Internet Services Architectures

Web 3.0

Szymon Olewniczak

szyolewn@pg.edu.pl

Web 3.0 vs Web3

- **Web 3.0 = Semantic Web** – an extension of the World Wide Web through standards set by the World Wide Web Consortium (W3C). The goal of the Semantic Web is to make Internet data machine-readable. (Wikipedia)
- **Web 3** – an idea for a new iteration of the World Wide Web which incorporates concepts such as decentralization, blockchain technologies, and token-based economics.

From Web 1.0 to Web 3.0

- **Web 1.0** – original Web created by Tim Berners-Lee at CERN in 1989:
  - Static Web pages written in HTML accessible through URLs.
  - Links based on URLs that allow to easily navigate between content stored on different machines.
  - The appearance of the first graphical browser Mosaic with support of submitting forms and HTTPd server with CGI started the Web 2.0 revolution.
- **Web 2.0** = participative/participatory web = social web – Web pages that allows the visitors to co-author the presented content:
  - New standards on top of the HTTP/HTML stack: CSS (1997), JavaScript (1995 with Netscape Navigator 2), cookies (1994 with Netscape Navigator 1.0).
  - Sometimes the introduction of Ajax (with IE feature called XMLHttpRequest (1999)) is considered the beginning of the Web 2.0 era.
  - Web sites are becoming more like desktop applications than static documents. But it rises a problem: the extraction of information from the Web sites in machine readable form becomes more complicated.

Is Web 2.0 still adequate?

- What has changed since introduction of Web 2.0 in 1999?
  - Introduction of Web sites based on processing the content of the other sites:
    - Google Search (1997) and other search engines.
    - Portals for comparing product prices from different resellers (e.g. ceneo.pl)
    - Many smaller project targeting smaller subdomains: marketing campaign reception, news aggregating etc.
  - Artificial Intelligence revolution (2010) - state-of-the-art Natural Language Processing techniques are learned through processing enormous amount of web content. See for example: https://chat.openai.com/
  - Web ubiquity and accessing the web from different devices: desktops, smartphones, tablets, not to mention many specific IoT devices.
- We want to Web be more machine readable that it is in its Web 2.0 form.

Web 3.0

- *"I have a dream for the Web [in which computers] become capable of analyzing all the data on the Web – the content, links, and transactions between people and computers. A "Semantic Web", which makes this possible, has yet to emerge, but when it does, the day-to-day mechanisms of trade, bureaucracy and our daily lives will be handled by machines talking to machines. The "intelligent agents" people have touted for ages will finally materialize."* - Tim Berners-Lee, 1999
- Build on top of two technologies that enables encoding of semantics with the data:
  - Resource Description Framework (RDF)
  - Web Ontology Language (OWL)
- Move from **web of documents** to **web of data**.

Resource Description Framework (RDF)

- Resource Description Framework (RDF) is a World Wide Web Consortium (W3C) standard originally designed as a data model for metadata. (Wikipedia)
- RDF is a directed graph composed of triple statements. Every statement can be either URI or literal value:
  - a node for the subject
  - an arc that goes from a subject to an object for the predicate
  - a node for the object

RDF Example: Description of a person named Eric Miller (from W3C website)

- "There is a Person identified by http://www.w3.org/People/EM/contact#me, whose name is Eric Miller, whose email address is e.miller123@example, and whose title is Dr."
- The objects are:
  - "Eric Miller" (with a predicate "whose name is"),
  - mailto:e.miller123@example (with a predicate "whose email address is"),
  - "Dr." (with a predicate "whose title is").
- The subject is a URI.
- The predicates also have URIs. For example, the URI for each predicate:
  - "whose name is" is http://www.w3.org/2000/10/swap/pim/contact#fullName,
  - "whose email address is" is http://www.w3.org/2000/10/swap/pim/contact#mailbox,
  - "whose title is" is http://www.w3.org/2000/10/swap/pim/contact#personalTitle.
- In addition, the subject has a type (with URI http://www.w3.org/1999/02/22-rdf-syntax-ns#type), which is person (with URI http://www.w3.org/2000/10/swap/pim/contact#Person).

RDF Example: Description of a person named Eric Miller (N-Triples format)

<http://www.w3.org/People/EM/contact#me> <http://www.w3.org/2000/10/swap/pim/contact#fullName> "Eric Miller" .

