ReFinED

ReFinED: An Efficient Zero-shot-capable Approach to End-to-End Entity Linking

Tom Ayoola, Shubhi Tyagi, Joseph Fisher, Christos Christodoulopoulos, Andrea Pierleoni
Amazon Alexa AI

Summary by Omar Khattab

Nov 2023

Entity Linking

“Barack Obama, born in Hawaii, served as the 44th President of the United States. He graduated from Harvard Law School and won the Nobel Peace Prize in 2009.”

Entity Linking

“**Barack Obama**, born in **Hawaii**, served as the 44th President of the **United States**.

He graduated from **Harvard Law School** and won the **Nobel Peace Prize** in 2009.”

Entity Linking

“**Barack Obama**, born in **Hawaii**, served as the 44th President of the **United States**.

He graduated from **Harvard Law School** and won the **Nobel Peace Prize** in 2009.”

- **Mention Detection.** Recognizing mentions of entities in text.
- **Entity Disambiguation.** Linking each mention to its entry in a knowledge base, like Wikidata.

3.1 Task Formulation

Given a KB[5] with a set of entities $E = \{e_1, e_2, \ldots, e_{|E|}\}$, let $X = [x_1, x_2, \ldots, x_{|X|}]$ be a sequence of tokens in the document, and $M = \{m_1, m_2, ...m_{|M|}\}$ be a set of entity mentions. The goal of ED is to create a function $\mathcal{M} : M \rightarrow E$ which assigns each mention the correct entity label. In EL, both the mention spans and entity labels need to be predicted. We only consider mentions with a valid gold entity in the KB during evaluation.

Entity Linking Applications

- Question Answering
- Relation Extraction
- Automated Construction of Knowledge Bases

Traditional Parametric Approach: **Learning purely from text!**

Broscheit. 2019. Investigating Entity Knowledge in BERT with Simple Neural End-To-End Entity Linking.

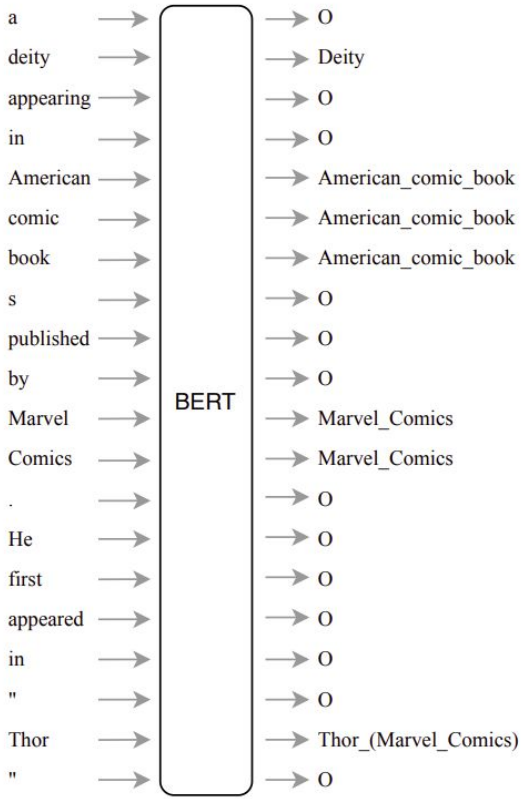
We can train a **token-level classifier**.

Encodes the sequence with **BERT**, and applies a **linear layer** on top to predict one of the set of entities or “O” (i.e., not applicable).

✔ **Efficient** and simple architecture.

✘ No semantic grounding => **poor generalization across contexts**!

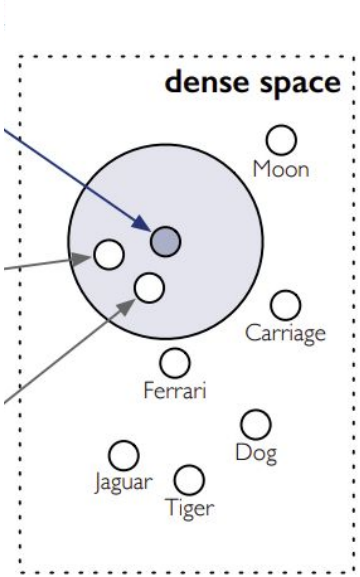
✘ By extension of this, incapable of generalizing to **unseen entities**.



