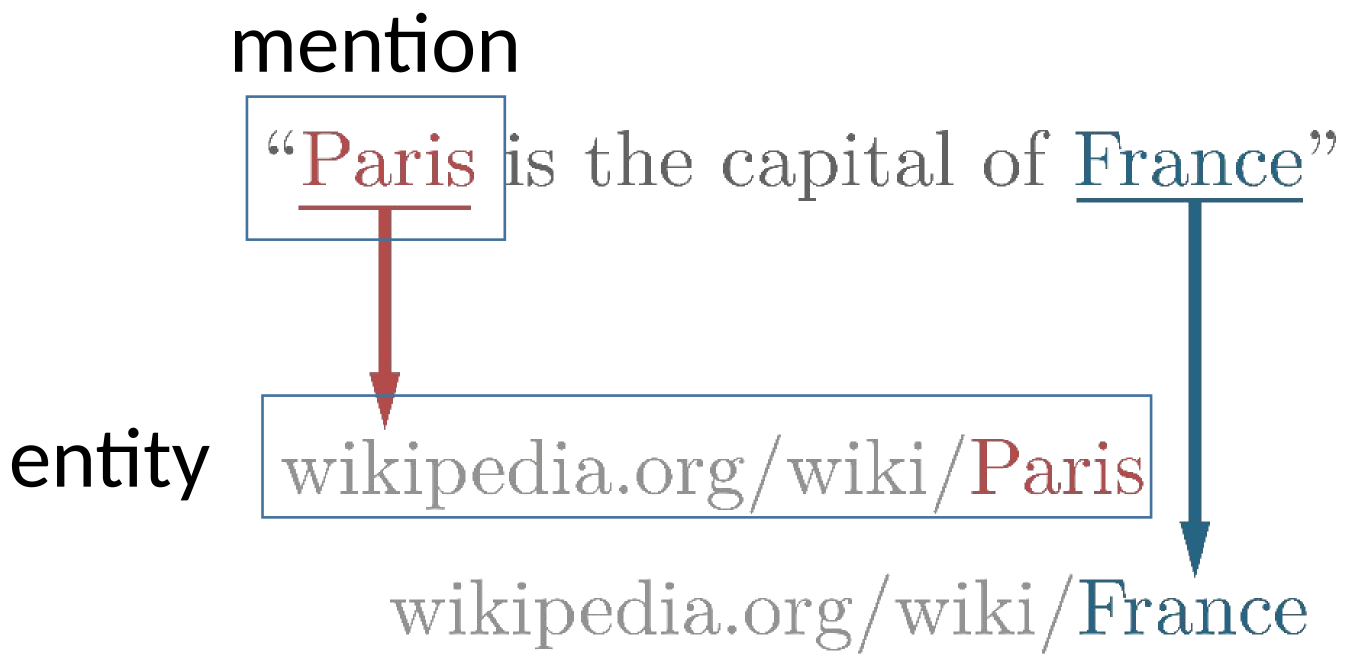
Entity Linking

Entity linking

- Assign a unique identity to entities mentioned in the text
    - Knowledge graph
    - Ontology
    - Dictionary
- Aka
    - named-entity disambiguation
    - named-entity normalization
    - Concept Recognition



