UNIVERSITÄT PADERBORN

DATA SCIENCE RESEARCH GROUP

RECENT ADVANCES IN NATURAL LANGUAGE PROCESSING

TOPIC: INVESTIGATING ENTITY KNOWLEDGE IN BERT WITH SIMPLE NEURAL END-TO-END ENTITY LINKING

Presented by Siddhanth Janadri

Agenda

- Background
- Motivation
- Problem definition
- Approach
- Experiments and Results
- Discussion
- Conclusion

UNIVERSITÄT PADERBORN

Background

What is Entity Linking?

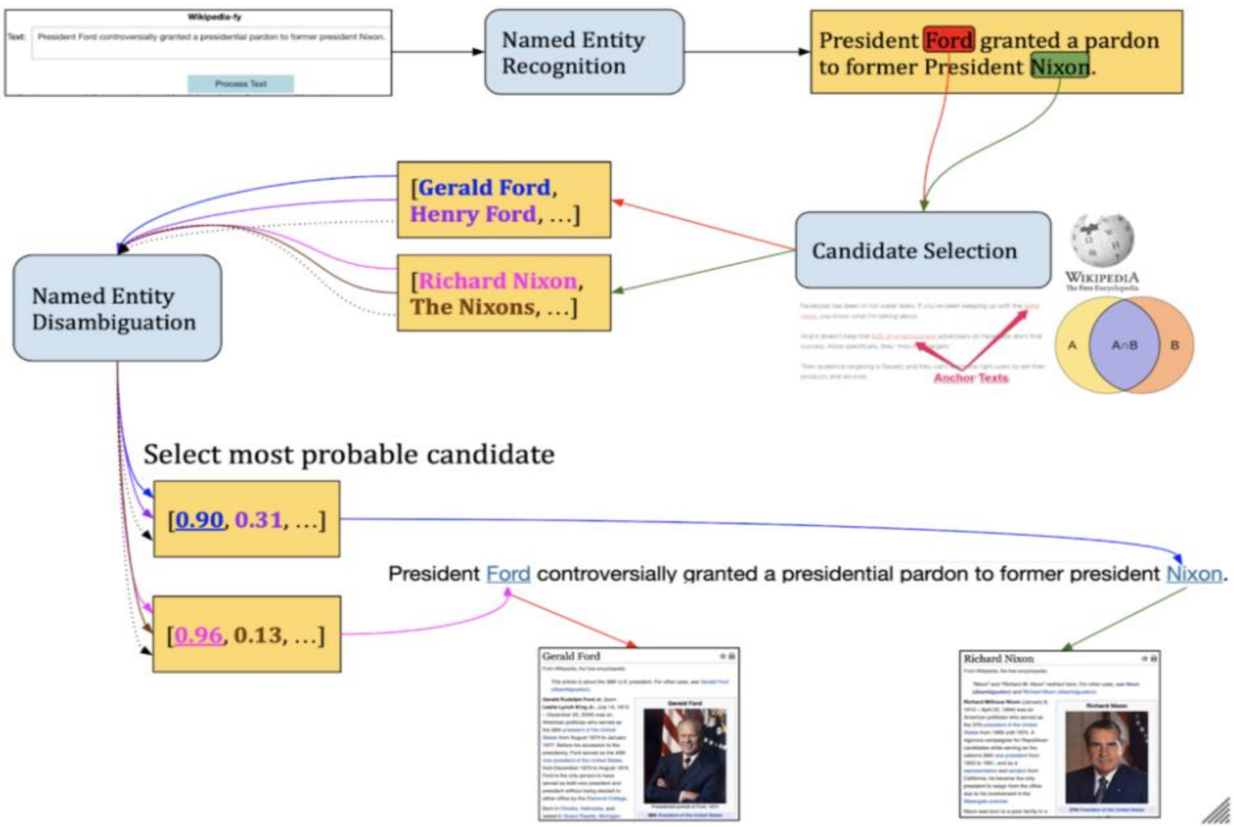

Figure 1: Overview of Entity linking[1]

1 https://towardsdatascience.com/named-entity-disambiguation-boosted-with-knowledge-graphs-4a93a94381ef

Background

What is Entity Linking?