Advanced Decomposition Approach: **Learning from descriptions!**

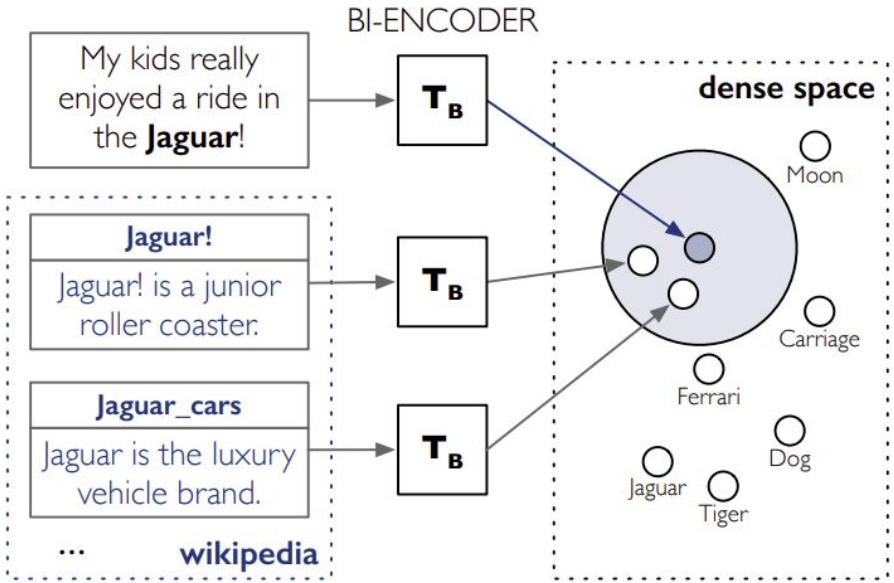
Wu et al. 2020. Scalable Zero-shot Entity Linking with Dense Entity Retrieval.

To insert some degree of semantic grounding, we ought to use more information about each entity, like its name and description.



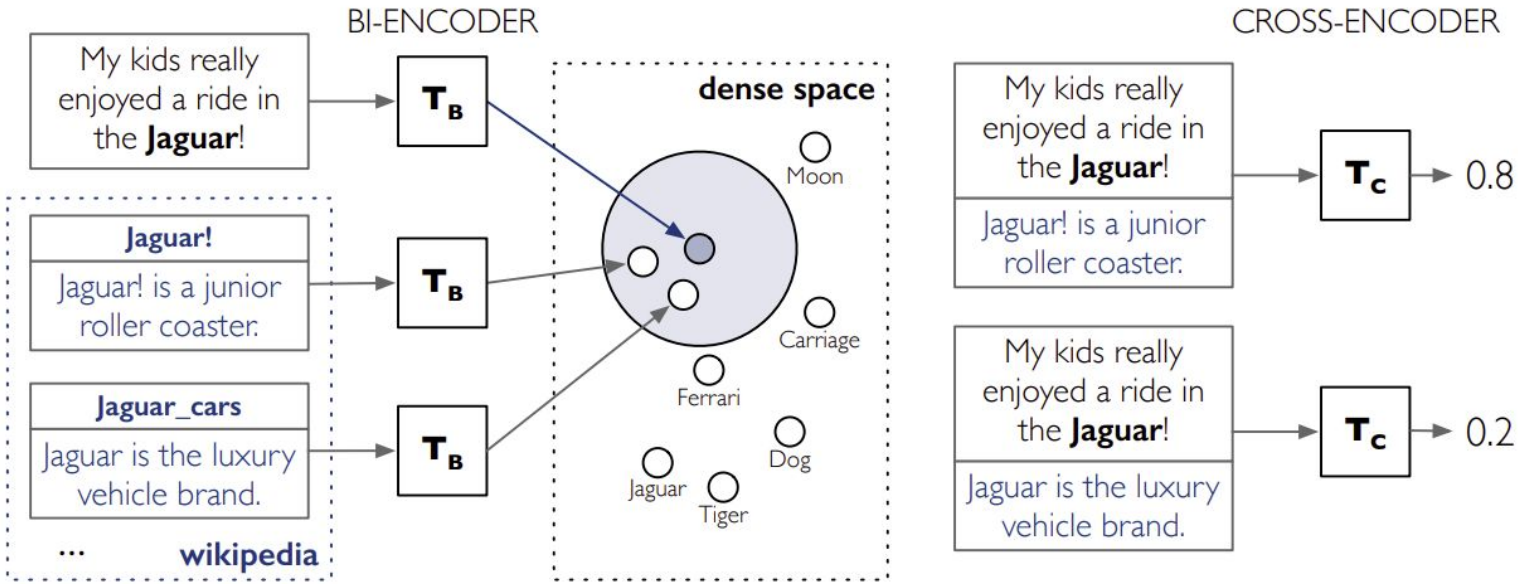
Advanced Decomposition Approach: **Learning from descriptions!**