Related tasks

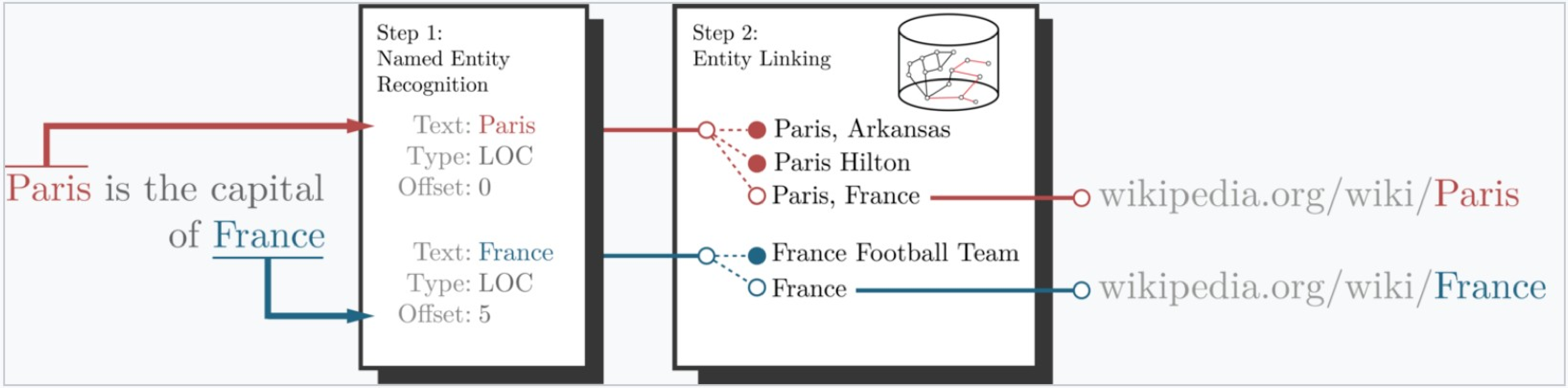
- Co-reference resolution
  - understands whether multiple words in a text refer to the same entity

Paris is the capital of France. It is also the largest city in France.

- Entity resolution
  - Record linkage, data matching, data linkage
  - Database

Data set	Name	Date of birth	City of residence
Data set 1	William J. Smith	1/2/73	Berkeley, California
Data set 2	Smith, W. J.	1973.1.2	Berkeley, CA
Data set 3	Bill Smith	Jan 2, 1973	Berkeley, Calif.

Typical steps



Typical entity linking steps. 1. NER - find named entities in the text (here, Paris and France); 2. link found named named entities to corresponding unique identifiers (here, Wikipedia pages). (2) is often done by: 2.1. defining a metric for comparing candidates in the system; 2.2. creating a small set of candidate identifiers for each named entity, and 2.3. Scoring candidates with the metric and choosing one that has the highest score.

https://en.wikipedia.org/wiki/Entity_linking

Typical Modules

- Candidate entity generation
  - Filter out irrelevant entities
  - Retrieve a candidate set
  - Requirement: fast
- Candidate entity ranking (Disambiguation)
  - rank the candidate entities
- Unlinkable mention prediction (NIL)
  - whether the top-ranked entity identified in the Candidate Entity Ranking module is the target entity for mention m.

Applications

- Information Extraction
- Information Retrieval
  - Semantic entity-based search
  - Query disambiguation
- Content Analysis
  - recommendation systems
- Question Answering
- Knowledge Base Population

Heuristic Based Methods (Candidate entity generation)

- acronym that is in parenthesis adjacent to the expansion
  - Iowa State University (ISU)
- identify named entities from the document and if some identified named entity contains the entity mention as a substring
  - Michael I. Jordan = Jordan
- Name Dictionary (Wikipedia)
  - Entity page
  - Redirect pages
  - Disambiguation pages
  - Bold phrases from the first paragraphs (nick names, alias, etc)
  - Hyperlinks in Wikipedia articles

Disambiguation

- Features

Disambiguation

- Features
  - Similarity
    - Lexicon (name string comparison)
    - semantic
  - Anchors statistic (Mention-entity prior)
  - Entity type
  - Context
  - Entity description
  - KG graph

Supervised Ranking Methods

- Binary classification

- Learning to rank
  - Pointwise (regression problem)
    - : predicting the real-value or ordinal score of x
    -
  - **Pairwise**
    - : classification problem for a given pair
    - Usually implemented with a scoring function:

