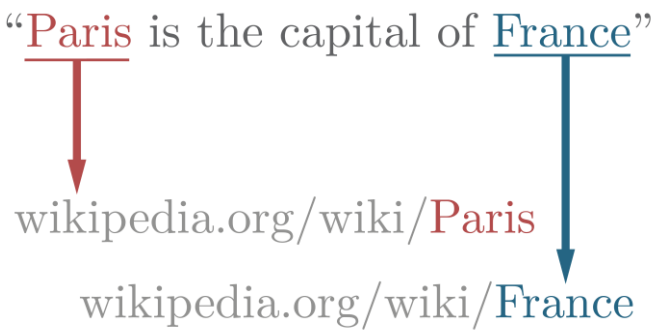
Investigating Entity Knowledge in BERT with Simple Neural End-to-End Entity Linking

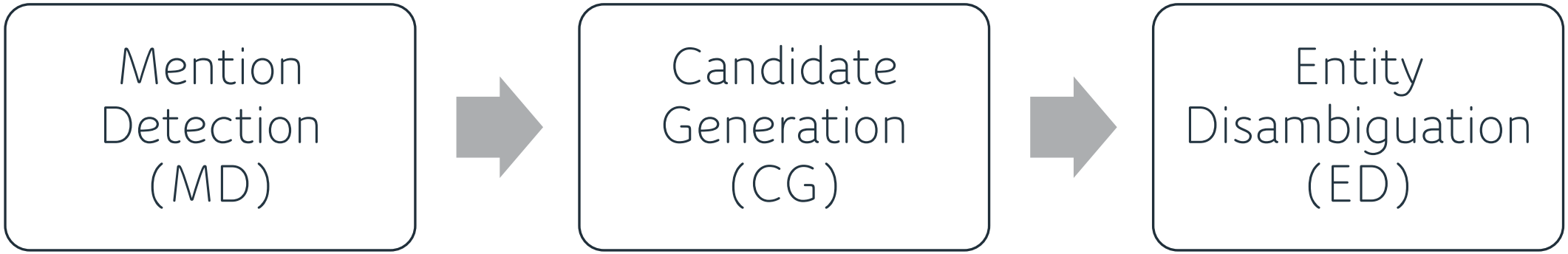
SAMUEL BROSCHEIT, CoNLL 2019

Mariana Leite
me.leite@campus.fct.unl.pt
NOVA Search Reading Group

Entitity Linking?



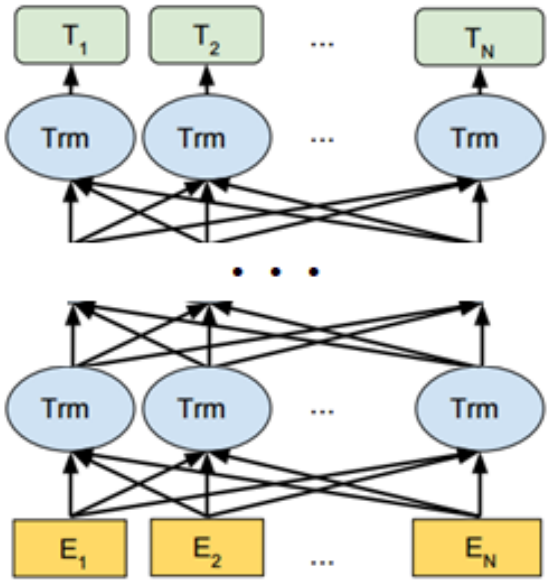
- **Goal** - given a knowledge base (KB) and unstructured data, e.g. text, to **detect** mentions of the KB's entities in the data and **link** them to the correct KB entry.



BERT? (Bidirectional Encoder Representations from Transformers [1])

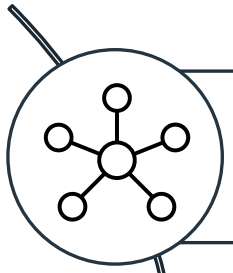
- Deep self attention-based architecture pretrained on large amounts of data.
- Provides very rich linguistic text-representations.

  (very useful for many NLP tasks!)
- BERT is being analyzed and applied in various domains [2][3].

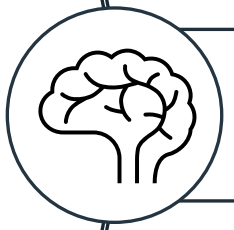


[1] Devlin et al., 2018 ; [2] Beltagy et al., 2019 ; [3] Lee et al., 2019.

The goal in this study...