To insert some degree of semantic grounding, we ought to use more information about each entity, like its name and description.



Advanced Decomposition Approach: **Learning from descriptions!**

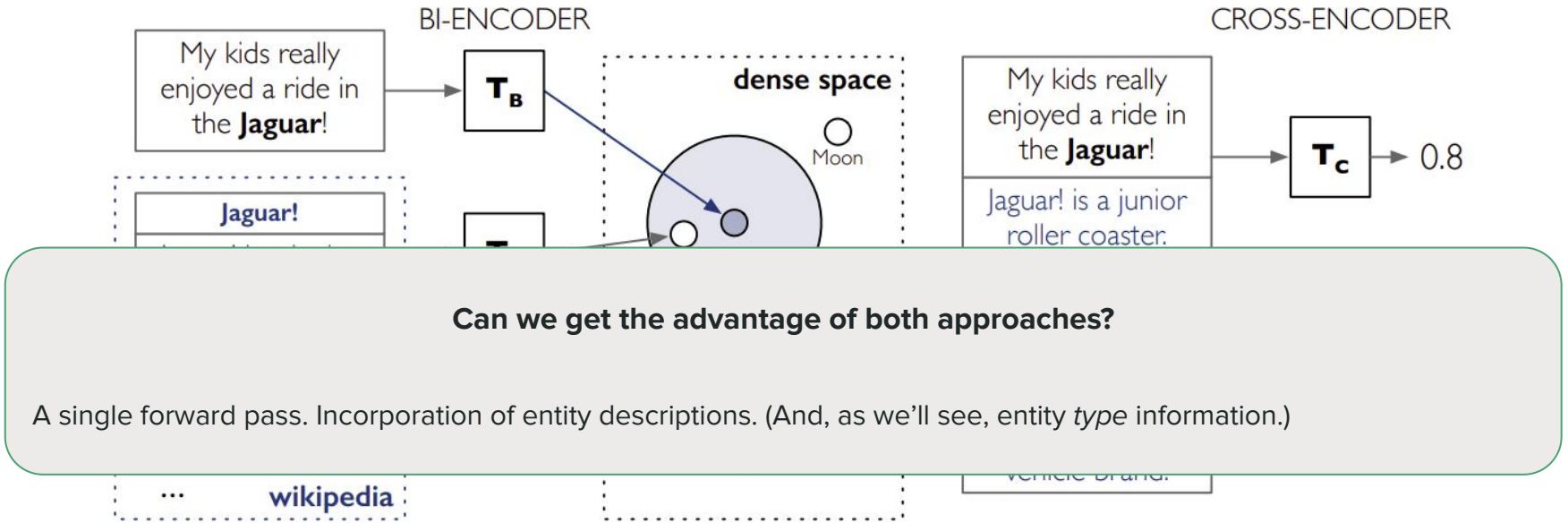
To insert some degree of semantic grounding, we ought to use more information about each entity, like its name and description.