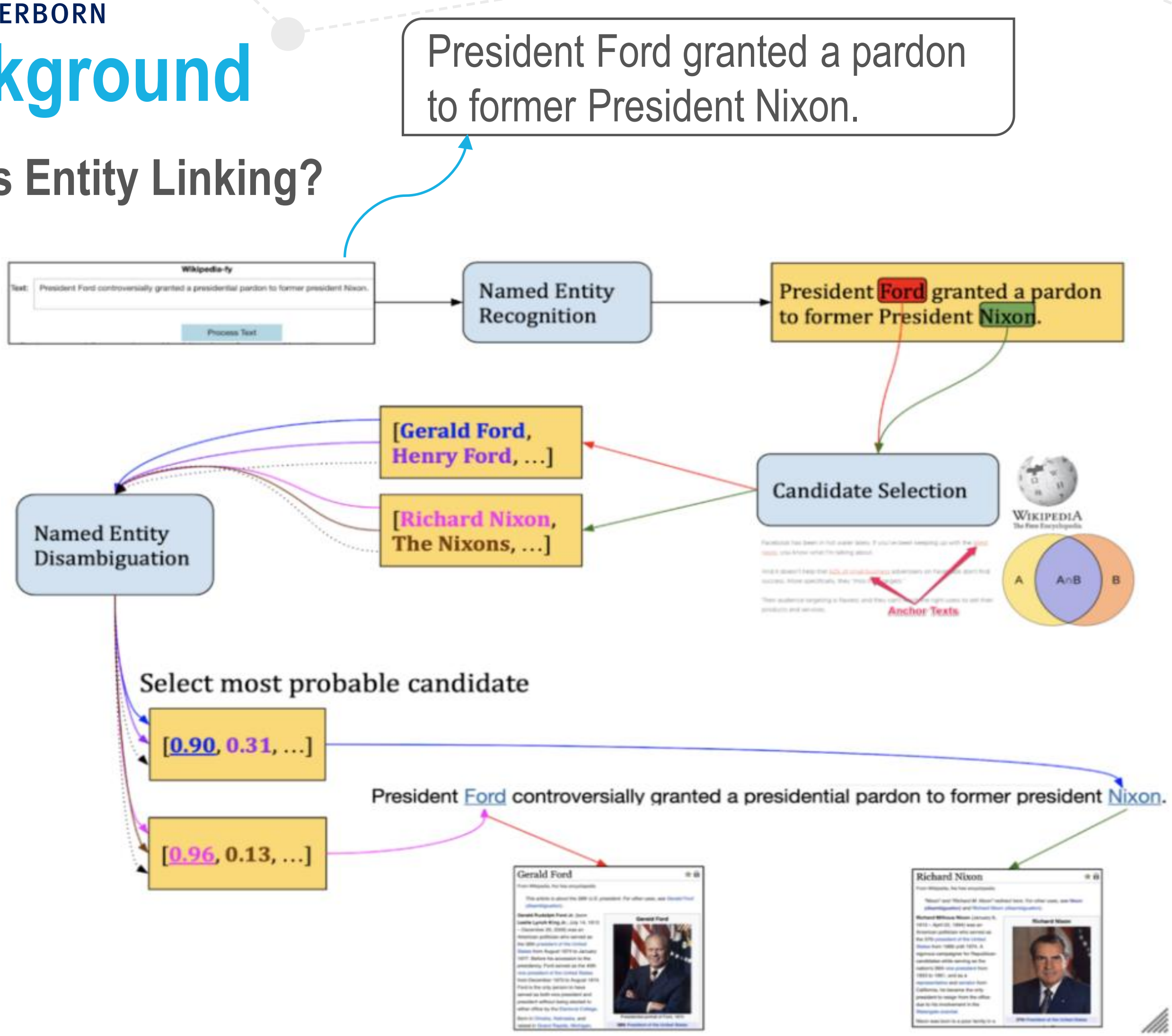

Figure 1: Overview of Entity linking[1]

1 https://towardsdatascience.com/named-entity-disambiguation-boosted-with-knowledge-graphs-4a93a94381ef

Background

What is Entity Linking?

Recognizing entity mentions in text and linking them to corresponding entries in a KB

- Named Entity Recognition (NER)
- Candidate Generation
- Entity Disambiguation

End-To-End Entity Linking: The process of performing all these tasks together as a single task leveraging mutual dependency.

UNIVERSITÄT PADERBORN

Background

What is BERT?

