Injecting Gender Knowledge into Debiased Model

Faezeh Hosseini
Ali Sarmadi

Problem:

After debiasing, the model may lose its sense to gender

[MASK] is pregnant ⟶ He is pregnant

Model:

Zari(Debiased Bert-large)

Data:
Google Knowledge Graph

Triple:

('Vanessa Kerry', 'father', 'John Kerry')

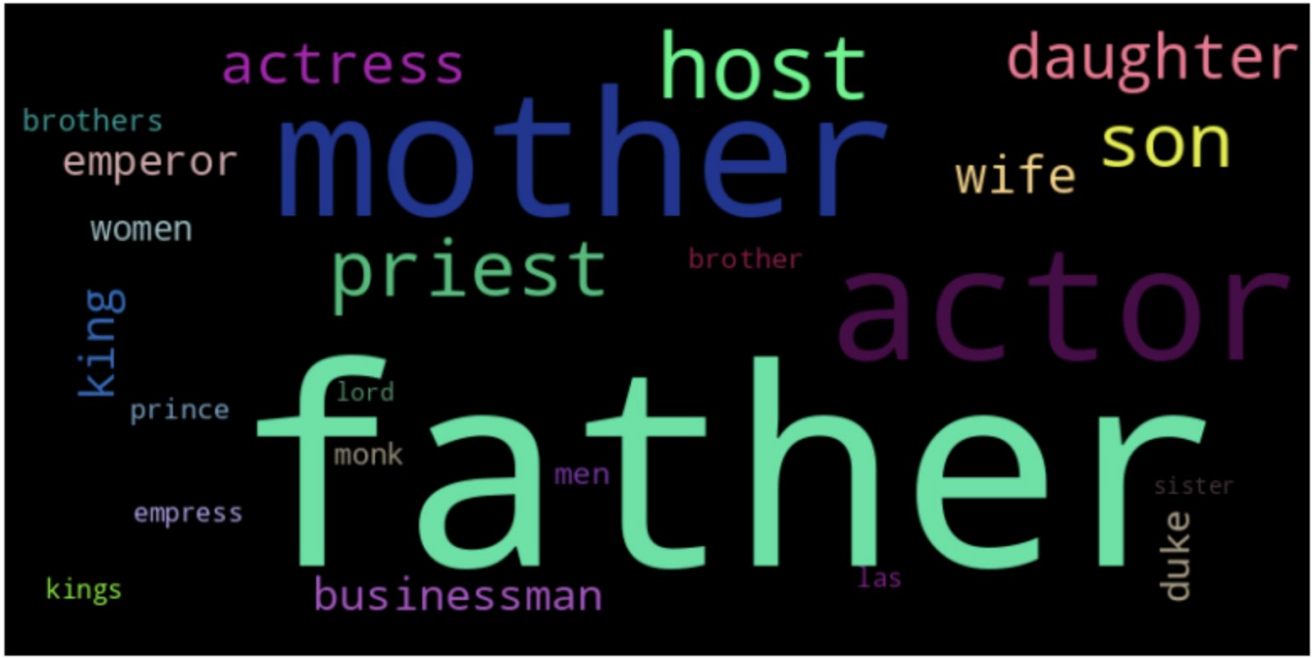
Subject **Relation** **Object**

Description:

Vanessa Kerry's father is John Kerry, who served as the 68th United States Secretary of State.

Gender Specific Knowledge DataSet

Filter the google knowledge graph dataset on 191 Gender-Specific keywords(109 one token)