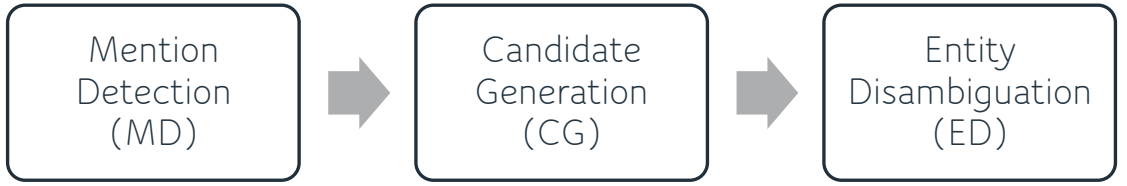
Can BERT's architecture learn all entity linking steps jointly?



How much entity knowledge is already contained in BERT?

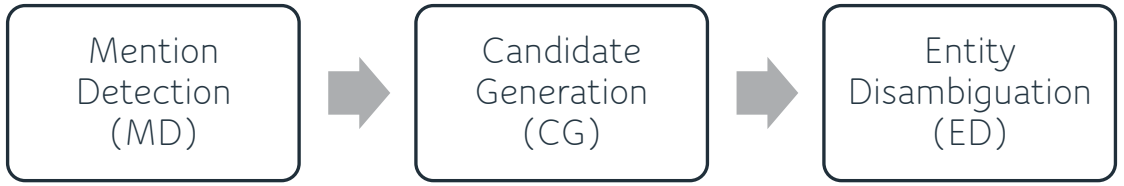
Does additional entity knowledge improve BERT's performance in downstream tasks?

EL – Related Work



- **Durrett and Klein** (2014) - model interactions between the MD, CG and ED tasks jointly.
- **Nguyen et al.** (2016) - jointly modelling MD and ED with a graphical model and show that it improves ED performance and is more robust.
- **Kolitsas et al.** (2018) - proposed the first neural model to learn MD and ED jointly.

EL – Related Work



- **Durrett and Klein** (2014) - model interactions between the MD, CG and ED tasks jointly.
- **Nguyen et al.** (2016) - jointly modelling MD and ED with a graphical model and show that it improves ED performance and is more robust.
- **Kolitsas et al.** (2018) - proposed the first neural model to learn MD and ED jointly.

Only yield entity links!

- **Yamada et al.** (2017) – 1st to investigate Neural Text Representations + EL (**limited to ED**)

BERT+Entity Model

- Straightforward extension on top of BERT.
  - ➢ BERT is initialized with the publicly available weights;
  - ➢ Added an output classification layer on top of the architecture.

d - BERT's token embedding size;

$|\mathrm{KB}|$ - number of entities in the KB;

$E \in \mathbb{R}^{|KB| \times d}$ - entity classification layer.