Advanced Decomposition Approach: **Learning from descriptions!**

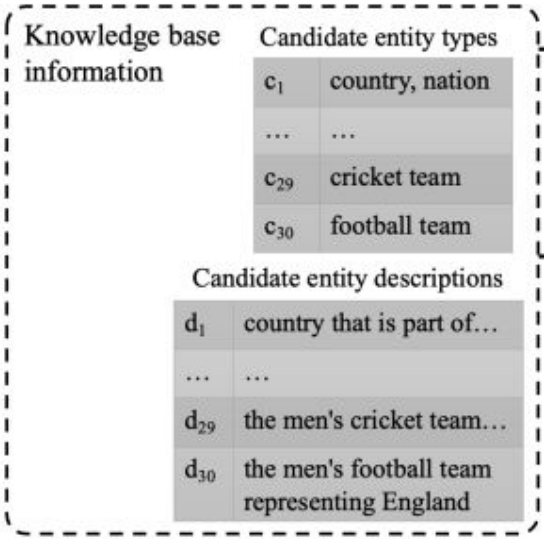
✔ Much better at generalizing to unseen and infrequent entities.

✘ Much more expensive: requires at least one forward pass per mention.

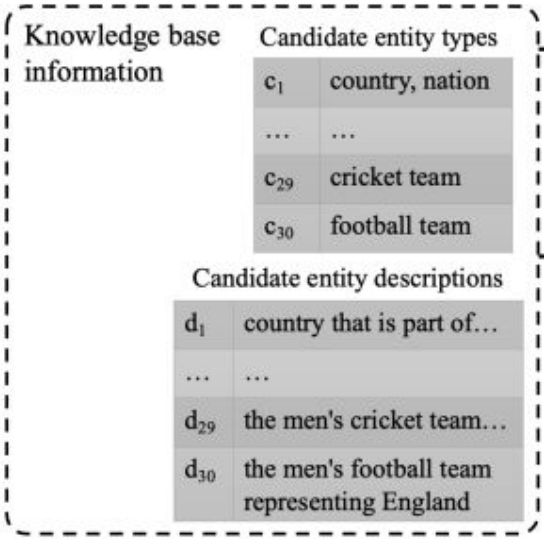


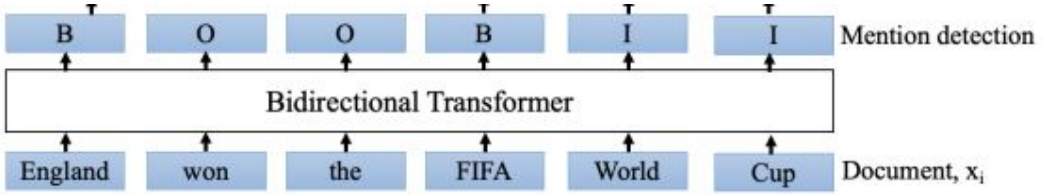
ReFinED: Representation and Fine-grained typing for ED

ReFinED: Representation and Fine-grained typing for ED



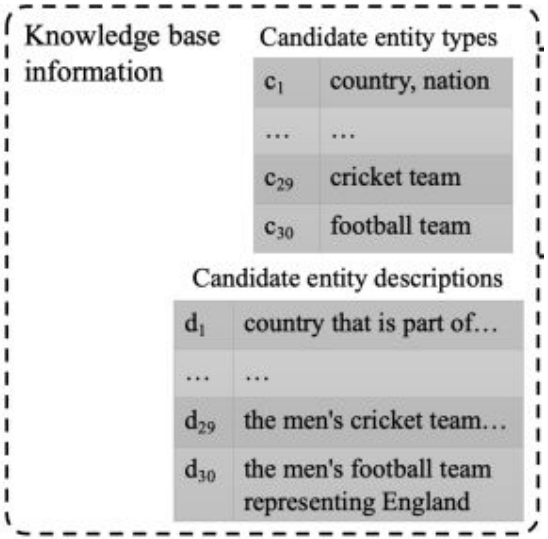
ReFinED: Representation and Fine-grained typing for ED