- BERT stands for Bidirectional Encoder Representations from Transformers
- It is pre-trained from unlabelled text[2]
- Learns contextual relations between words and sentences
- It is pre-trained on two prediction tasks:
  - Masked Language Modelling
  - Next sentence prediction

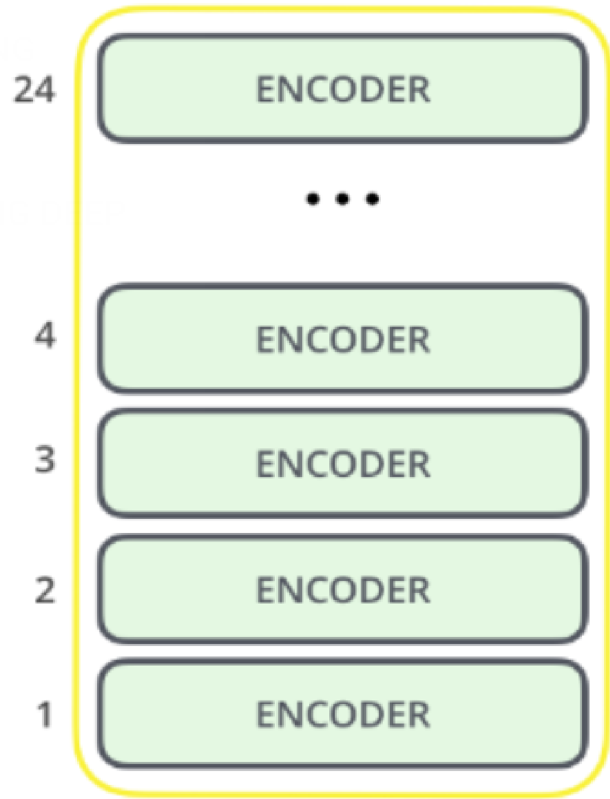

Fig 2: $BERT_{Large}$[1]

1 https://www.analyticsvidhya.com/blog/2019/09/demystifying-bert-groundbreaking-nlp-framework/?utm_source=blog&utm_medium=fine_tune_BERT

2 Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding.

Motivation

- BERT can be used for a wide variety of language tasks by only adding a small layer to the core model.
- This way it is possible to fine-tune BERT model that can be used as a purpose-specific model.

This paper explains the effects of fine-tuned BERT model[1].

- Is it possible for BERT's architecture to perform End-to-End Entity Linking?
- As BERT is a pre-trained model, how much entity knowledge is already present in it?
- Is it possible to improve the performance of BERT with additional entity knowledge?

1 Samuel Broscheit.2020. Investigating Entity Knowledge in BERT with Simple Neural End-To-End Entity Linking.

UNIVERSITÄT PADERBORN

Problem Definition

- BERT+Entity model – direct extension of BERT
- Additional output classification layer is added on top of it.
- Works on the principle of per token classification[1]
- The main ultimatum:
  - Generation of training data
- BERT-base-uncased model is used for experimentation that differs by token embedding size and self-attention layer depth

1 Samuel Broscheit.2020. Investigating Entity Knowledge in BERT with Simple Neural End-To-End Entity Linking.

Problem definition

Is it possible for BERT's architecture to perform End-to-End Entity Linking?

- This is dealt using per token classification over entire entity vocabulary.
- The fine-tuned model is later compared with baselines of entity linking for evaluation.

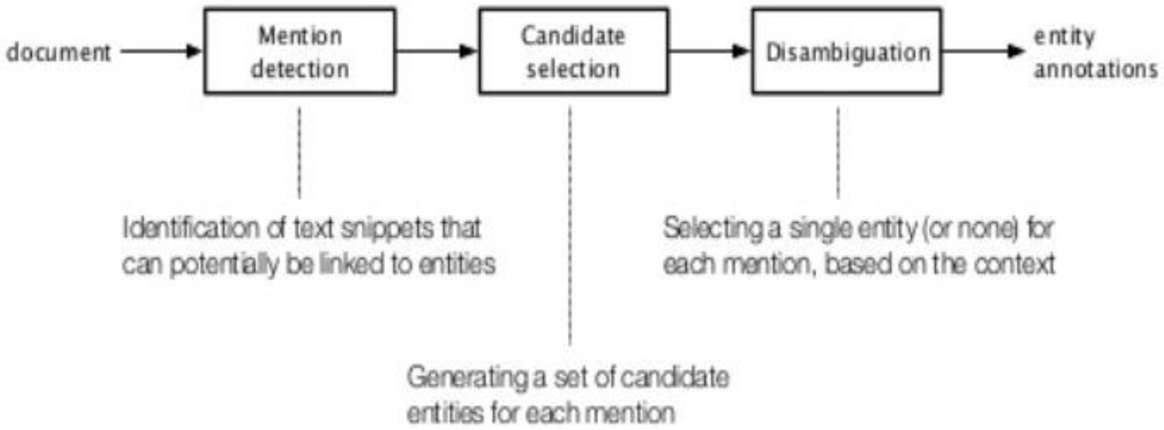

Figure 3: Entity linking[1]

1 https://www.slideshare.net/krisztianbalog/entity-linking-65308055

Problem definition

As BERT is a pre-trained model, how much entity knowledge is already present in it?

