Text-Symbol Alignment and Semantic Annotations

Entity Linking

April 6th, 2023
Prof.dr.ir. Arjen P. de Vries
github.com/laura-dietz/neurosymbolic-representations-for-IR/

Radboud University

TUTORIAL OUTLINE

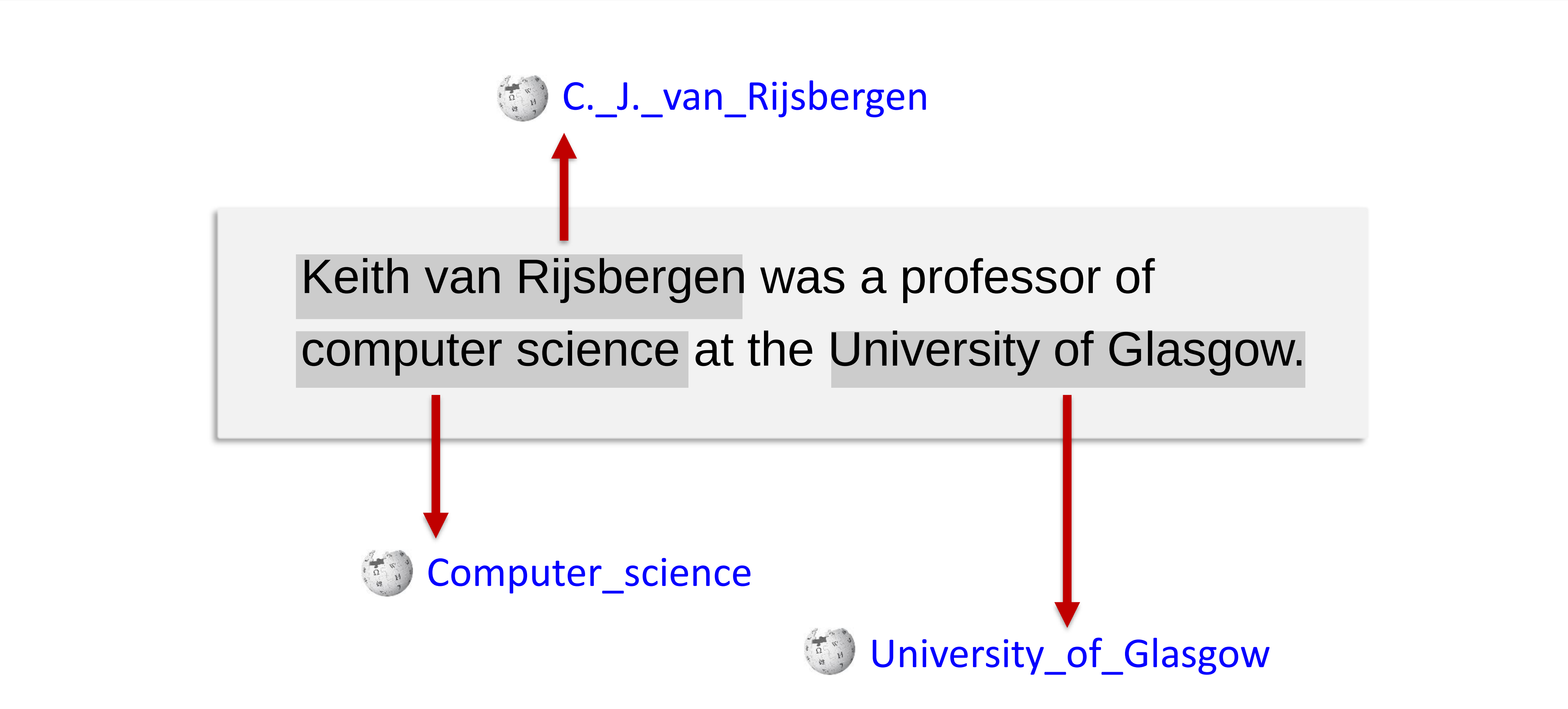
NEUROSYMBOLIC REPRESENTATIONS FOR INFORMATION RETRIEVAL

- Part 1: Symbolic AI representations and tasks
    - Welcome/Purpose of this tutorial
    - (Sub)symbolic AI, and representations
    - Question Answering on Knowledge Graphs
- Part 2: Text-to-symbols and Ranking
    - Neural Text Representations
    - Text-Symbol Alignment and Semantic Annotations ← we are here ☺
    - Entity Representations and Entity Ranking
- Part 3: Neuro-symbolic representations for Reasoning
    - Reasoning about Relevance
    - Neuro Pseudo-Relevance Feedback with Explainability
- Part 4: Applications for Neuro-symbolic approaches
    - Use Case: Knowledge Discovery
    - Use Case: Task-based Assistance
    - Use Case: Generating relevant (long-form) Articles
- Panel & Discussion

github.com/laura-dietz/neurosymbolic-representations-for-IR/

ENTITY LINKING

HOW TO TRANSITION FROM TEXT TO SYMBOLS?