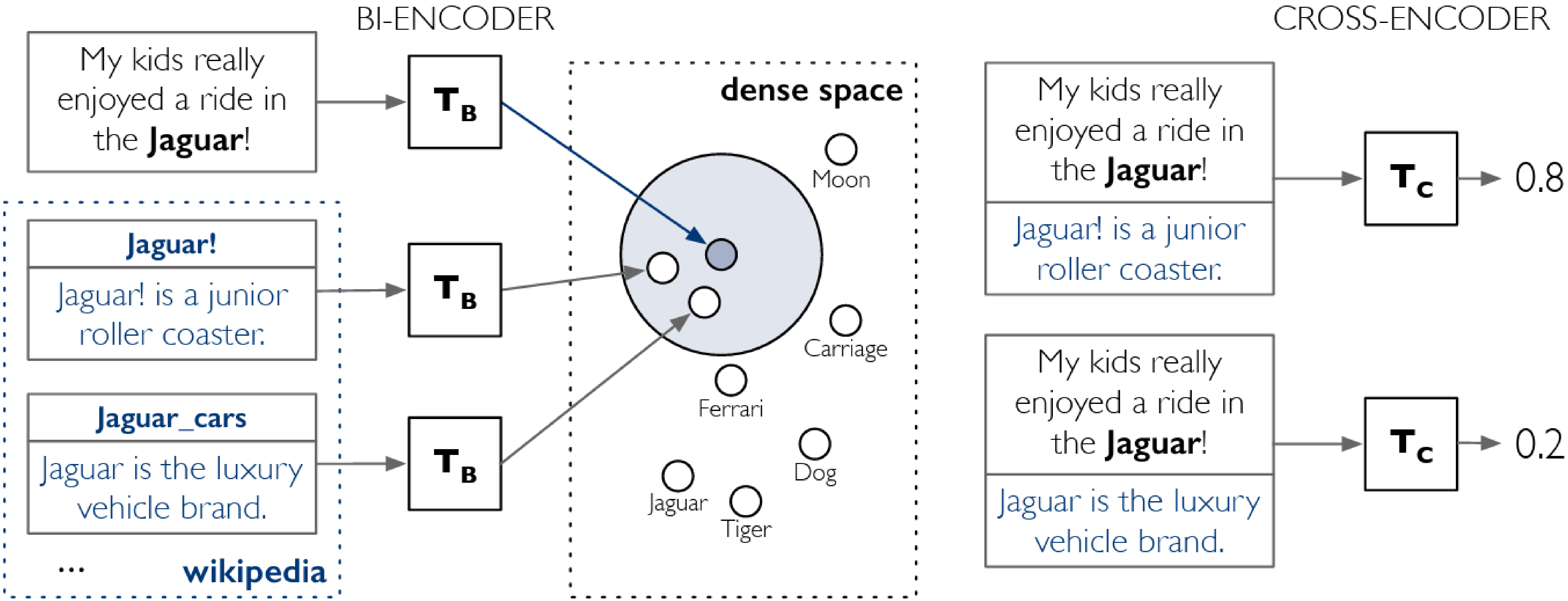
  - Listwise

EL Settings

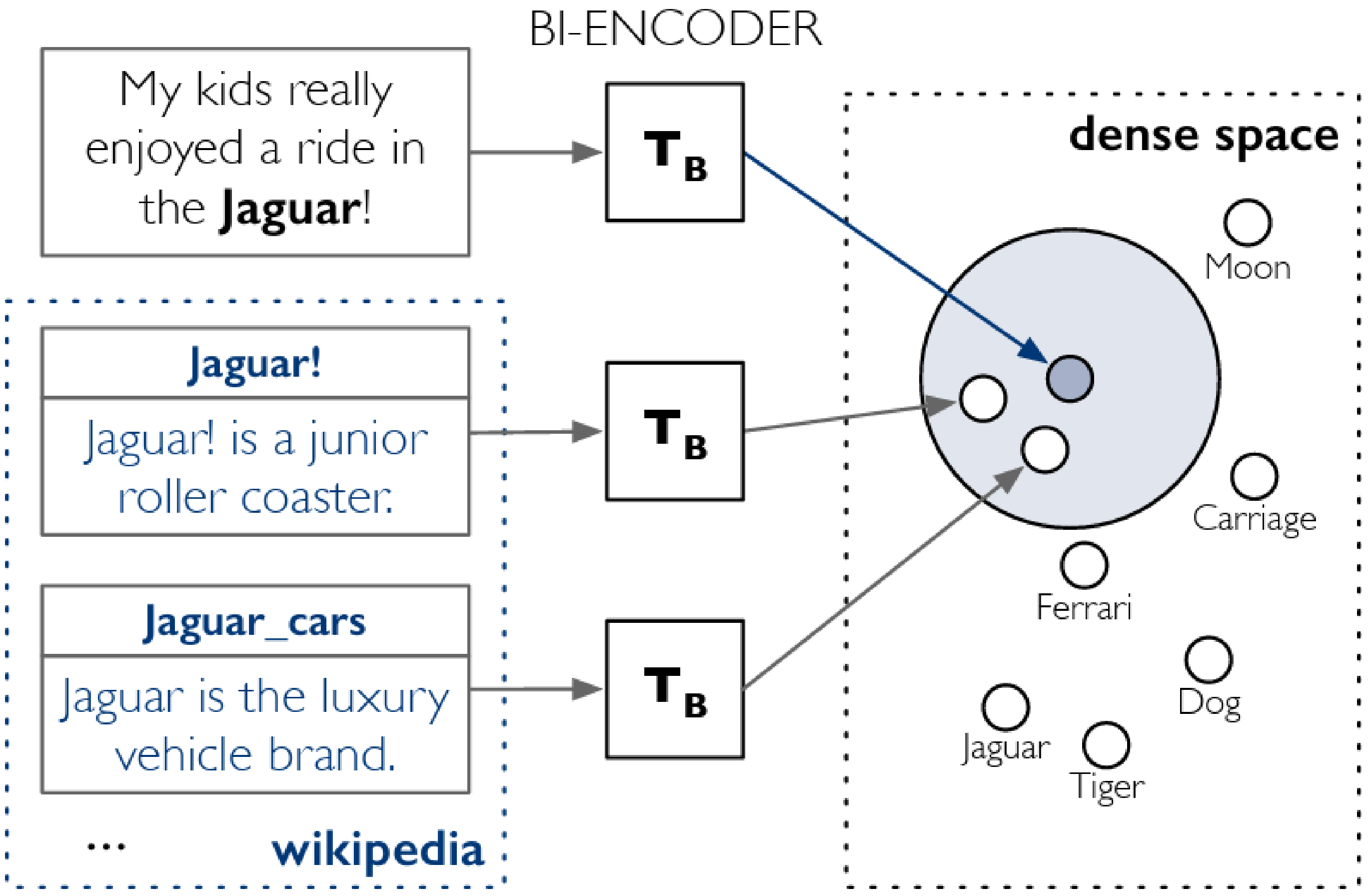
- Zero-shot
  - Train in Domain A
  - Inference in Domain B
  - Unseen mentions & entities
- End-to-End
  - Combine NER with EL

BLINK (by Facebook 2020)

