ellis unit STUTTGART

University of Stuttgart Germany

Anastasiia Iurshina
Contact: anastasiia.iurshina@ipvs.uni-stuttgard.de

Neural NIL-linking

Problem definition

Entity Linking (EL) is a task of matching an occurrence of a named entity in text (known as mention) with the corresponding entity in a knowledge base.

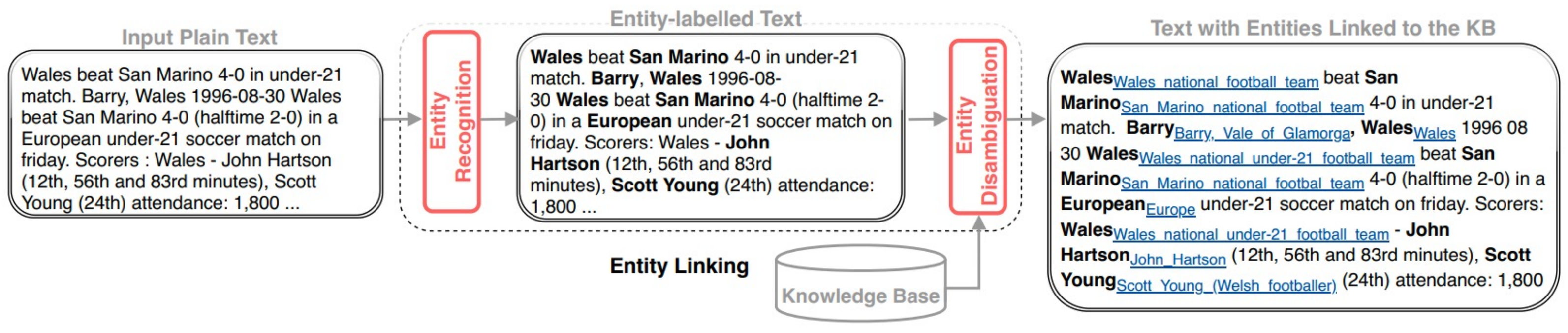


Figure 1. Entity Linking system [4]

In a real-world scenario, the given knowledge base is often incomplete, which leads to cases where *the mention in the text does not have a matching entity in the knowledge base.*

These mentions are called *NIL-mentions*, and the task of dealing with NIL-mentions is *NIL-linking*. In the NIL-linking task, we distinguish two sub-tasks: *NIL-detection* and *NIL-disambiguation*.

NIL-detection determines if a mention is an NIL-mention, and if its corresponding entity exists in the knowledge base. If not, NIL-disambiguation will distinguish between NIL-mentions by determining which of them refers to the same out-of-knowledge-base entity.

Related work

There are a lot of works that address the task of Entity Linking with neural approaches [2 3]. The core idea of many of them can be summarized as follows: two vector representations are built, one of the mentions from the text and the other of the entities in the knowledge base. Then the best match between these two representations is found.

However, even though NIL-detection and NIL-disambiguation tasks have been considered in the older works, they have not been addressed in recent Neural Entity Linking approaches and are often left for future work [1 2].

Contribution 1: NIL-linking dataset

To address the tasks of NIL-detection and NIL-disambiguation, we developed a new dataset, called NILK. It is constructed from WikiData and Wikipedia dumps from two different timestamps. The NILK dataset has two main features:
1) It marks NIL-mentions for NIL-detection by extracting mentions which belong to newly added entities in Wikipedia text.
2) It provides an entity label for NIL-disambiguation by marking NIL-mentions with WikiData IDs from the newer dump.

