Gendered Pronoun Resolution

Dhruv Patel (14528) Prateek Sachan (15754)

Computer Science and Automation, Indian Institute of Science

May 2, 2019

Problem

- We are given a sentence or sentences, two candidate nouns within a sentence and a pronoun

Problem

- We are given a sentence or sentences, two candidate nouns within a sentence and a pronoun
- Task is to identify which of these two candidates refers to that pronoun

Problem

- We are given a sentence or sentences, two candidate nouns within a sentence and a pronoun
- Task is to identify which of these two candidates refers to that pronoun

Example

Impressed by her beauty, her warrior skills, and the fact that she was able to locate him, she is promoted to a position similar to that later held by her half-sister, Talia. As a right hand associate, she accompanies him during his adventures. Ra's is so impressed with her abilities, he even allows Nyssa to use his Lazarus Pits. Like **her** sister **Talia**, **Nyssa** eventually becomes disenchanted with Ra's genocidal plans to "cleanse the Earth", and disassociates herself from her father sometime in the early 20th century.

Problem

- We are given a sentence or sentences, two candidate nouns within a sentence and a pronoun
- Task is to identify which of these two candidates refers to that pronoun

Example

Impressed by her beauty, her warrior skills, and the fact that she was able to locate him, she is promoted to a position similar to that later held by her half-sister, Talia. As a right hand associate, she accompanies him during his adventures. Ra's is so impressed with her abilities, he even allows Nyssa to use his Lazarus Pits. Like **her** sister **Talia**, **Nyssa** eventually becomes disenchanted with Ra's genocidal plans to "cleanse the Earth", and disassociates herself from her father sometime in the early 20th century.

- Introduced as Kaggle competition by Google AI

Datasets

Datasets

GAP Webster et al. (2018)[1]

Datasets

GAP Webster et al. (2018)[1]

- Balanced Dataset

Datasets

GAP Webster et al. (2018)[1]

- Balanced Dataset
- 2000 Development, 2000 Test and 454 validation sentences

Datasets

GAP Webster et al. (2018)[1]

- Balanced Dataset
- 2000 Development, 2000 Test and 454 validation sentences

  Kathleen *first appears when* **Theresa** *visits* **her** *in a prison in London.*

Datasets

GAP Webster et al. (2018)[1]

- Balanced Dataset
- 2000 Development, 2000 Test and 454 validation sentences

  Kathleen *first appears when* **Theresa** *visits* **her** *in a prison in London.*

Definite Pronoun Resolution, Rahman et al. (2012), [2]

Datasets

GAP Webster et al. (2018)[1]

- Balanced Dataset
- 2000 Development, 2000 Test and 454 validation sentences

  Kathleen *first appears when* **Theresa** *visits* **her** *in a prison in London.*

Definite Pronoun Resolution, Rahman et al. (2012), [2]

- Unbalanced dataset

Datasets

GAP Webster et al. (2018)[1]

- Balanced Dataset
- 2000 Development, 2000 Test and 454 validation sentences

  Kathleen *first appears when* **Theresa** *visits* **her** *in a prison in London.*

Definite Pronoun Resolution, Rahman et al. (2012), [2]

- Unbalanced dataset
- 1886 sentences in total (943 sentence pairs) (When used, we used all for training)

Datasets

GAP Webster et al. (2018)[1]

- Balanced Dataset
- 2000 Development, 2000 Test and 454 validation sentences

  Kathleen *first appears when* **Theresa** *visits* **her** *in a prison in London.*

Definite Pronoun Resolution, Rahman et al. (2012), [2]

- Unbalanced dataset
- 1886 sentences in total (943 sentence pairs) (When used, we used all for training)

  James asked **Robert** *for a favor, but* **he** *refused.*
  James *asked Robert for a favor, but* **he** *was refused.*

Data Augmentation

- 2000 sentences, two less data to train.

Data Augmentation

- 2000 sentences, two less data to train.
- Our models quickly overfitted. We were unable to pass our own baseline.