<http://www.w3.org/People/EM/contact#me> <http://www.w3.org/2000/10/swap/pim/contact#mailbox> <mailto:e.miller123(at)example> .

<http://www.w3.org/People/EM/contact#me> <http://www.w3.org/2000/10/swap/pim/contact#personalTitle> "Dr." .

<http://www.w3.org/People/EM/contact#me> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://www.w3.org/2000/10/swap/pim/contact#Person> .

RDF Example: Description of a person named Eric Miller (RDF/XML format)

```
<?xml version="1.0" encoding="utf-8"?>
<rdf:RDF xmlns:contact="http://www.w3.org/2000/10/swap/pim/contact#"
xmlns:eric="http://www.w3.org/People/EM/contact#" xmlns:rdf="http://www.w3.org/1999/02/22-rdf-
syntax-ns#">
   <rdf:Description rdf:about="http://www.w3.org/People/EM/contact#me">
      <contact:fullName>Eric Miller</contact:fullName>
   </rdf:Description>
   <rdf:Description rdf:about="http://www.w3.org/People/EM/contact#me">
      <contact:mailbox rdf:resource="mailto:e.miller123(at)example"/>
   </rdf:Description>
   <rdf:Description rdf:about="http://www.w3.org/People/EM/contact#me">
      <contact:personalTitle>Dr.</contact:personalTitle>
   </rdf:Description>
   <rdf:Description rdf:about="http://www.w3.org/People/EM/contact#me">
      <rdf:type rdf:resource="http://www.w3.org/2000/10/swap/pim/contact#Person"/>
   </rdf:Description>
</rdf:RDF>
```

RDF in large scale: DBpedia

- https://dbpedia.org/page/Hypertext

Web Ontology Language (OWL)

- RDF only defines the way in which knowledge graphs are constructed, not what they can contains.
- The question remains: what predicates can have different subjects?
- Ontologies are the formal way to describe the structure of knowledge for various domains.
- OWL is one of the standards that allows to express ontologies.
- Example DBPedia ontology: https://www.dbpedia.org/resources/ontology/

Schema.org - Semantic data embeded in HTML

- Schema.org is a reference website that publishes documentation and guidelines for using structured data mark-up on web-pages (called microdata). Its main objective is to standardize HTML tags to be used by webmasters for creating rich results (displayed as visual data or infographic tables on search engine results) about a certain topic of interest. (Wikipedia)
- Standardizes the items that can be marked up. For example:
  - Article
  - Breadcrumb
  - Course
  - Event
  - FAQ
  - LocalBusiness
  - Logo
  - Movie
  - Product
  - Recipe
  - Review
  - Video

Example: Schema.org + RDF

```
<div vocab="http://schema.org/" typeof="Movie">
   <h1 property="name">Avatar</h1>
   <div property="director" typeof="Person">
      Director: <span property="name">James Cameron</span>
      (born <time property="birthDate" datetime="1954-08-16">August 16,
1954</time>)
   </div>
   <span property="genre">Science fiction</span>
   <a href="../movies/avatar-theatrical-trailer.html" property="trailer">Trailer</a>
</div>
```

Web 3.0 adaptation problem

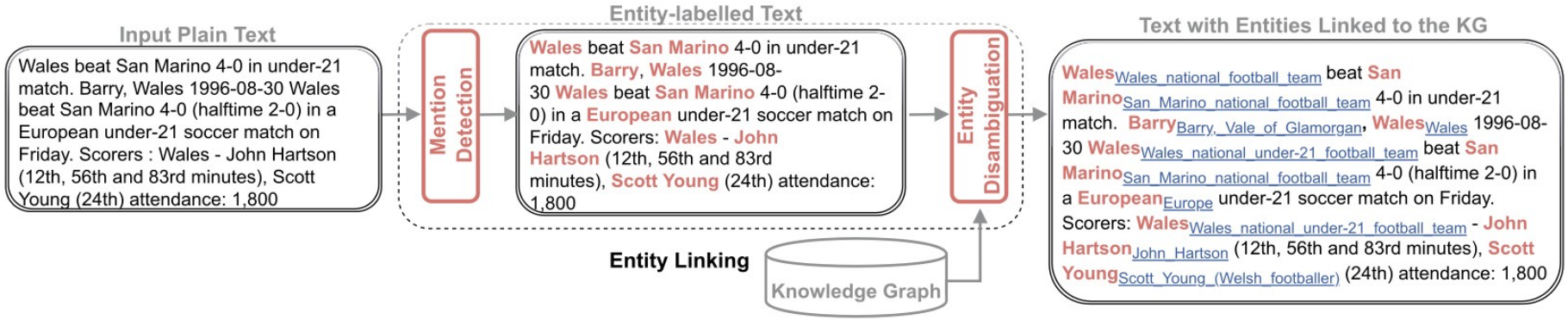
- The idea is good but requires a lot of work from the web sites creators and in practice was never widely adopted.
- Alternative idea is to create the algorithm that can extracted semantic information from web pages automatically The technique is called Entity Linking.