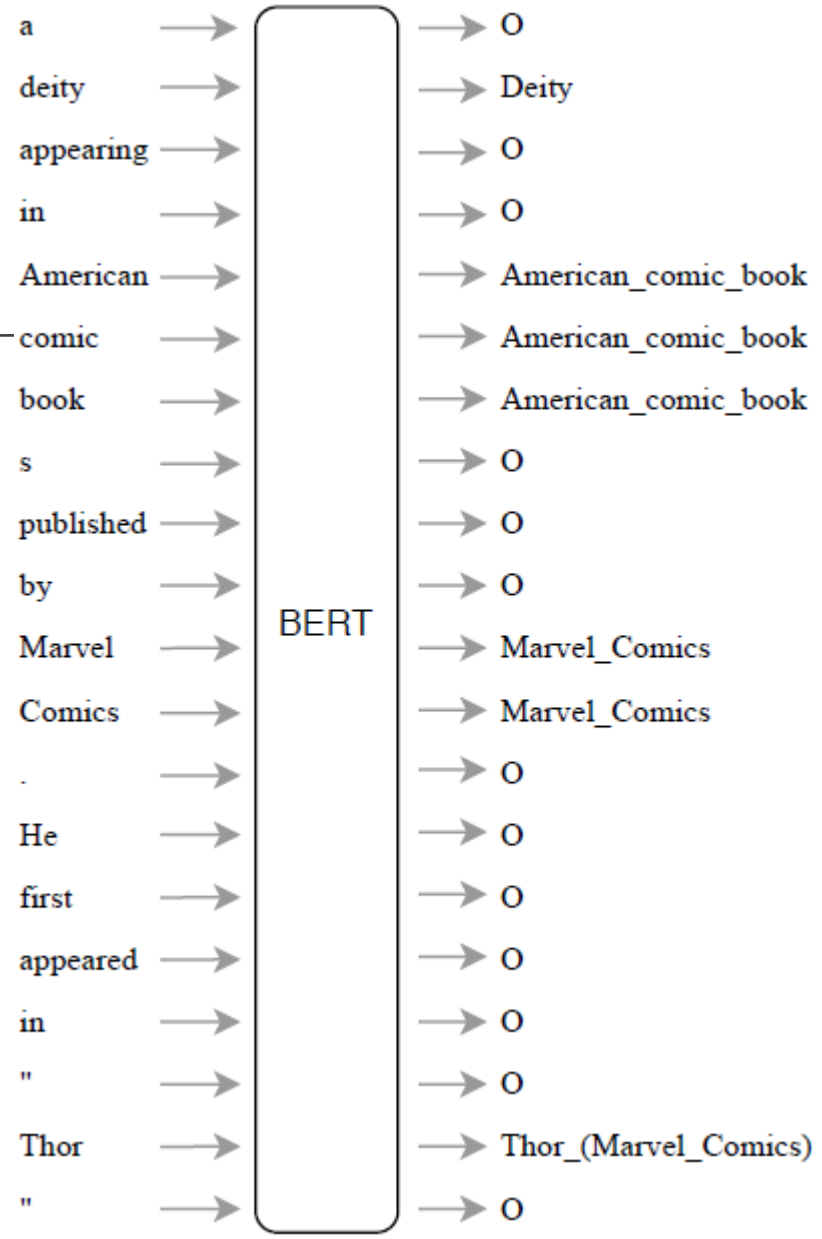
BERT+Entity Model

V - sub-word vocabulary

$c_i = \text{BERT}(h)[i]$ is the:

- i-th contextualized token computed by BERT
- From context $h = [v_1, v_2, \ldots, v_{i-1}, v_i, v_{i+1}, \ldots, v_m],\ v \in V$

The probability $p(j|v,h)$ of word v linking to entity j is computed by $\sigma(E_j c_i)$.



Training Data

- Derived from English Wikipedia texts.
  - ➢ Extracted expanded text spans that are associated with an internal Wikipedia link to use as annotation.
- A new challenge arises!
  - ➢ Most entities do not have all their mentions annotated.
- So:
  - ➢ Only select text fragments that contain a minimum count of annotated Wikipedia links.
  - ➢ Annotate all occurrences of linkable strings that were collected.

Experiments

SETTING I

- **700K** topmost frequent entities
- **Missing 30** entities
- Text fragments of **110** tokens
- Keep fragments that contain at least 1 infrequent linked E or 3 frequent
- **8.8M** training instances

SETTING II

- **500K** topmost frequent entities
- **Added 1000** entities
- Text fragments of **250** tokens
- Keep fragments that contain at least 1 linked E + at most 500 fragments per entity
- **2.2M** training instances

