

Cheryl convinces Marlene as she never

antecedent - A antecedent - B pronoun

Input Sentence

Fig. 2: BERT vs ALBERT models are trained, incl. BERT (base) + NN.

5. Results

Table 2: Exapmle outputs of BERT (base) + NN on the GAP dataset.

Text	A	B	N
... and Olympic-medalist Bob Suter are Dehner's uncles. His cousin is ...	Bob Suter P(A)=.84 Incorrect	Dehner P(B)=.16 Correct	Neither P(N)≈0 Incorrect
... Grenfell's career as a monologuist was directly inspired by Draper. Her nephew ...	Grenfell P(A)=.99 Correct	Draper P(B)≈0 Incorrect	Neither P(N)≈0 Incorrect
... Swedish divas Robyn and Lykke Li. Perry did a great job of letting us know she's ...	Robyn P(A)=.00 Incorrect	Lykke P(B)≈.1 Incorrect	Neither P(N)≈.99 Correct

Table 3: Performance of BERT and its variants on the GAP dataset.

Model	Accuracy	Precision	Recall
Mention-ranking baseline	0.17	0.46	0.37
ALBERT (base) Vanilla	0.59	0.51	0.55
ALBERT (base) + NN	0.53	0.53	0.53
CorefMulti BERT (base)	0.56	0.54	0.56
BERT (base) Vanilla	0.63	0.55	0.63
BERT (base) + NN	**0.77**	0.77	**0.76**
Late Fusion Model	0.50	**0.80**	0.48
CorefMulti BERT (large) [1]	0.87	0.87	0.87

Table 4: Training time BERT vs. ALBERT on Tesla P100-PCIE-16GB.

	BERT	ALBERT
Model	bert-base-uncased	albert-base-v2
Time to train	15 minutes	10 minutes

6. Conclusion and Future Work

- BERT finetuned with multiclass classifiers improves performance substantially for GAP. The much larger CorefMulti BERT does best.
- BERT variants like ALBERT allows parameter sharing to reduce model size and take slightly less time to finetune but performs worse.
- Late fusion integrating BERT and ALBERT improves precision only.
- Next steps involves further work to resolve full coreference chains.
- Next steps involves implementing SpanBERT like model for non-english languages like, German, etc..

7. References

[1] Rakesh Chada. “Gendered Pronoun Resolution using BERT and an Extractive Question Answering Formulation”. In: *First Workshop on Gender Bias in Natural Language Processing.*

[2] Jacob Devlin u. a. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.*

[3] Mandar Joshi u. a. “SpanBERT: Improving Pre-training by Representing & Predicting Spans”. In:

[4] Zhenzhong Lan u. a. *ALBERT: A Lite BERT for Self-supervised Learning of Language Representations.* 2020. arXiv: 1909.11942 [cs.CL].

[5] Ekaterina Lapshinova-Koltunski, Cristina España-Bonet und Josef van Genabith. “Analysing Coreference in Transformer Outputs”. In:

[6] Kellie Webster u. a. *Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns.* 2018. arXiv: 1810.05201 [cs.CL].