Entity Linking

- "Entity linking is the task of assigning a unique identity to entities (such as famous individuals, locations, or companies) mentioned in text." (Wikipedia)
- Very hot topic in the last decade. Applications in various NLP tasks:
  - Text analysis
  - Recommender systems
  - Semantic search
  - Chatbots
  - Web 3.0
- Examples:
  - "Paris[Paris] is the capital of France[France]."
  - "Paris[Paris_(band)] was an American rock music power trio formed in 1975."
  - "Paris[Paris_(plant)] is a genus of flowering plants described by Linnaeus[Carl_Linnaeus] in 1753."

Standard Architecture

- Takes plain text as an input.
- First stage: Mention Detection (MD) – discover all phrases that can be linked to Knowledge Graph (KG).
- Second stage: Entity Disambiguation (ED) – match the detected mentions with their correct definitions in KG.
- Optionally detect the mentions that are not defined in KG, so called NIL detection.



Source: SEVGILI, Özge, et al. Neural entity linking: A survey of models based on deep learning. *Semantic Web*, 2022, Preprint: 1-44.

Knowledge Graph

- Sometimes called Knowledge Base (KB), usually when structure is less formal.
- Formally can be defined as set of RDF triples: (subject, predicate, object).
- Wikipedia is still the most popular:
  - slow update cycle of other KBs. Wikipedia is updated very often and quickly incorporates new emerging facts and entities.
  - Wikipedia's entities contains much more textual content than entities in other KBs which are more predicate-oriented. What can be a little bit surprising, textual content is often more important than more formal predicate notation since it provides us a good training data.
  - Wikipedia is very popular service in the Web which can give us some important metadata like the page view statistics which can be very valuable in our EL pipeline.
- Other alternatives (sorted by decreasing popularity):
  - DBPedia: https://www.dbpedia.org/
  - Yago: https://yago-knowledge.org/
  - FreeBase (discontinued)

Mention Detection

- Detecting all phrases that can be potentially linked to KB.
- Usually we assumes that the phrases are:
  - continuous – Mentions cannot have "holes". One mention is a continuous fragment of text.
  - non-overlapping – We forbid two mentions to share the same text fragment.
  - word-level – Every mention must start with the beginning of a word and end with the end of some word.
- Usual approaches:
  - Named Entity Recognition (NER) algorithms
  - Dictionary methods
  - Successive n-grams

Candidate generation

- For every detected mention we must generate the list of possible "senses" – the entities from KB that the detected phrase may be linked to.
- For example the phrase "Paris" may be reasonable linked to the following Wikipedia articles: "Paris", "Paris (mythology)", "Paris, Ontario", "Paris (novel)", "Paris (plant)" and many others.
- Three common strategies:
  - surface form matching – we use techniques such as Levenshtein distance, n-grams overlap or normalization to find the entities nearest to the mention - doesn't work for aliases (e.g. New York - Big Apple)
  - expansion using aliases - we construct aliases table, for example by using Wikipedia disambiguation/redirect pages
    - coreference resolution – expand the given mention to the longest mention in a context
    - YAGO-means - YAGO KG provides a "means" relation that can be used to build an aliases table
  - probability + expansion using aliases - we calculate prior probabilities between certain mention and entity p(e|m), for example using Wikipedia hyperlinks.
  - Of course mixes are possible.