- Retrieve and rerank



BLINK



- Mention Encoder

Input

Context [B] Mention [E] Context

Output

Last layer

- Entity Encoder

Input

Entity Title [X] Description

Output

Last layer

Optimize dot-product of the two output

BLINK

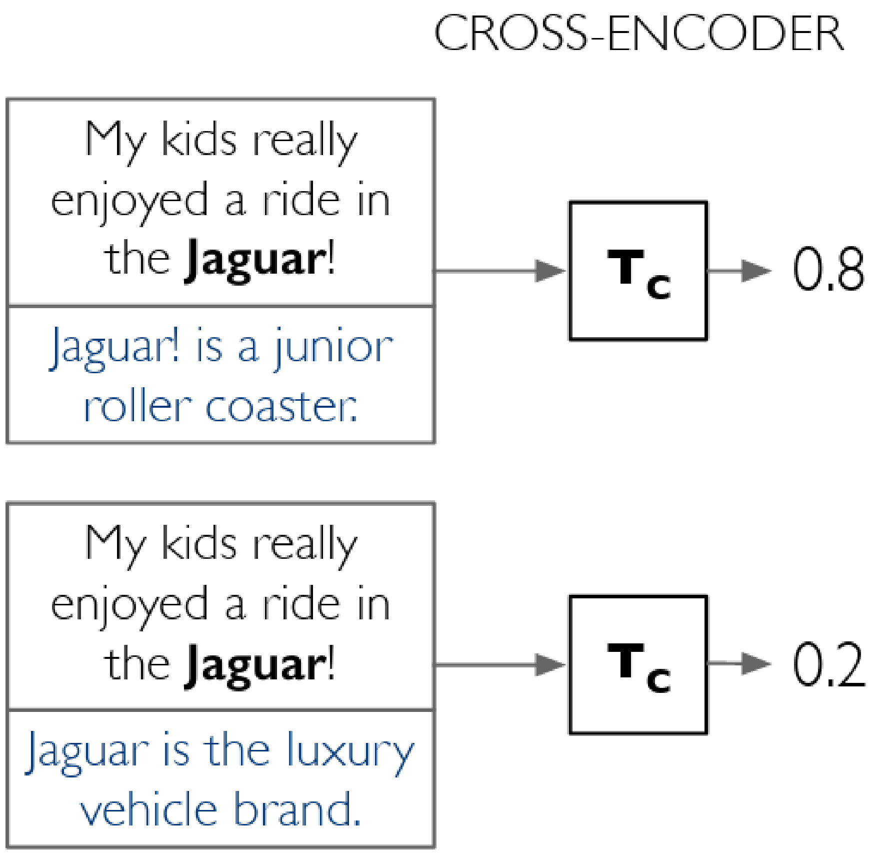
- Cross Encoder

Input

Context [B] Mention [E] Context [S] Entity Title [X] Description

Output

Probability distribution among entity choices



BiEncoder VS CrossEncoder

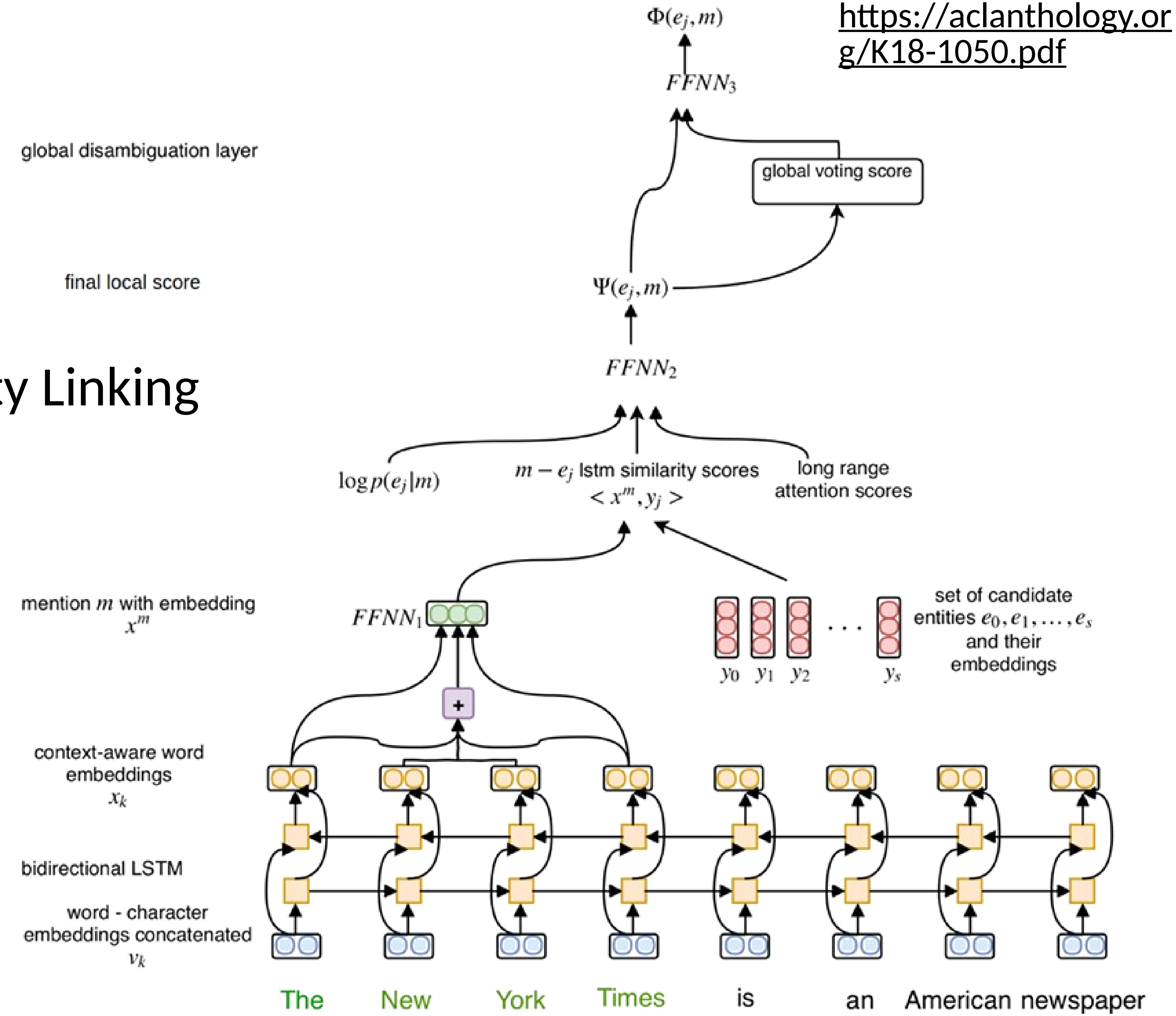
- Bi-Encoder
  - Fast
    - Pre-compute entity representations (off-line)
  - Less accurate comparing to cross-encoder