19498 rows

Injecting GS-Knowledge to Zari through Soft-Prompts

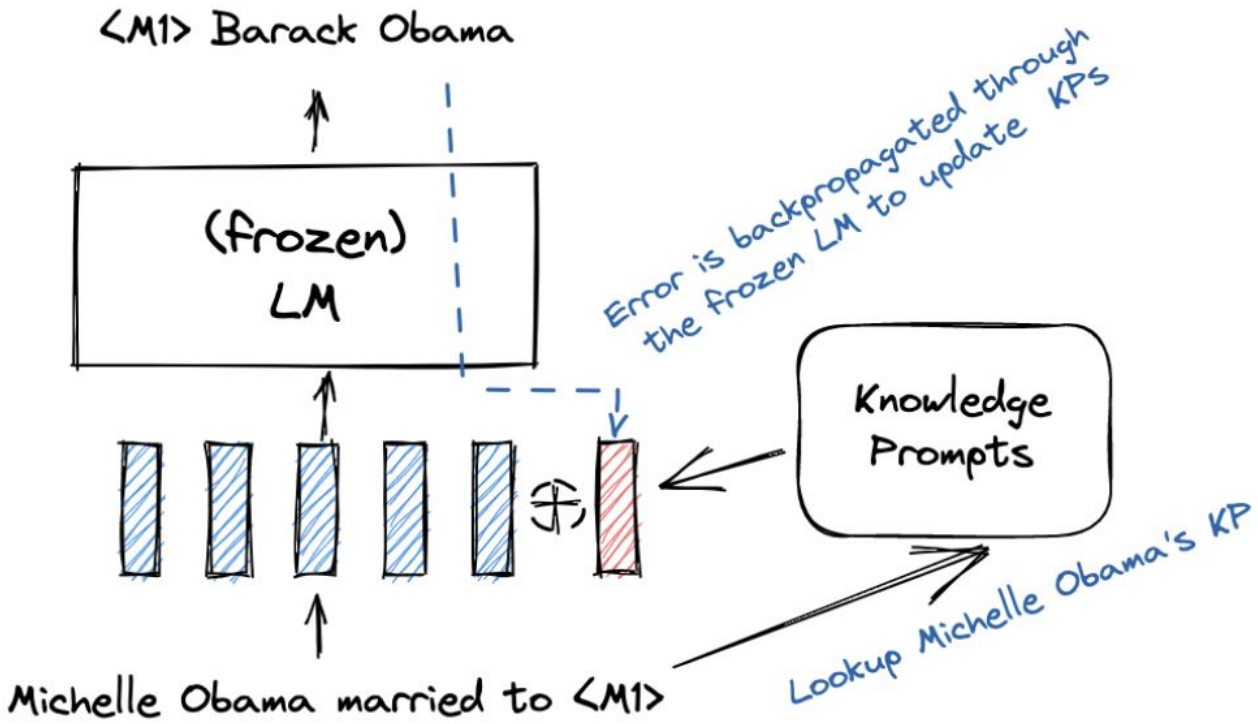
Input: (Tiples) : Vanessa Kerry [MASK] John Kerry

Label : father

Training:

1. Pass the subject or object to the Knowledge-Prompt.
2. Get the Soft-Prompt out of KP.
3. Add it to the input and input it to the MLM
4. Get the prediction of [MASK] and calculate the loss.
5. Backward the loss through KP.

Knowledge Prompts: Injecting World Knowledge into Language Models through Soft Prompts



Implementation of KP

Paper's suggestion(KP table): **not proper in our scale**

Progress report2:

First idea(Neural Network): **failed**

Progress report3:

Second idea(Bert word-embedding): **successful**

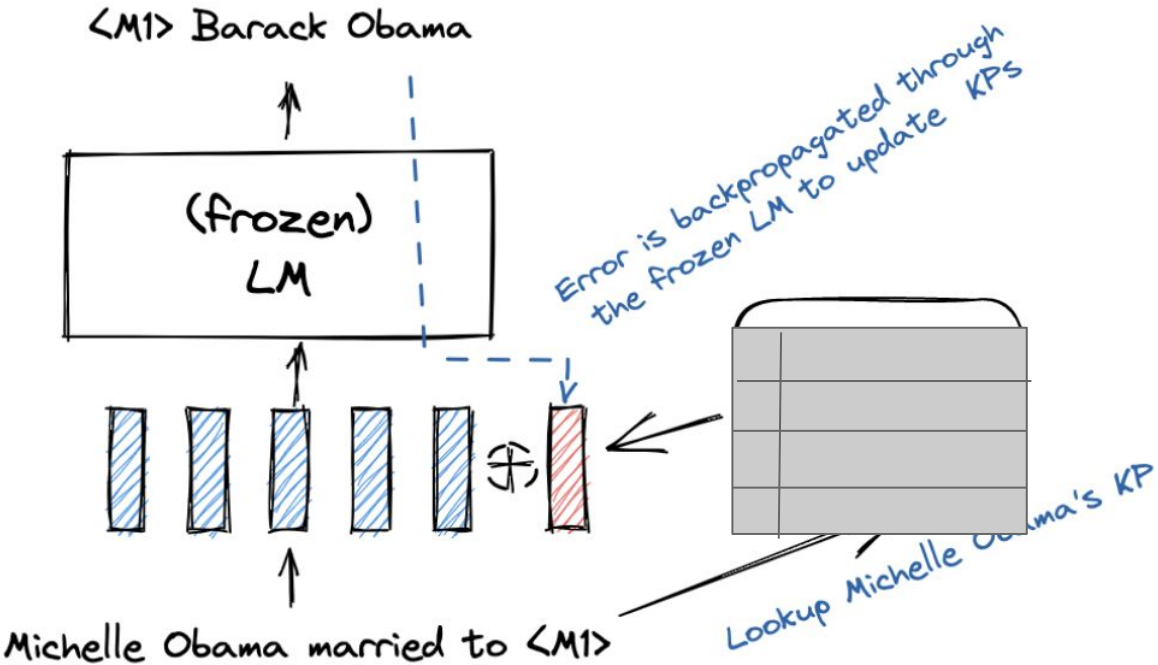
-But it changed the models entire embedding