Candidate generation

Method	5 candidate entities for the mention "Big Blue"
surface form matching – random matches from DBpedia labels dataset	Big_Blue_Trail Big_ Bluegrass Big_Blue_Spring_cave_crayfish Dexter_Bexley_and_the_Big_Blue_Beastie IBM_Big_ Blue_(X-League)
expansion using aliases from YAGO-means	Big_Blue_River_(Indiana) Big_ Blue_River_(Kansas) Big_Blue_(crane) Big_ Red_(drink) **IBM**
probability + expansion using aliases from KnowBert (Anchor prob. + CrossWikis + YAGO)	**IBM** Big_Blue_River_(Kansas) The_Big_Blue Big_Blue_River_(Indiana) Big_Blue_(crane)

Source: SEVGILI, Özge, et al. Neural entity linking: A survey of models based on deep learning. *Semantic Web*, 2022, Preprint: 1-44.

Entity Disambiguation

- The most active field of research, many different techniques. We can divide approaches according to several criteria.
- By the main source of features:
  - Text-based approaches – we use textual features from large text corpora (e.g. word co-occurrence probabilities)
  - Graph-based approaches – we exploit the structure of KG to represent the context and the relation of entities (e.g. Wikipedia hyperlinks).
- Local and Global approaches:
  - Local approaches – we disambiguate every mention separately (better for short texts, resistant to context drift).
  - Global approaches – we disambiguate entire textual document as a whole, by assuming that mentions in a one document are more-or-less coherent (better for longer texts but sensitive to context drift problem).
- By the used methodology:
  - Without learning component - usually based on some statistical features like conditional probabilities.
  - Machine Learning - systems based on shallow ML algorithms such as Naive Bayes, SVM or Decision Trees.
  - Deep Learning - systems using Deep Neural Networks.