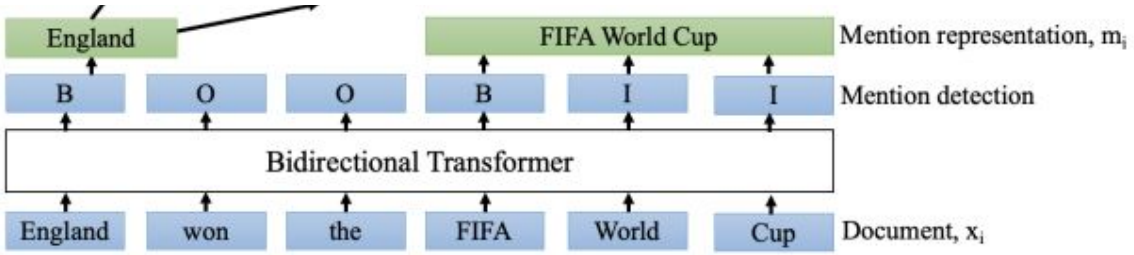




Mention Detection with Begin/Inside/Outside (BIO) Tagging.

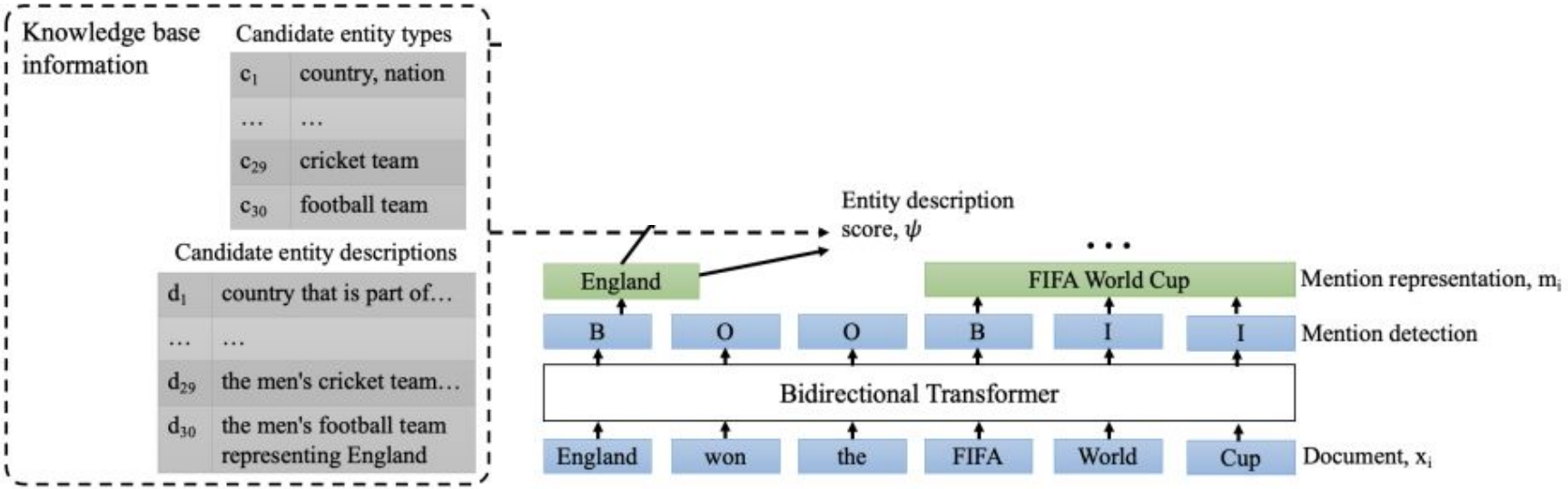
ReFinED: Representation and Fine-grained typing for ED