Data Augmentation

- 2000 sentences, two less data to train.
- Our models quickly overfitted. We were unable to pass our own baseline.

Hypothesis

To neural network, if only "Jon Snow doesn't know anything" is given, the effect should be similar to when "John Wick doesn't know anything" is given instead.

Data Augmentation

- 2000 sentences, two less data to train.
- Our models quickly overfitted. We were unable to pass our own baseline.

Hypothesis

To neural network, if only "Jon Snow doesn't know anything" is given, the effect should be similar to when "John Wick doesn't know anything" is given instead.

So we replace all occurrences of Jon with John and of Snow with Wick. There could be same occurrence for different **unrelated** objects by chance. We assume that, that doesn't happen often.

Data Augmentation

- 2000 sentences, two less data to train.
- Our models quickly overfitted. We were unable to pass our own baseline.

Hypothesis

To neural network, if only "Jon Snow doesn't know anything" is given, the effect should be similar to when "John Wick doesn't know anything" is given instead.

So we replace all occurrences of Jon with John and of Snow with Wick. There could be same occurrence for different **unrelated** objects by chance. We assume that, that doesn't happen often.
As a side effect Ramsey Snow will become Ramsey Wick. In most cases this is not a problem.

Method

- If both candidate A and candidate B has less than four words and neither of them contains characters from ",(*)"
  - If pronoun is **he**, **him** or **his**
    - Find alternative male candidate of same length as A, such that no word of old A or old B is contained in new proposal.
    - Find alternative male candidate of same length as B, such that no word of old A or old B is contained in new proposal.
  - If pronoun is **she**, **her** or **hers**
    - Find alternative female candidate of same length as A, such that no word of old A or old B is contained in new proposal.
    - Find alternative female candidate of same length as B, such that no word of old A or old B is contained in new proposal.
- if old A and old B had any common word, modify proposals to behave similarly to old candidates.
- replace old A with new A, replace old B with new B.

Example

- Tony Markham, a high school senior and the "Tall Dark Stranger" Betsy fell in love with as a freshman, who has since become a good friend not only to Betsy but the entire Ray family. Mrs. Ray, Betsy's mother. Mr. Ray, Betsy's father, who owns a shoestore. **Margaret Ray**, **Betsy**'s sister who is five years younger than she is.
- Tony Markham, a high school senior and the "Tall Dark Stranger" Alyssa fell in love with as a freshman, who has since become a good friend not only to Alyssa but the entire Jolie family. Mrs. Jolie, Alyssa's mother. Mr. Jolie, Alyssa's father, who owns a shoestore. Angelina Jolie, Alyssa's sister who is five years younger than she is.
- Tony Markham, a high school senior and the "Tall Dark Stranger" Booth fell in love with as a freshman, who has since become a good friend not only to Booth but the entire Delgado family. Mrs. Delgado, Booth's mother. Mr. Delgado, Booth's father, who owns a shoestore. Pam Delgado, Booth's sister who is five years younger than she is.