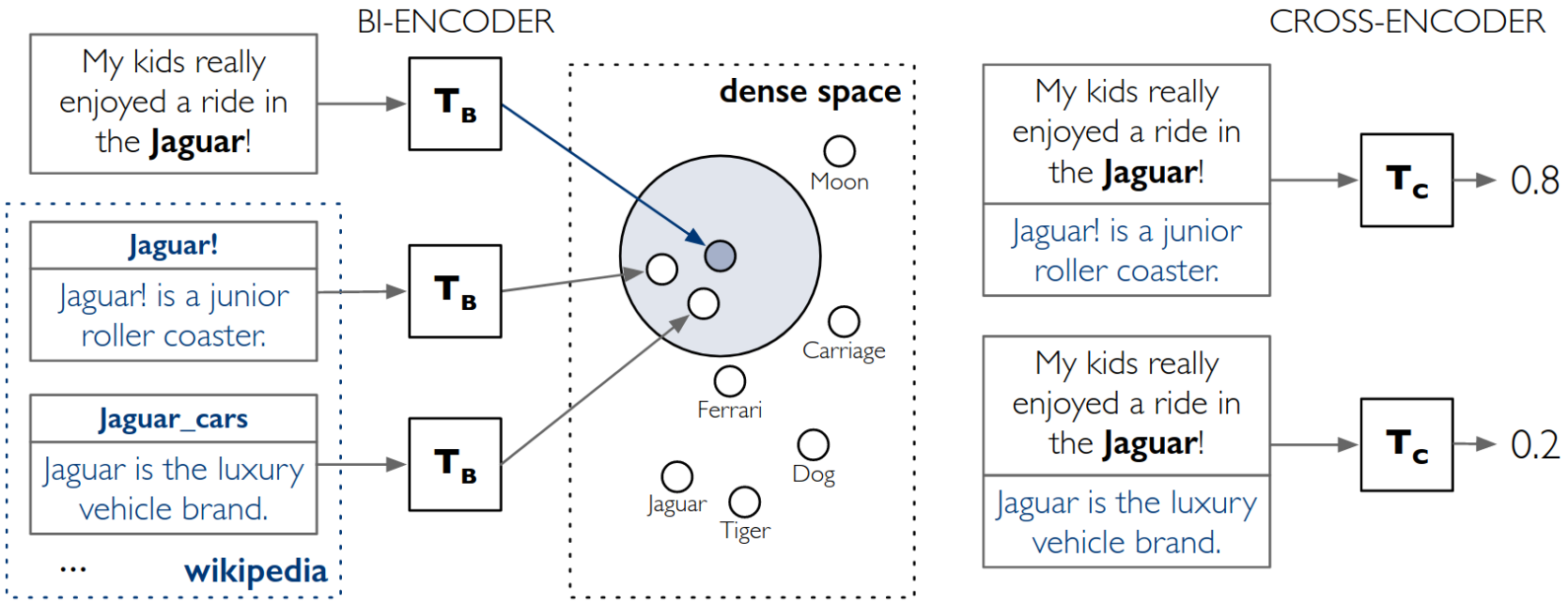
Example: DBPedia Spotlight

- Uses DBPedia as a KB.
- Allow us to retrieve only subset of entities based on Ontology or SPARQL.
- https://www.dbpedia-spotlight.org/

GDAŃSK UNIVERSITY OF TECHNOLOGY

Example: BLINK

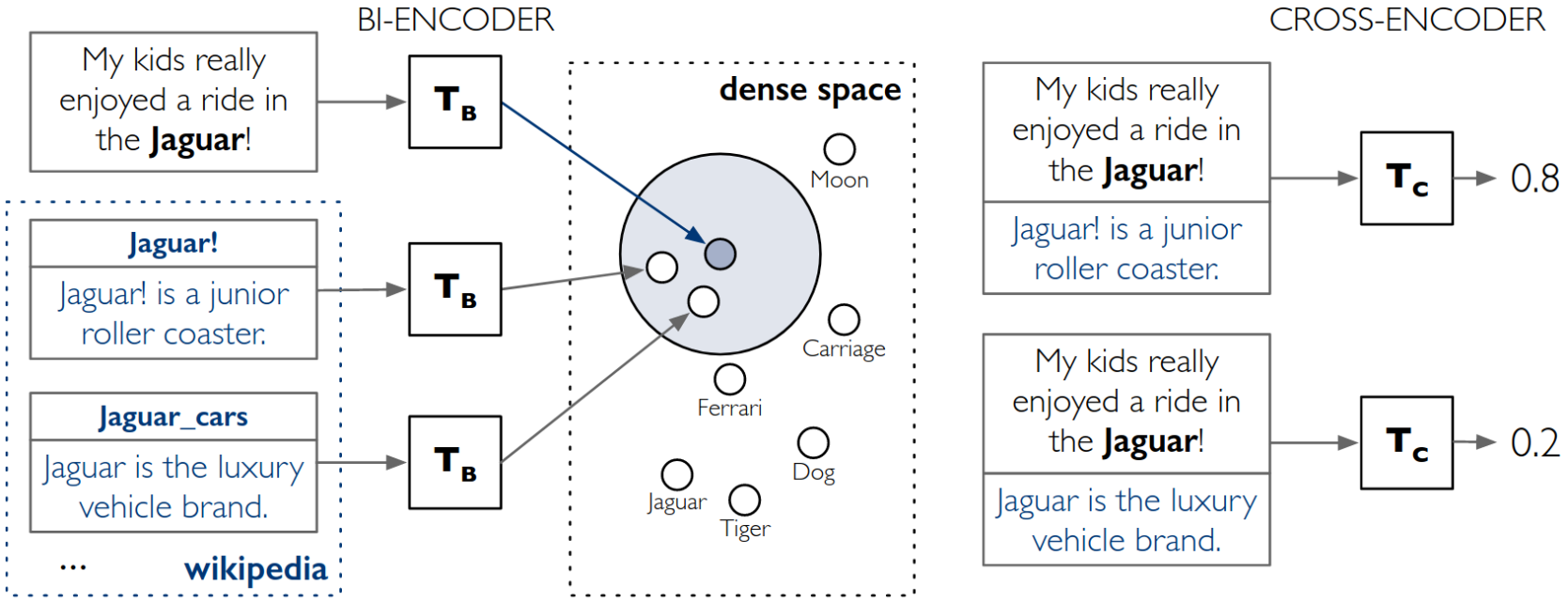
- Two-stage zero-shot linking algorithm based on BERT (Wu et al., 2020)
- Example of state-of-the-art DNN-based local approach.
- Zero-shot entity linking – KB is separated in training and test time.
- Works for Entity linking with gold mentions setup – no mention detection phrase.



Source: WU, Ledell, et al. Scalable zero-shot entity linking with dense entity retrieval. arXiv preprint arXiv:1911.03814, 2019.