- Cross-Encoder
  - Accurate
  - Expensive to compute
    - Linear to number of candidates

https://aclanthology.org/K18-1050.pdf

End-to-End

- End-to-End Neural Entity Linking



Kolitsas N, Ganea OE, Hofmann T. End-to-End Neural Entity Linking. InProceedings of the 22nd Conference on Computational Natural Language Learning 2018 Oct (pp. 519-529).

Training loss

- max-margin loss

$$\underbrace{\mathcal{L} = \max(0, \gamma - \Phi(m, e^{+}) + \Phi(m, e^{-})),}$$

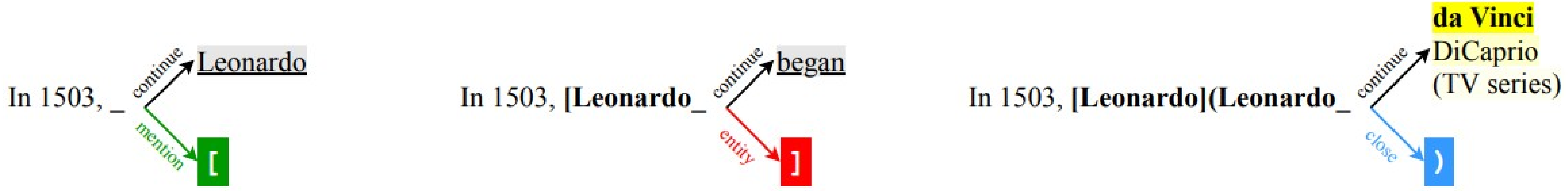
ranking score

- Cross-entropy loss

$$\mathcal{L} = -(y \log \Phi(m, e) + (1 - y) \log (1 - \Phi(m, e))),$$

Autoregressive Entity Retrieval

sequence-to-sequence model



(a) Outside: we can either continue to generate the input or start a new mention.

(b) Inside a mention: we can either continue to generate the input or end the current mention.

(c) Inside an entity link: we can either generate from the entities prefix trie or close if the generated prefix is a valid entity.

Figure 2: Example of dynamically constrained *Markup* decoding for entity linking using "*In 1503, Leonardo began painting the Mona Lisa.*" as input. There are 3 cases: when we are outside a mention/entity (a), inside a mention generation step (b), and inside an entity link generation step (c). The model is supposed to output the input source annotating mentions and pointing them to the respective entities: "*In 1503,* ***[Leonardo]***(*Leonardo da Vinci*) *began painting the* ***[Mona Lisa]***(*Mona Lisa*)".

Autoregressive Entity Retrieval

$$\text{score}(e|x) = p_\theta(y|x) = \prod_{i=1}^{N} p_\theta(y_i|y_{<i}, x)$$

- standard seq2seq objective: maximizing the output sequence likelihood
- Inference: constrained Beam search