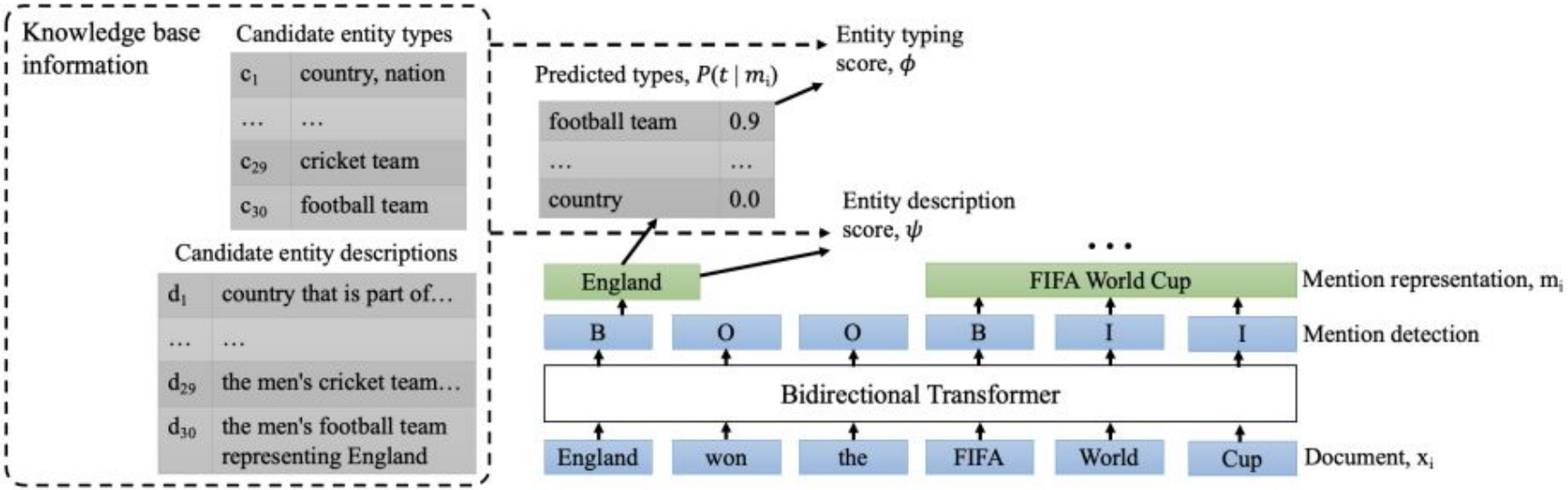
Mention Representation via mean pooling.

ReFinED: Representation and Fine-grained typing for ED



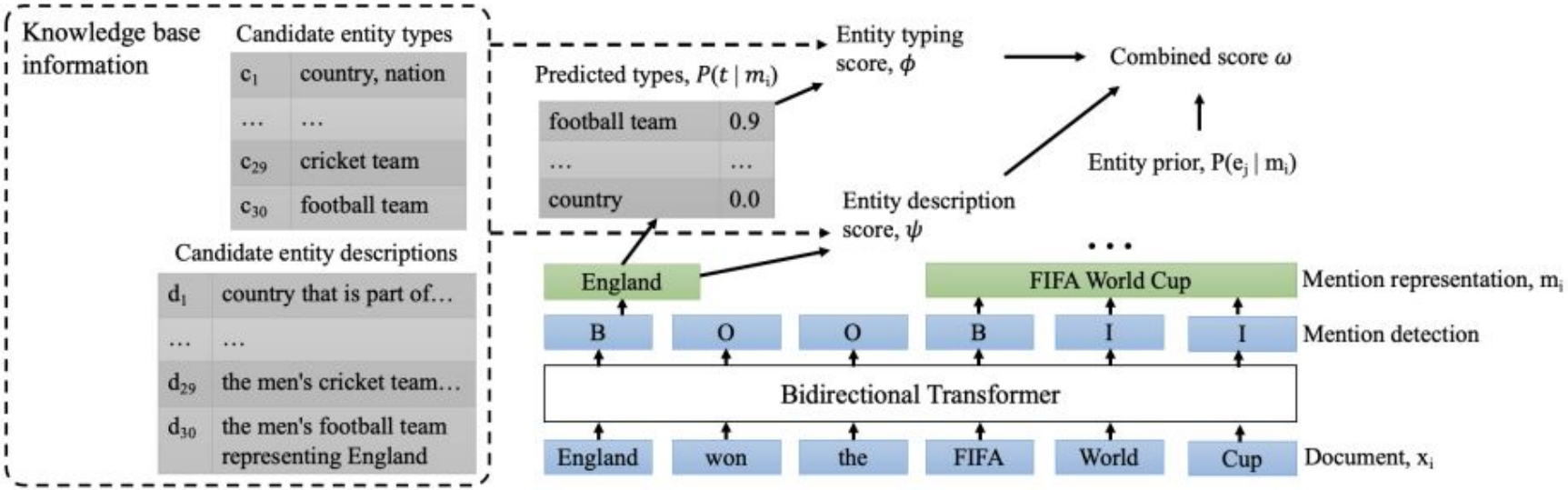
Entity Description score: multiple simultaneous bi-encoder representations