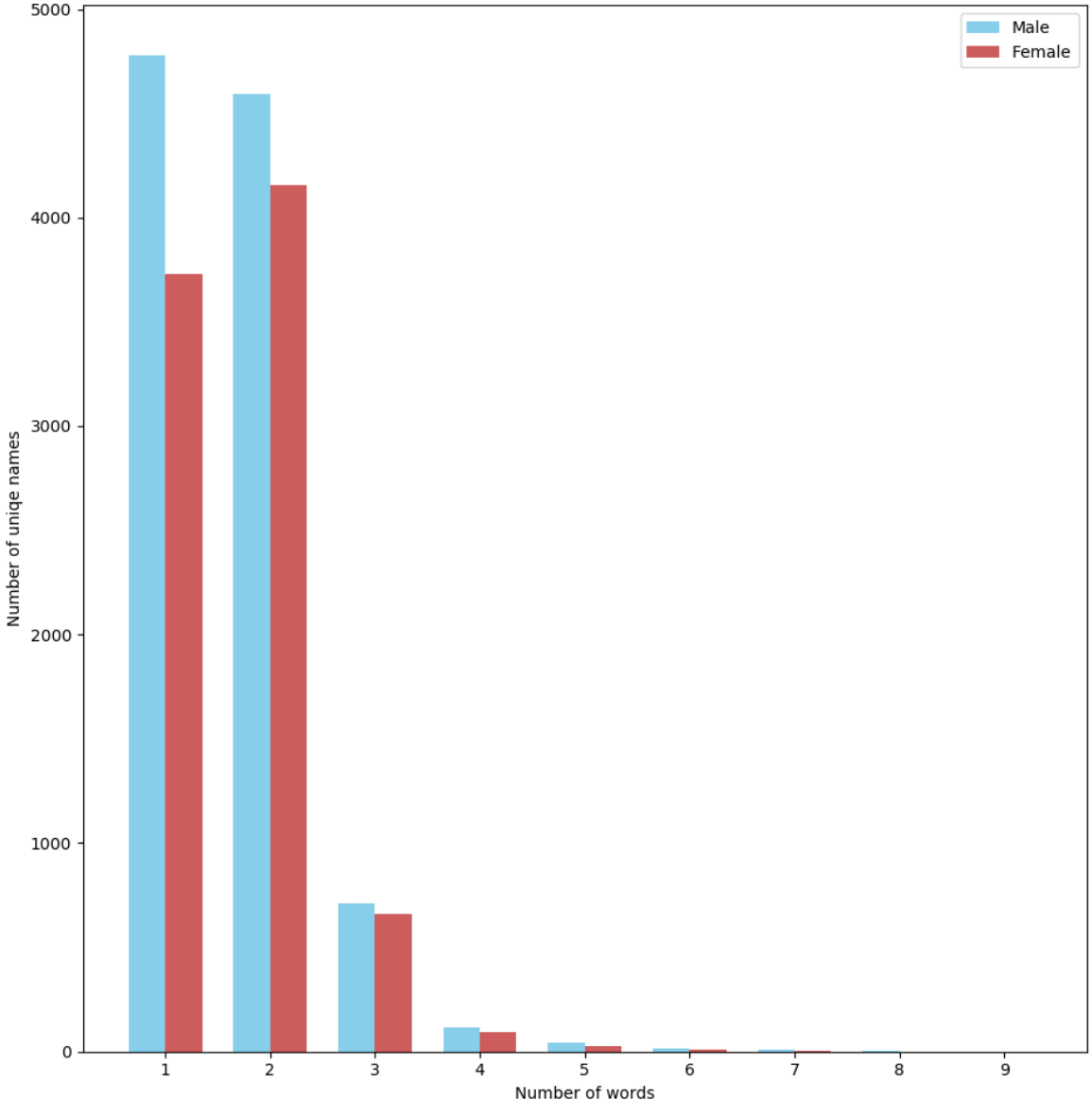


Figure: Distribution of names

Metrics

To compare our results with baseline proposed by Webster et al.[1], we use micro average of F1-scores. To compare our results with other Kaggle competitors, we used cross entropy loss $\mathcal{L}$.

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j\in\{A,B,N\}}(y_i^j * log(\sigma(\hat{y_i^j}))).$$

Where y_i^j is 1 if j is correct candidate for i^{th} example. N denotes neither case. $\hat{y_i^j}$ denotes predicted probability for class j for i^{th} example.

Method

- Earlier we started with ambitious goal. We chose all candidates using Named Entity Recognition.

Method

- Earlier we started with ambitious goal. We chose all candidates using Named Entity Recognition.
  - Class imbalance

Method

- Earlier we started with ambitious goal. We chose all candidates using Named Entity Recognition.
  - Class imbalance
  - Models didn't perform well. So we used simpler setting.