Progress Report4:

Changing configs and handle the exceptions

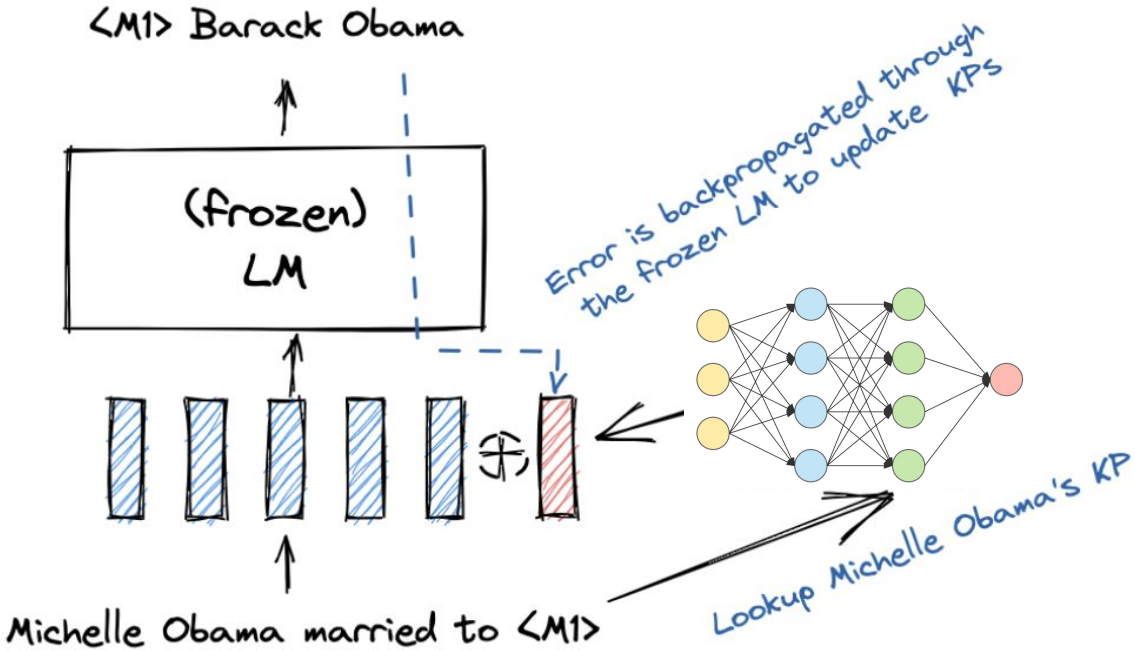
-Changed the embedding in the right place