C._J._van_Rijsbergen

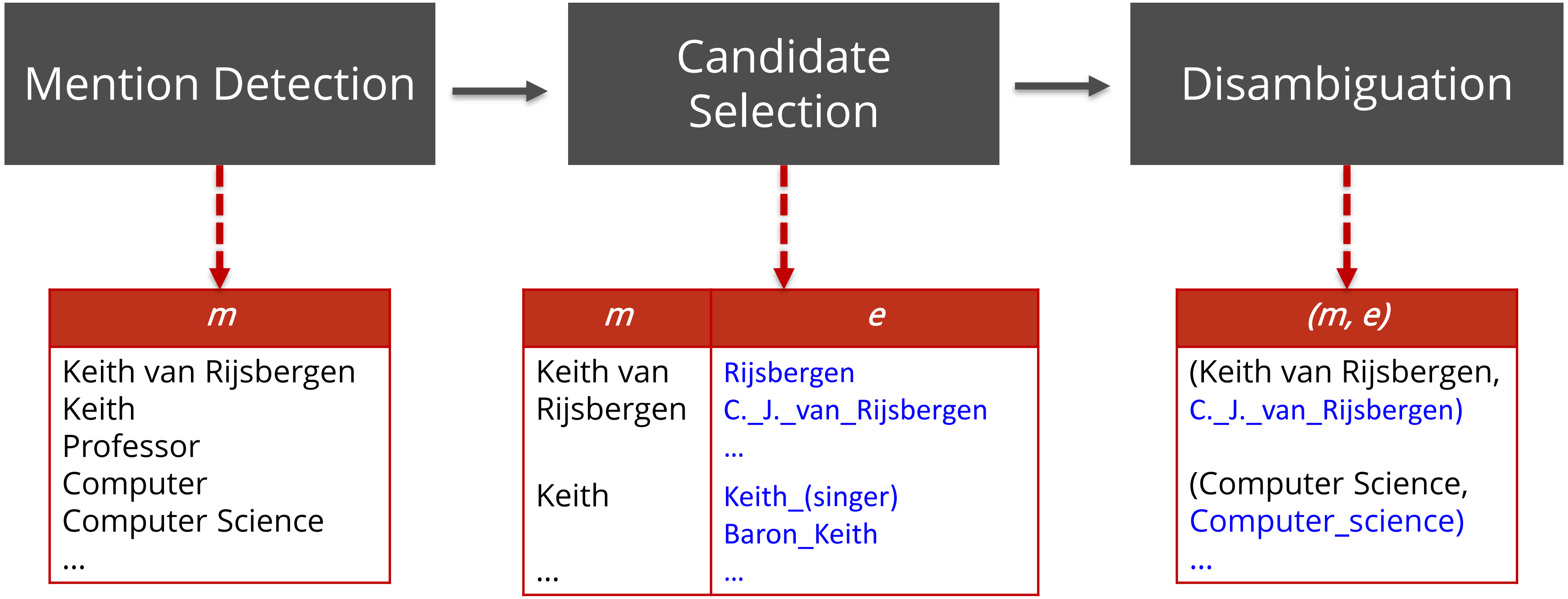
Keith van Rijsbergen was a professor of computer science at the University of Glasgow.

Computer_science

University_of_Glasgow

ENTITY LINKING

STANDARD PIPELINE



m
Keith van Rijsbergen
Keith
Professor
Computer
Computer Science
…

m	*e*
Keith van Rijsbergen	Rijsbergen C._J._van_Rijsbergen …
Keith	Keith_(singer) Baron_Keith
…	…

(m, e)
(Keith van Rijsbergen, C._J._van_Rijsbergen)
(Computer Science, Computer_science)
…

Krisztian Balog, Entity-Oriented Search (2018).
https://eos-book.org/

ENTITY LINKING
MENTION DETECTION

- Identify all “linkable phrases” (mentions) in the text

- Recall oriented
  - Do not miss any mention that should be linked, because we cannot recover in later stages
- Find entity name variants
  - E.g. “jlo” is a name variant of [Jennifer Lopez]

ENTITY LINKING
SURFACE FORM DICTIONARY

- Page title
  - The most common name of the entity
- Redirect pages
  - Alternative name for referring to the entity
- Disambiguation pages
  - Entities that share the same name
- Anchor text
  - Wikipedia hyperlinks

- Use other sources, e.g. YAGO, annotated corpora (e.g., ClueWeb with Freebase annotations, shared by Google), *etc.*

ENTITY LINKING

MENTION DETECTION

- Probability of a word being linked to an entity

number of times mention *m*
appears as a link

$$P(\text{link}|m) = \frac{\text{link}(m)}{\text{freq}(m)}$$

number of times mention *m*
appears in the text (linked or not)