ReFinED: Representation and Fine-grained typing for ED



Entity Typing score: L2 distance between type vectors

ReFinED: Representation and Fine-grained typing for ED



Linear Combination of {Prior, Types, Description} scores.

ReFinED: Recap!

- Mention Detection with Begin/Inside/Outside (BIO) Tagging.
- Mention Representation via mean pooling.
- Entity Typing score.
- Entity Description score.
- Linear Combination of {Prior, Types, Description}.

Entity Linking Results

Method	AIDA
Hoffart et al. (2011)	72.8
Kolitsas et al. (2018)	82.4
van Hulst et al. (2020)	80.5
Cao et al. (2020)	83.7
ReFinED (Wikipedia)	77.8
ReFinED (fine-tuned)	**84.0**

Entity Linking Results

Method	AIDA	MSNBC*	DER*	K50*
Hoffart et al. (2011)	72.8	65.1	32.6	55.4
Kolitsas et al. (2018)	82.4	72.4	34.1	35.2
van Hulst et al. (2020)	80.5	72.4	41.1	50.7
Cao et al. (2020)	83.7	**73.7**	**54.1**	60.7
ReFinED (Wikipedia)	77.8	70.0	49.0	**65.9**
ReFinED (fine-tuned)	**84.0**	71.8	50.7	64.7

Entity Linking Results

Method	AIDA	MSNBC*	DER*	K50*	R128*	R500*	OKE15*	OKE16*	Avg.
Hoffart et al. (2011)	72.8	65.1	32.6	55.4	46.4	**42.4**	63.1	0.0	47.2
Kolitsas et al. (2018)	82.4	72.4	34.1	35.2	50.3	38.2	61.9	52.7	53.4
van Hulst et al. (2020)	80.5	72.4	41.1	50.7	49.9	35.0	63.1	58.3	56.4
Cao et al. (2020)	83.7	**73.7**	**54.1**	60.7	46.7	40.3	56.0	50.0	58.2
ReFinED (Wikipedia)	77.8	70.0	49.0	**65.9**	52.6	40.1	**65.0**	**59.5**	60.0
ReFinED (fine-tuned)	**84.0**	71.8	50.7	64.7	**58.1**	42.0	64.4	59.1	**61.9**

ReFinED has been deployed by Amazon Alexa at “web scale”

- Populate a KB from a **billion web pages**, *multiple times per year*.
- Requires **2 days of processing**, using **500 T4 GPUs**.
- Pro: Uniform architecture is **easy to scale**!
- Pro: Generalizes well to **90M entities** at scale.

Method	Time taken (s)	Avg. ED F1
Cao et al. (2020)	2100	88.7
Wu et al. (2020) bi-encoder	93	80.4
Wu et al. (2020) cross-encoder	917	87.2
Orr et al. (2021)	438	77.6
ReFinED	**15**	**89.4**

Table 5: Time taken in seconds for EL inference on AIDA-CoNLL test dataset.