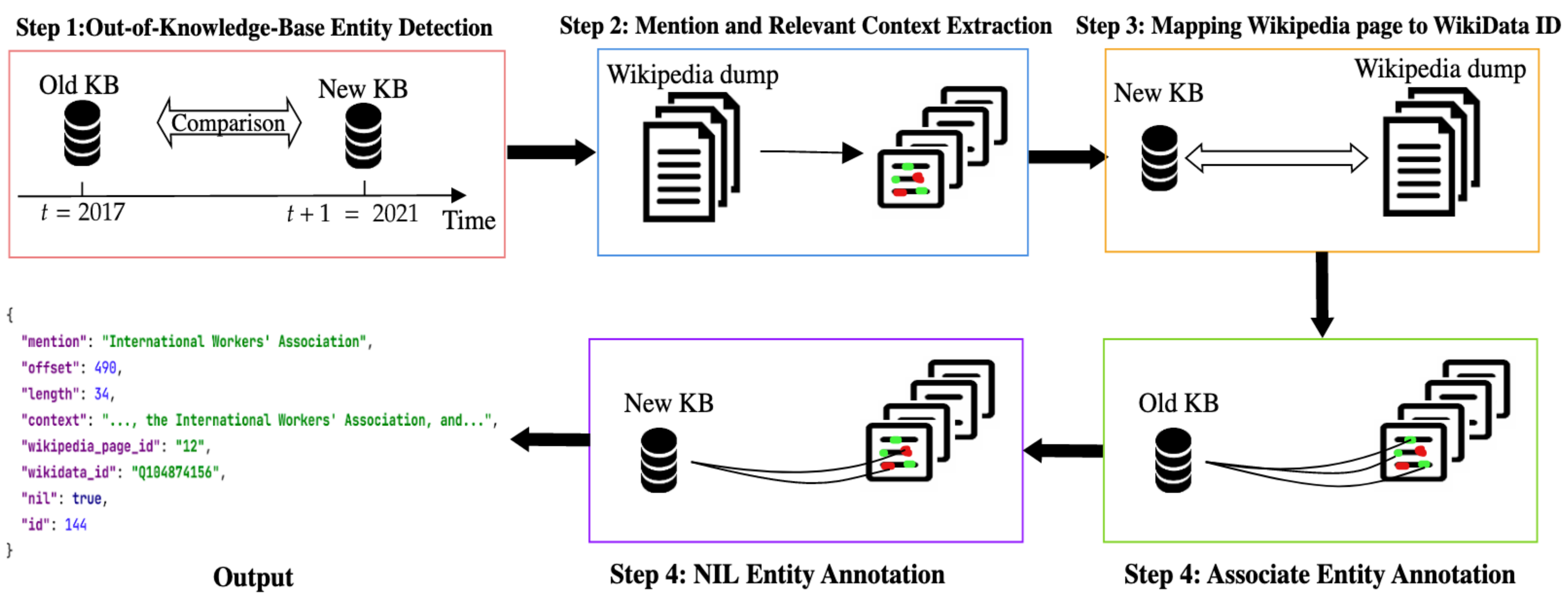


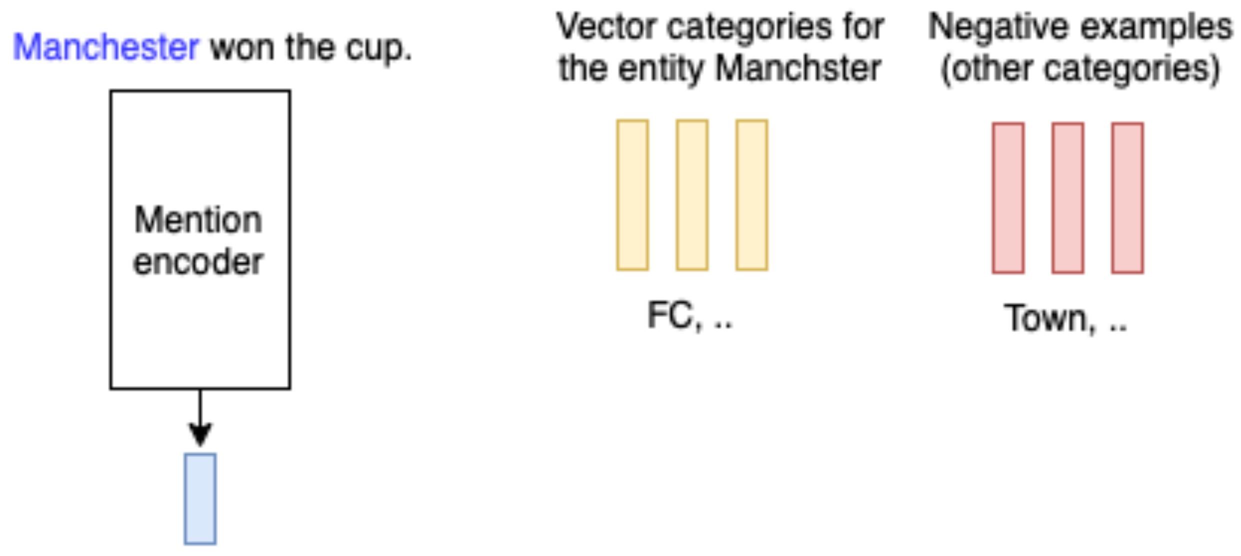
Figure 2. The pipeline of NILK dataset construction

Type	Count
Linked entities	4,228,124
Out-of-knowledge-base entities	352,765
Linked mentions	106,028,997
NIL-mentions	1,652,484

Table 1. NILK-dataset statistics

Contribution 2: Typing for Entity Linking and NIL-linking

Entity Typing is a powerful mechanism that can be helpful for both Entity Linking and NIL-linking tasks. Currently we are working on an model for mentions type prediction.



negative sampling objective

$$\log \sigma({v'_{w_O}}^\top v_{w_I}) + \sum_{i=1}^{k} \mathbb{E}_{w_i \sim P_n(w)} \left[\log \sigma(-{v'_{w_i}}^\top v_{w_I})\right]$$

Figure 3. Type prediction model

- Mention representation: vector from the encoder
- Entity representation: average of categories vectors
- Linking: find the closest entity vector to the mention
- BERT-based mention encoder
- Self-attention for merging mention tokens into one token

In addition to using this model for Entity Linking and NIL-linking, we want to investigate how the process of type prediction differs for linked and NIL-mentions.

Planned contribution 3: NIL-mention detection

To address the task of NIL-detection, we look into a simple threshold approach where we determine a similarity threshold below which a mention is considered to be NIL-mention.

However, it is unlikely that a single threshold will be sufficient. So we want to combine the threshold with the type prediction. For different types, we will have different thresholds ensuring a more reliable approach for NIL-detection.

Planned contribution 4: NIL-mention disambiguation

As a baseline, for NIL-disambiguation we will perform hierarchical clustering of the vector representations of detected NIL-mentions (with and without linked mentions). We will also try adding type information to those representations.

References

[1] Feng Hou, Ruili Wang, Jun He, and Yi Zhou. 2020. Improving entity linking through semantic reinforced entity embeddings. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics
[2] Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable zero-shot entity linking with dense entity retrieval. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).
[3] Daniel Gillick, Sayali Kulkarni, Larry Lansing, Alessandro Presta, Jason Baldridge, Eugene Ie, Diego Garcia-Olano 2019. Learning Dense Representations for Entity Retrieval, CoNLL
[4] Özge Sevgili, Artem Shelmanov, Mikhail Arkhipov, Alexander Panchenko, and Chris Biemann. 2022. Neural entity linking: a survey of models based on deep learning. Semantic Web

ANALYTIC COMPUTING