KP structure in the paper



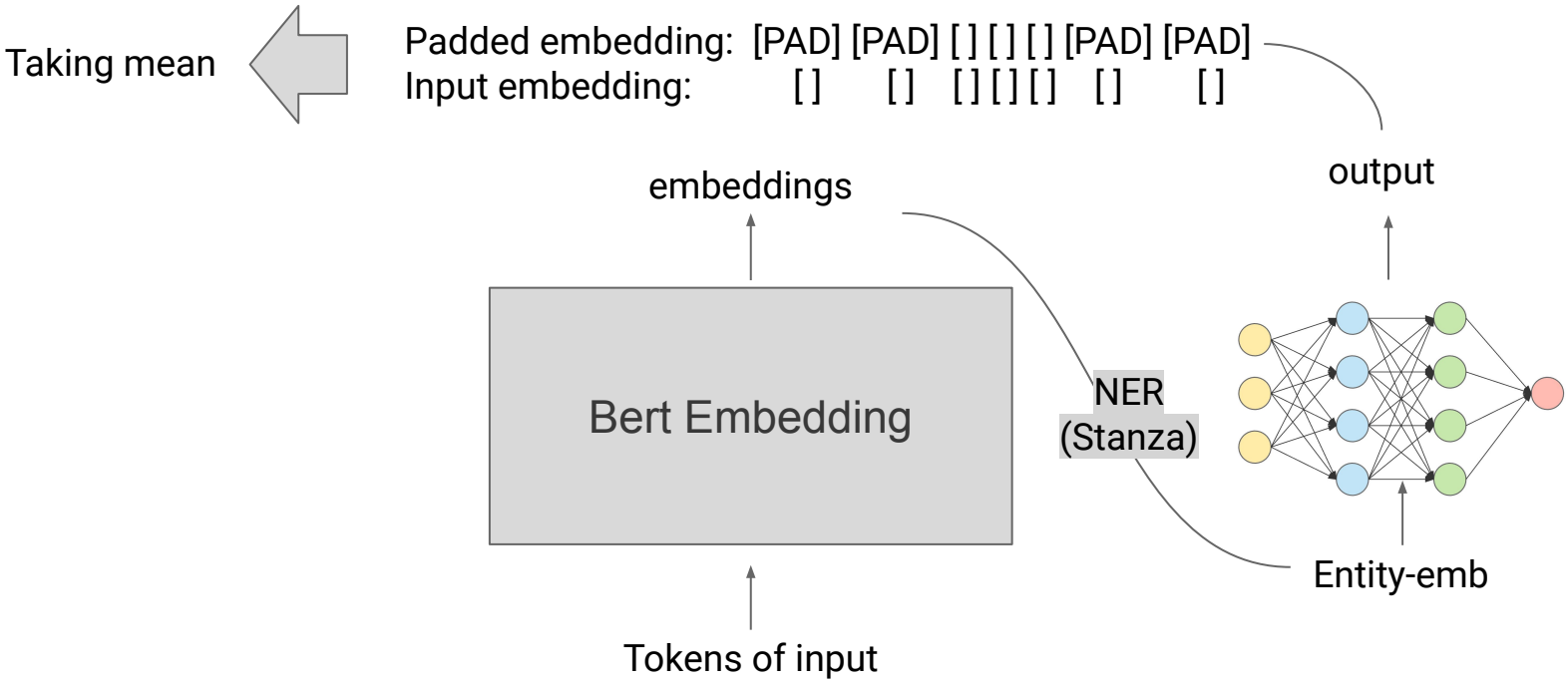
First idea:
Using Neural network as a KP

Progress Report 2:



Second idea:
Using NN as a KP in the Bert_embedding

Progress Report 3:



Improved Soft_embedding Module

Progress Report 4:

-Initialize learned_embedding in the Soft_embedding Module
    So the entire embedding won't change

-Handling the position alignment for combining the learned_embedding and input_embedding
    Different tokenization between Stanza and Bert-large-uncased tokenizers
    =Many to many relation

-handling Exceptions of Stanza tokenization:

Handling Exceptions of Stanza Tokenizer

Example:

william boleyn [MASK] geoffrey boleyn


```
{
    "id": 7,
    "text": "boley",
    "lemma": "boley",
    "upos": "X",
    "xpos": "FW",
    "head": 1,
    "deprel": "flat",
    "start_char": 31,
    "end_char": 36,
    "ner": "O"
  },
  {
    "id": 8,
    "text": "n",
    "lemma": "n",
    "upos": "PUNCT",
    "xpos": ".",
    "head": 1,
    "deprel": "punct",
    "start_char": 36,
    "end_char": 37,
    "ner": "O"
  }
```

Evaluation Results

Name	Gender Specific Score						Gender Neutral Score	Bias Score
	T1	T2	T3	T4	T5	Overall		
Zari Dropout	25.54%	43.06%	17.47%	21.36%	21.31%	25.40%	86.93%	39.31%
Zari Dropout SP	56.02%	53.73%	52.58%	59.14%	59.14%	55.60%	46.68%	50.75%
Zari cda	55.73%	79.66%	64.46%	50.97%	44.33%	59.12%	75.61%	66.36%
Zari cda SP	57.79%	55.70%	50.25%	61.36%	47.01%	56.50%	42.61%	48.58%

Evaluation Results

After 30 epochs with lr=0.001

Name	Gender Specific Score						Gender Neutral Score	Bias Score
	T1	T2	T3	T4	T5	Overall		
Zari cda	55.73%	79.66%	64.46%	50.97%	44.33%	59.12%	75.61%	66.36%
Zari cda SP	37.56%	52.28%	38.21%	33.07%	21.59%	37.41%	85.38%	52.31%

T1: Historical or contextual preservation
T2: Name replaced with pronoun or possessive adjective
T3: Name replaced with a gendered name
T4: Name replaced with a random name
T5: Biological fact indicating gender

Looking Forward

1. Train longer with more epochs so we can get better accuracy and Scores on this task

2. Extend the learned_embeddings for other gender-specific words and use them to inject more knowledge to the model

Thanks for your attention

References:

Knowledge Prompts: Injecting World Knowledge into Language Models through Soft Prompts:
https://arxiv.org/abs/2210.04726

Data:
Google Knowledge Graph
https://developers.google.com/knowledge-graph

Evaluation dataset and metric:
DIFAIR: A Benchmark for Disentangled Assessment of Gender Knowledge and Bias