- Evaluated by training only classification layer of BERT+Entity model by freezing BERT.

Is it possible to improve the performance of BERT with additional entity knowledge?

- Improves performance by additional entity knowledge.
- Not beneficial for many tasks.

Approach

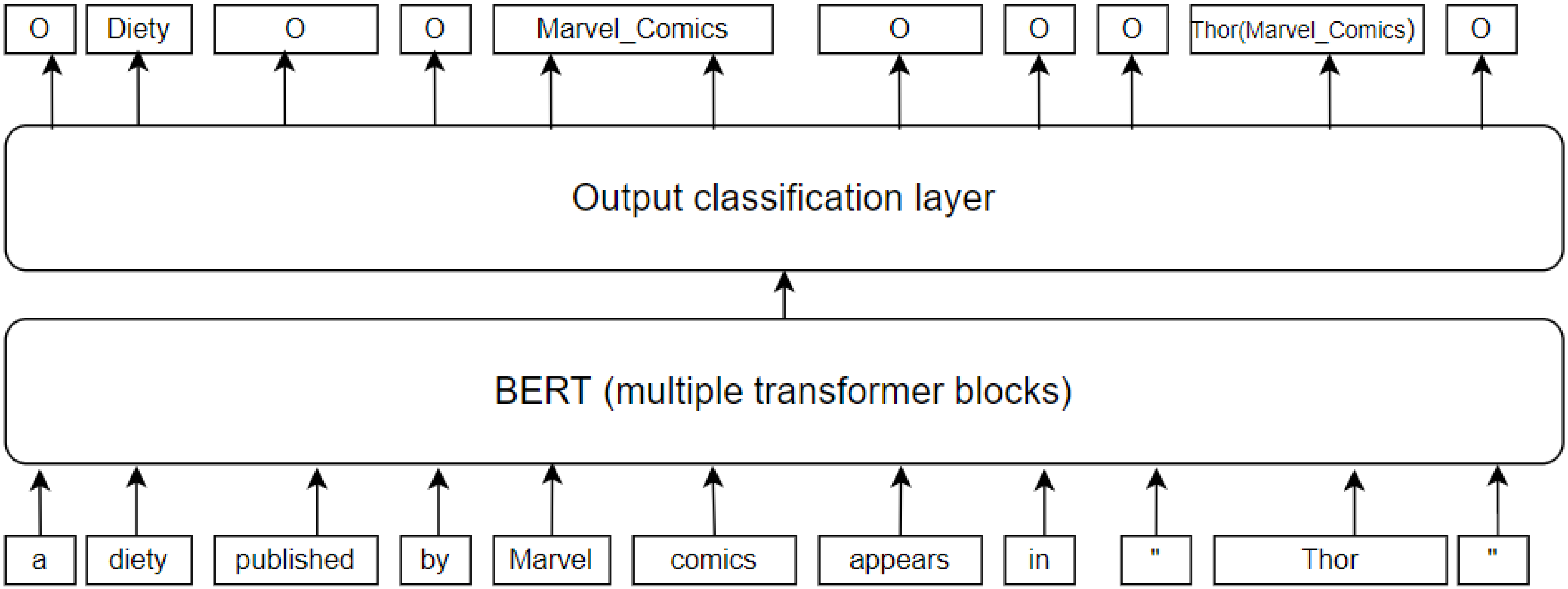

Figure 3: BERT+Entity model

This diagram shows how BERT+Entity model is linking Thor to Thor_(Marvel_Comics) based on the context. O indicates that nothing is predicted for that particular token.

Approach

- The entitiy classification layer is denoted by $E \in \mathbb{R}^{|KB|\times d}$ [1] where |KB| denotes the number of entities in KB d denotes the token‘s embedding size.
- The probability of entity link for each entry in the entire vocabulary is given by
- $p(j\,|\,v,h)$ where word v is the i-th token in context h. The probability is calculated by $\sigma(E_j c_i)$

1 Samuel Broscheit.2020. Investigating Entity Knowledge in BERT with Simple Neural End-To-End Entity Linking.

Approach

For better entity disambiguation:

- A larger context that spans multiple sentences are preferred.
- Text fragments which have less annotated Wikipedia links are chosen.
- Trie-based matcher is used for annotating all occurrences of entities' mentions that are collected as linkable strings.
- (m, e) tuples of entities e and their mentions m are collected.
- Mentions of less frequent entities have a non-zero probability to link to nothing
- Average of the probability of linking to Nil is calculated as follows:

$$\bar{p}_{Nil} = \frac{1}{k}\sum_{j}\frac{\#(m_j, Nil)}{\#m_i}$$ [1]

1 Samuel Broscheit.2020. Investigating Entity Knowledge in BERT with Simple Neural End-To-End Entity Linking.

UNIVERSITÄT PADERBORN

Experiments and Results

- To investigate if BERT+Entity model learns something additional on top of BERT.