Example: BLINK

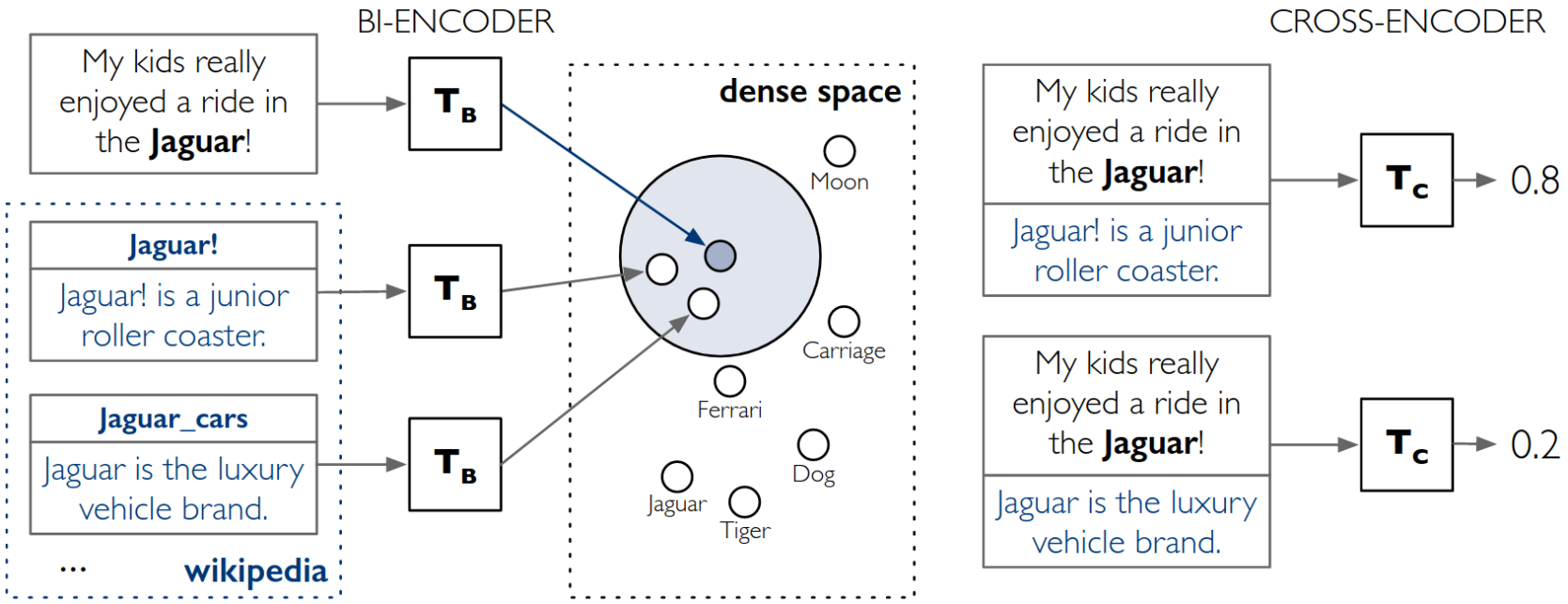
- Bi-encoder – inputs and KB entities gets encoded in the same dense space.
- Then the k near neighbors (<100) are retrieved and supplied to the cross encoder for final evaluation.
- Bi-encoder – two BERT models for input context and candidates entities trained jointly by minimizing the dot-product for the positive pairs using random and hard negatives.
- At inference time, the entity representation for all the entity candidates can be precomputed and cached.



Source: WU, Ledell, et al. Scalable zero-shot entity linking with dense entity retrieval. arXiv preprint arXiv:1911.03814, 2019.

Example: BLINK

- Cross-encoder – input is the concatenation of input context and the candidate mention representation.
- Then the third BERT model is used that produces the T_c embedding.
- T_c embedding is then multiplied by additional weight matrix W to produce the scalar value – the final rating for the given mention.
- Additional Knowledge Distillation – cross-encoder as a teacher for bi-encoder.



Source: WU, Ledell, et al. Scalable zero-shot entity linking with dense entity retrieval. arXiv preprint arXiv:1911.03814, 2019.

GDAŃSK UNIVERSITY
OF TECHNOLOGY

HISTORY IS WISDOM
FUTURE IS CHALLENGE