Method

- Earlier we started with ambitious goal. We chose all candidates using Named Entity Recognition.
  - Class imbalance
  - Models didn't perform well. So we used simpler setting.
- We adapted Lee et al's architecture [3].

Method

- Earlier we started with ambitious goal. We chose all candidates using Named Entity Recognition.
  - Class imbalance
  - Models didn't perform well. So we used simpler setting.
- We adapted Lee et al's architecture [3].
  - No need for mention scoring.

Method

- Earlier we started with ambitious goal. We chose all candidates using Named Entity Recognition.
  - Class imbalance
  - Models didn't perform well. So we used simpler setting.
- We adapted Lee et al's architecture [3].
  - No need for mention scoring.
  - Different scoring functions were tried. Best we found was three layer fully connected network.

Architecture

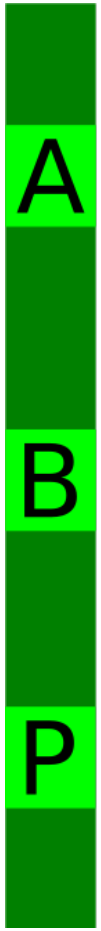


Figure: Architecture

Architecture

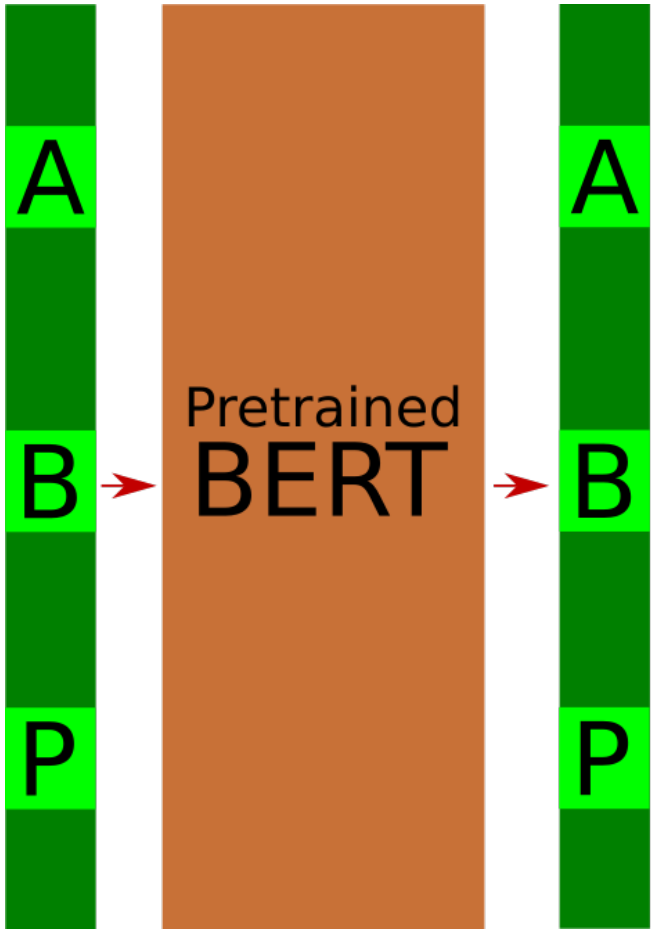


Figure: Architecture

Architecture

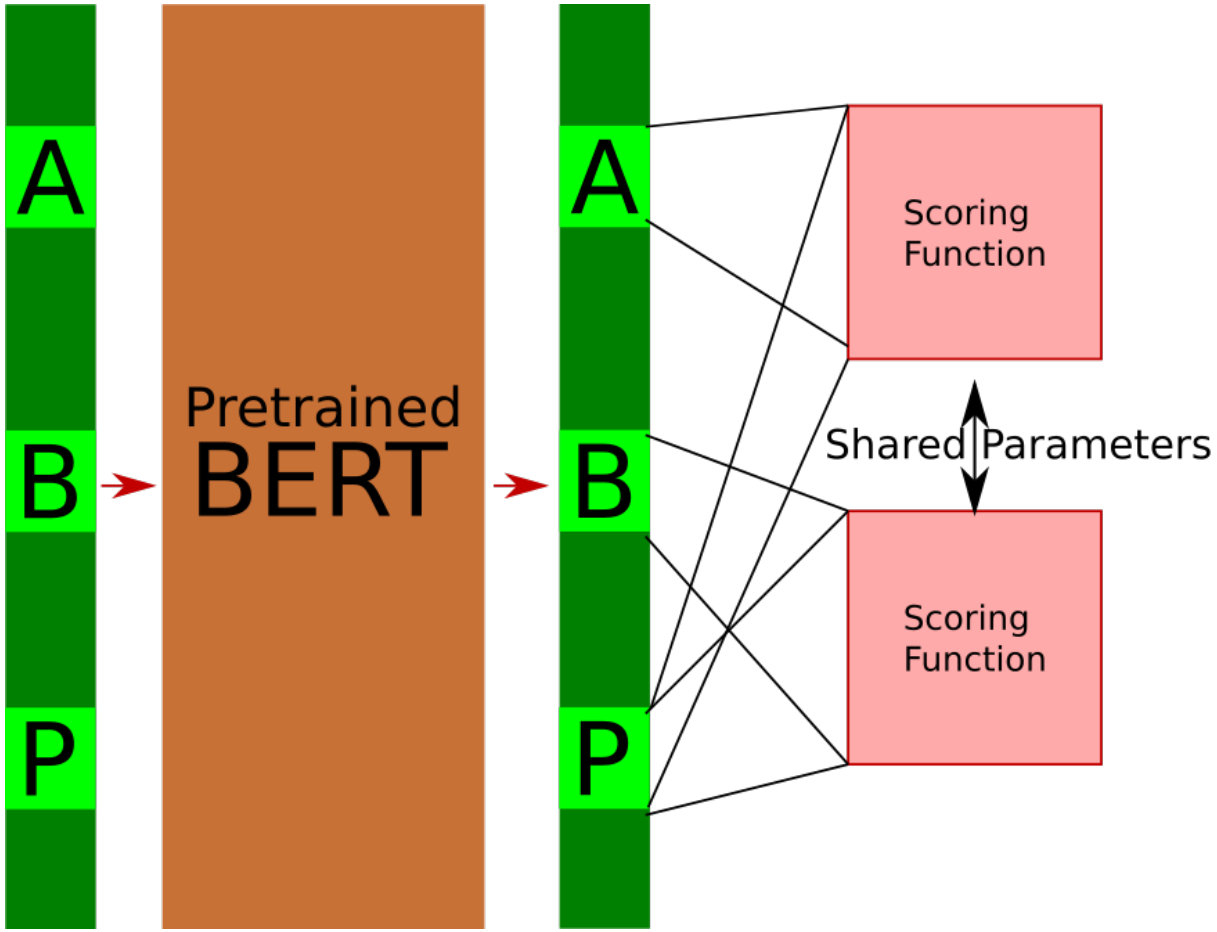


Figure: Architecture

Architecture

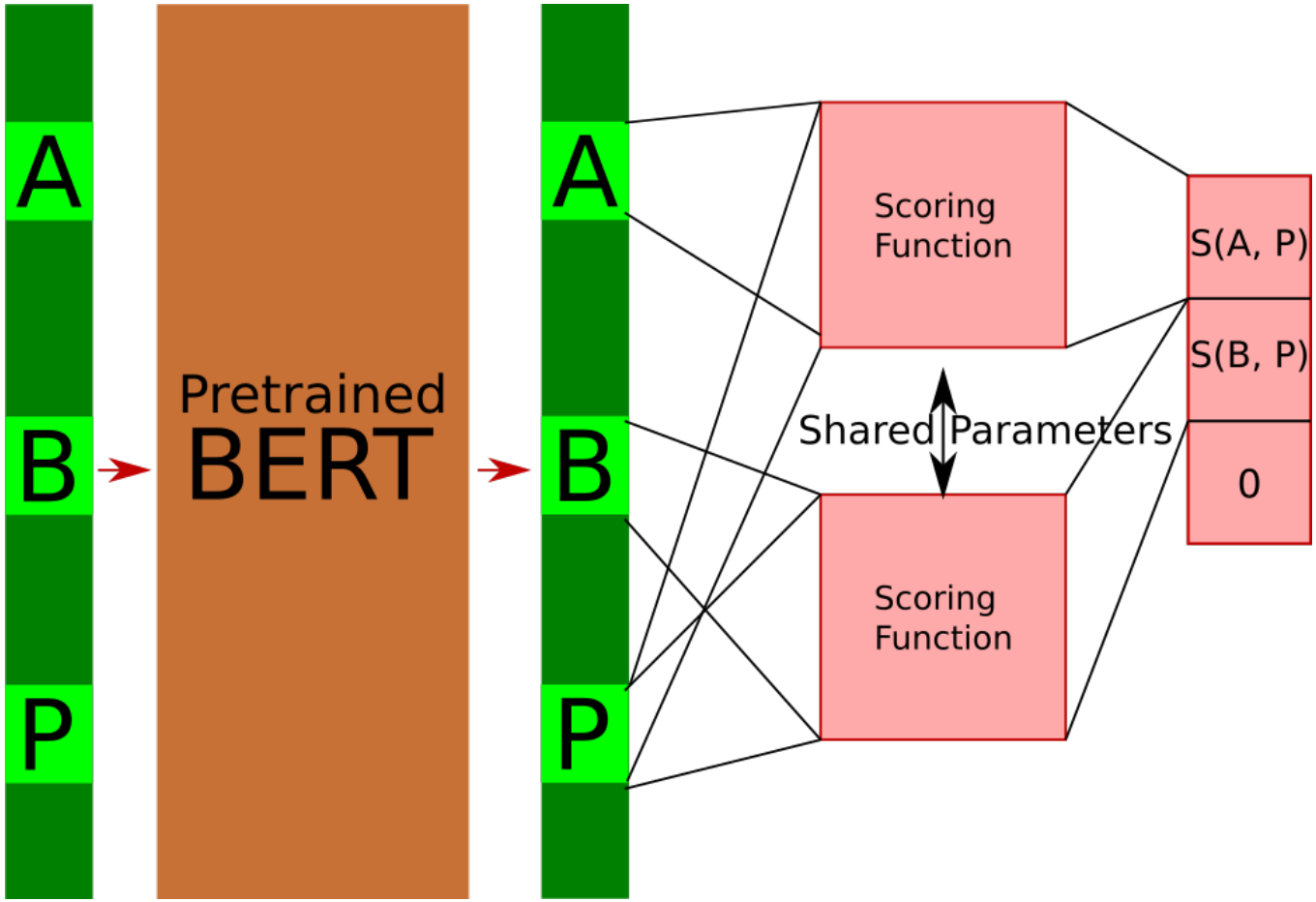


Figure: Architecture

Experiments

- Different layers of BERT large and base model were tried.

Experiments

- Different layers of BERT large and base model were tried.
- With and without DPR dataset

Experiments

- Different layers of BERT large and base model were tried.
- With and without DPR dataset
- Scoring function used three linear layers of size 62, 32, and 1 respectively with ReLU activations.

Experiments

- Different layers of BERT large and base model were tried.
- With and without DPR dataset
- Scoring function used three linear layers of size 62, 32, and 1 respectively with ReLU activations.
- Also tried LSTM with single layer of 256 units between BERT and scoring function.

Experiments

- Different layers of BERT large and base model were tried.
- With and without DPR dataset
- Scoring function used three linear layers of size 62, 32, and 1 respectively with ReLU activations.
- Also tried LSTM with single layer of 256 units between BERT and scoring function.
- To combine different tokens of A and B, we tried attention mechanism and simple mean method.