Setting 1	Setting 2
Wikipedia	CoNLL03/AIDA
700K frequent entities	500k frequent entities
Fragment size of 110 tokens	Fragment size of 250 tokens
3 frequent , 1 infrequent linked entities	1 linked entity
8.8M training instances	2.4M training instances

- Setting 1: For initial study
- Setting 2: Follow-up study to improve entity linking performance

Experiments and Results

- The steep increase at the 4th epoch happens because of switching the model from Frozen-BERT+Entity to BERT+Entity for training.

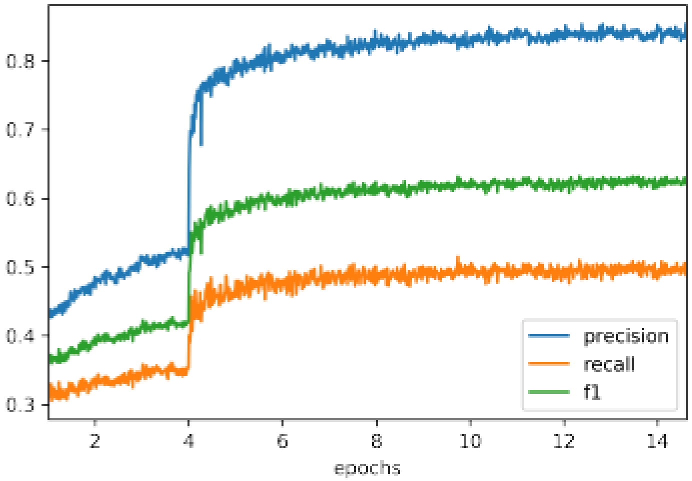

Figure 4: InKB scores on validation data in setting 2[1]

1 Samuel Broscheit.2020. Investigating Entity Knowledge in BERT with Simple Neural End-To-End Entity Linking.

UNIVERSITÄT PADERBORN

Experiments and Results

Results:

- Apparently, only the entity classifier in Frozen-BERT+Entity is trained.
- BERT+Entity learns more entity knowledge than frozen-BERT+ Entity and BERT

		AIDA/testa			AIDA/testb		
		strong F1	weak F1	ED	strong F1	weak F1	ED
Kolitsas et al. (2018) indep. baseline		80.3	80.5	-	74.6	75.0	-
Kolitsas et al. (2018)		89.4	89.8	93.7	82.4	82.8	87.3
BERT		63.3	66.6	67.6	49.6	52.4	52.8
Setting I	Frozen-BERT+Entity	76.8	79.6	80.6	64.7	68.0	68.6
	BERT+Entity	82.8	84.4	86.6	74.8	76.5	78.8
Setting II	Frozen-BERT+Entity	76.5	80.1	79.6	67.8	71.9	67.8
	BERT+Entity	86.0	87.3	92.3	79.3	81.1	87.9

Figure 5: comparison of results across different models[1]

- Scores of the models change based on the datasets that are used for training

1 Samuel Broscheit.2020. Investigating Entity Knowledge in BERT with Simple Neural End-To-End Entity Linking.

Discussion

Pros:

- Errors that occur due to interdependencies between MD, CD and ED can be avoided.
- The results of BERT+Entity comes very close to that of state-of-the-art model.
- The performance of BERT+Entity shows an increase of 23%-25% over BERT.

Cons:

- BERT+Entity predicts Nil to lot of entities instead of linking to something that is related.
- The performance of all the models drops from AIDA/testa to AIDA/testb due to model overfitting on validation data.

UNIVERSITÄT PADERBORN

Conclusion

- Performance improvement in setting 2 of data can be seen due to maximum fragments per entity.[1]
- Hardware specification of the model can be enhanced to tackle the challenges that this model face with respect to current state of the art.[2]
- First model that doesn't undergo any entity linking steps for learning.

1 Samuel Broscheit.2020. Investigating Entity Knowledge in BERT with Simple Neural End-To-End Entity Linking.

2 Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann. 2018. End-to-end neural entity linking. In Proceedings of the 22nd Conference on Computational Natural Language Learning,pages 519–529, Brussels, Belgium. Association forComputational Linguistics.

Thank you