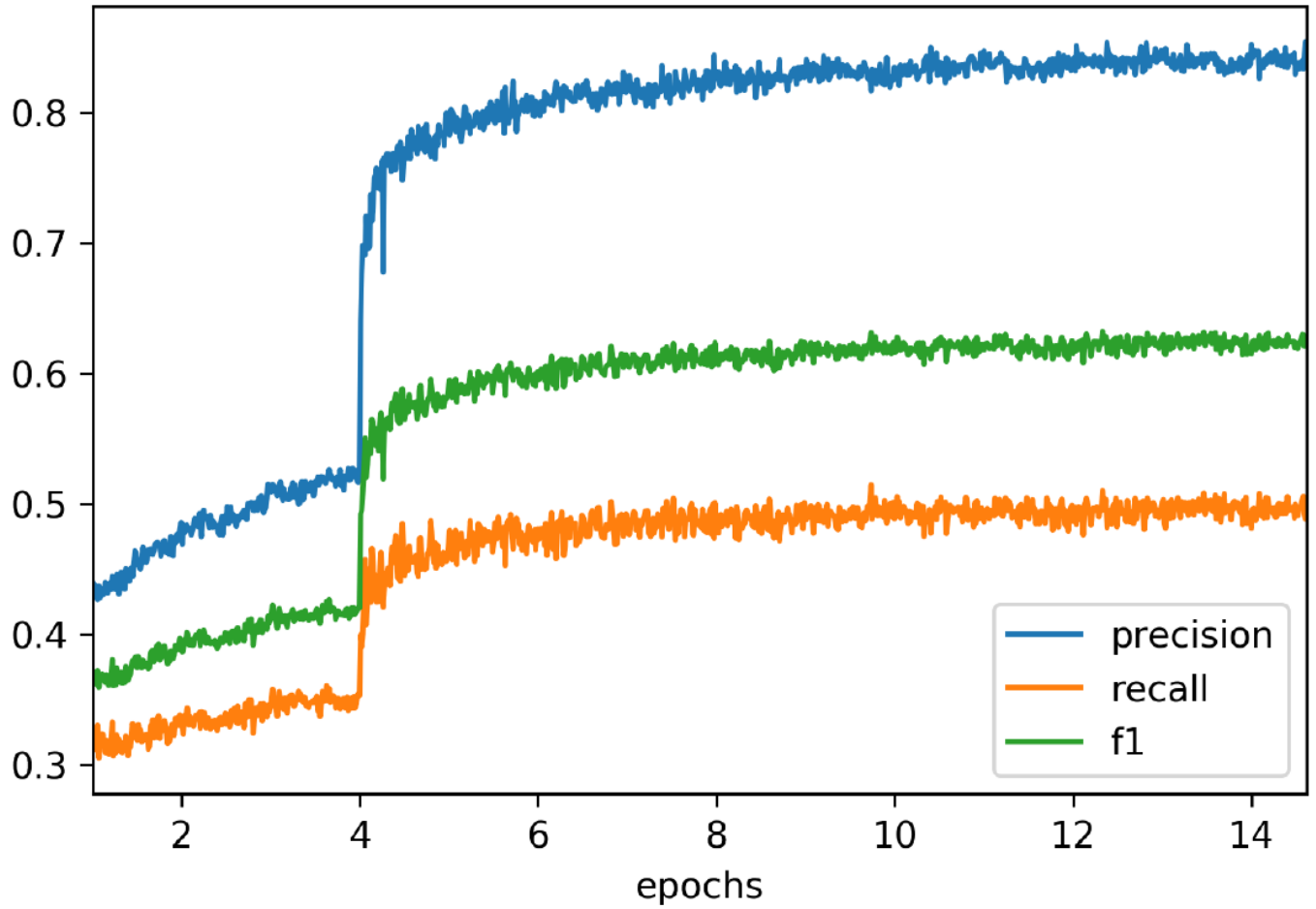
Training

- Multi-class classification over the entity vocabulary:
  - $y_{ij} = p(j|v_i)$ for $j \in \{1, \ldots, \|KB\|\}$

- Computing the loss over the whole entity vocabulary would be infeasible, because this is very large and the gradients for the entity classifier would exceed our GPU memory
  - Negative Sampling

Training

Per token classification InKB scores on the validation data during training on the Wikipedia dataset in Setting II for 40 days.



Results

		AIDA/testa			AIDA/testb		
		strong F1	weak F1	ED	strong F1	weak F1	ED
Kolitsas et al. (2018) indep. baseline		75.7	76.0	-	73.3	73.9	-
Kolitsas et al. (2018)		86.6	87.2	92.4	82.6	83.2	89.1
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

Investigating Errors

Investigating the types of strong precision errors of BERT+Entity trained in Setting I on CoNLL03/AIDA (testa) on 100 randomly sampled strong precision errors from the validation dataset.

Reason for error	#
no prediction	57
different than gold annotation	
no obvious reason	13
semantic close	4
lexical overlap	5
nested entity	5
gold annotation wrong	12
span error	3
unclear	1
	100

Downstream Tasks Experiments Results

Task	Metric	BERT-BERT-Ensemble	BERT+Entity-Ensemble
CoLA	Matthew's corr.	59.92	59.97
SST-2	accuracy	92.73	92.43
MRPC	F1/accuracy	89.16	90.13
STS-B	Pearson/Spearman corr.	89.90	89.60
QQP	accuracy	91.64	91.21
MNLI	matched acc./mismatched acc.	84.96	84.78
QNLI	accuracy	91.21	91.15
RTE	accuracy	71.48	73.64
WNLI	accuracy	56.33	56.33
SQUAD V2	matched/mismatched	76.89/73.83	76.36/73.46
SWAG	accurracy	80.70	80.76
WMT14 EN-DE	BLEU	22.51	22.20

Experiments on downstream tasks with BERT+Entity trained in Setting I.

Downstream Tasks Experiments Results

Additional entity knowledge is not helpful! Why?

- Entity overlap in training and testing such that also an entity unaware model can learn the necessary model?
- The entities are too scarce in the training data to make a difference?
- The tasks themselves do not require entity knowledge?

Conclusions

- This simplified approach worked surprisingly well!
  - Gap can be closed with larger hardware capacity.

- First model to perform EL without pipeline/heuristics.

- We can learn additional entity knowledge in BERT that helps in EL.

- Almost none of the downstream tasks really required entity knowledge.

(open question for future research!)

Q&A Time!