Experiments

- Different layers of BERT large and base model were tried.
- With and without DPR dataset
- Scoring function used three linear layers of size 62, 32, and 1 respectively with ReLU activations.
- Also tried LSTM with single layer of 256 units between BERT and scoring function.
- To combine different tokens of A and B, we tried attention mechanism and simple mean method.
- Best results were obtained without RNN, with weight decay of 0.001 and dropout of 0.5. Adam optimization was used.

Baselines

From Webster et al. [1]

	M	F	O
Wiseman et al. [4]	68.4	59.9	64.2
Lee et al. [3]	67.2	62.2	64.7

Baselines

From Webster et al. [1]

	M	F	O
Wiseman et al. [4]	68.4	59.9	64.2
Lee et al. [3]	67.2	62.2	64.7

However not trained on GAP dataset. Lee et al. was trained on Onto Notes.

Baselines

From Webster et al. [1]

	M	F	O
Wiseman et al. [4]	68.4	59.9	64.2
Lee et al. [3]	67.2	62.2	64.7

However not trained on GAP dataset. Lee et al. was trained on Onto Notes.

Ours

An SVM trained on the output of 8th layer BERT base. (input to SVM was 768*3 dimensional vector)

	M	F	O
SVM-8	**77.10**	**79.10**	**78.10**

Stage 1

	M	F	O
SVM-8	77.10	79.10	78.10
MLP	89.10	88.00	88.55
MLP-attn	89.50	88.40	88.95
MLP-dpr	88.90	88.70	88.80
MLP-dpr-attn	90.10	87.90	89.00

Table: F1 score Results

	M	F	O
SVM-8	0.5127	0.5077	0.5102
MLP	0.2669	0.3412	0.3041
MLP-attn	0.2828	0.3252	0.3040
MLP-dpr	0.2752	0.3187	0.2969
MLP-dpr-attn	0.2706	0.3367	0.3036

Table: Cross Entropy Loss on stage 1 test set

Stage 2

	O
SVM-8	0.7809
RNN-MLP	0.3529
MLP	0.2462
MLP-attn	0.2545
MLP-dpr	0.2727
MLP-dpr-attn	0.2517
Kaggle First	0.1366
Kaggle 25th	0.2403

Table: Cross Entropy Loss on stage 2 test set

	precision	recall	f1-score	support
A	0.8635	0.9533	0.9062	856
B	0.9219	0.8930	0.9072	925
Neither	0.8113	0.5890	0.6825	219

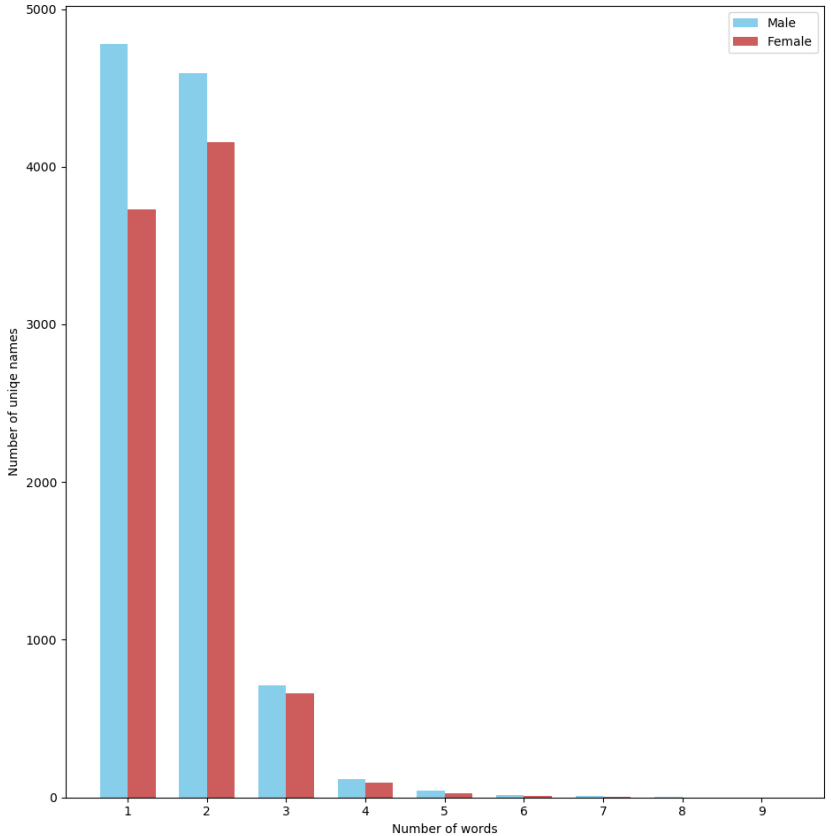
Table: Precision-Recall of MLP

Comments

- Lack of difference between attention mechanism and simple mean

Comments

- Lack of difference between attention mechanism and simple mean



Comments

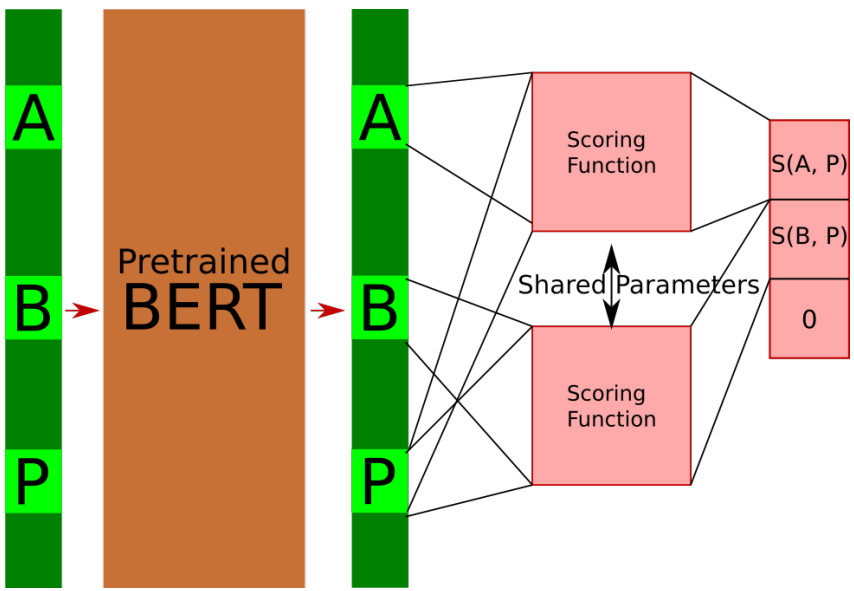
- Lack of difference between attention mechanism and simple mean
- Bias

Comments

- Lack of difference between attention mechanism and simple mean
- Bias
- Adaptability to more than two candidates

Comments

- Lack of difference between attention mechanism and simple mean
- Bias
- Adaptability to more than two candidates



Thank You

References

K. Webster, M. Recasens, V. Axelrod, and J. Baldridge, "Mind the gap: A balanced corpus of gendered ambiguou," in *Transactions of the ACL*, p. to appear, 2018.

A. Rahman and V. Ng, "Resolving complex cases of definite pronouns: the winograd schema challenge," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 777–789, Association for Computational Linguistics, 2012.

K. Lee, L. He, M. Lewis, and L. Zettlemoyer, "End-to-end neural coreference resolution," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, (Copenhagen, Denmark), pp. 188–197, Association for Computational Linguistics, Sept. 2017.

S. Wiseman, A. M. Rush, and S. M. Shieber, "Learning global features for coreference resolution," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (San Diego, California), pp. 994–1004, Association for Computational Linguistics, June 2016.