ENTITY LINKING

CANDIDATE SELECTION

- Narrow down the space of disambiguation possibilities, by ranking the entities for each mention
  - *Very important: exact formulation for entity disambiguation leads to an NP-complete problem*
- Types of features used:
  - Popularity
  - Textual similarities
  - Entity relatedness in the knowledge graph

ENTITY LINKING

CANDIDATE SELECTION

- Commonness:
  - Probability of a word referring to the entity:

number of times entity *e* is the link target of mention *m*

$$P(e|m) = \frac{n(m,e)}{\sum_{e'} n(m,e')}$$

total number of times mention *m* appears as link

ENTITY LINKING
DISAMBIGUATION

- Disambiguation strategies
  - **Individually**: one-mention-at-a-time
    E.g., rank candidates for each mention, take the top ranked one (or NIL)
    Take into account **entity coherency**?
  - **Collectively**: all mentions in the document jointly
    E.g., using graph based approaches

ENTITY LINKING
OFTEN: A “BLACK BOX”

- *Quite a complex task, as we will see!*
- Typically delegated to a third-party toolkit, in the hope that it does this well
  - *EM-BERT – details later – does not work well on MS Marco because only a very low percentage of queries have entities that are detected by these tools!*

DBpedia Spotlight

Nordlys

SMAPH
Linking Queries to Knowledge

TAGME

WAT
The Wikipedia Annotator of Text

ENTITY LINKING

SHORTCOMINGS

- Not keeping up with the recent NLP progress
  - More recent neural approaches are rarely integrated
- Designed for short texts and inefficient for long texts
- Lack of speed (throughput)
- Requiring large computational power
- Reliance on external sources like web search engines

ENTITY LINKING

RADBOUD ENTITY LINKER (REL)

Advertorial!

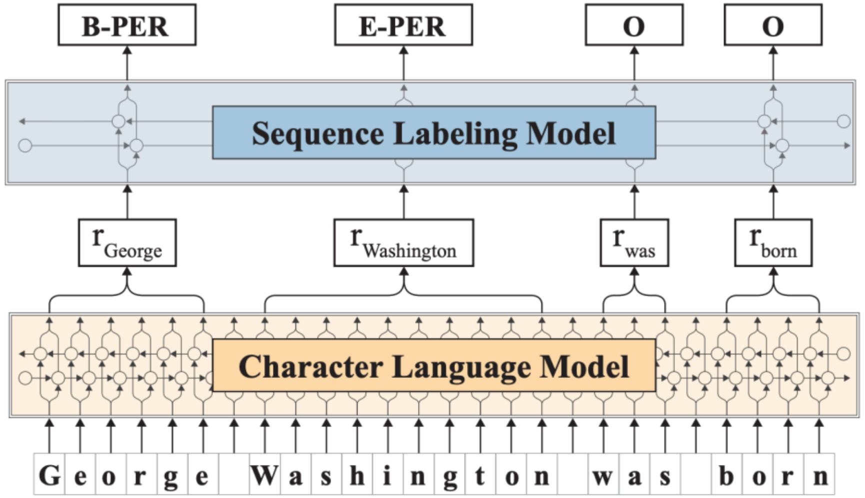
- Fast and lightweight
  - Can be deployed on an average laptop/desktop machine
  - Does not need much RAM and GPU

- Modular, available as a python package
  - NER component can be replaced by other algorithms/tools
  - Can be used for NER, ED, and end-to-end entity linking

van Hulst, J. M., Hasibi, F., Dercksen, K., Balog, K., & de Vries, A. P. (2020). REL: An Entity Linker Standing on the Shoulders of Giants. SIGIR '20.

ENTITY LINKING
MENTION DETECTION

- Mention detection uses the Flair named entity recognizer: a bidirectional character-level language model and a sequence labeling module to generate NER tags



Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual String Embeddings for Sequence Labeling. In Proceedings of the 27th International Conference on Computational Linguistics, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

https://alanakbik.github.io/flair.html
https://towardsdatascience.com/contextual-embeddings-for-nlp-sequence-labeling-9a92ba5a6cf0

ENTITY LINKING

MENTION DETECTION

- Flair uses “stacked embeddings” –
  - Concatenates the contextual embeddings with GloVe embeddings to represent words for the sequence tagging stage

MENTION DETECTION

- Alternatives for mention detection:
  - Dictionary-based *(e.g., when we needed Dutch medical entity linking...)*
  - Spacy – may be more efficient than Flair, models for many languages, usually very quick on integrating new ideas from NLP community: https://github.com/explosion/spaCy
  - BERT as a classifier to assign NER labels;
    see also https://github.com/google-research/bert/issues/223 (and this comment)

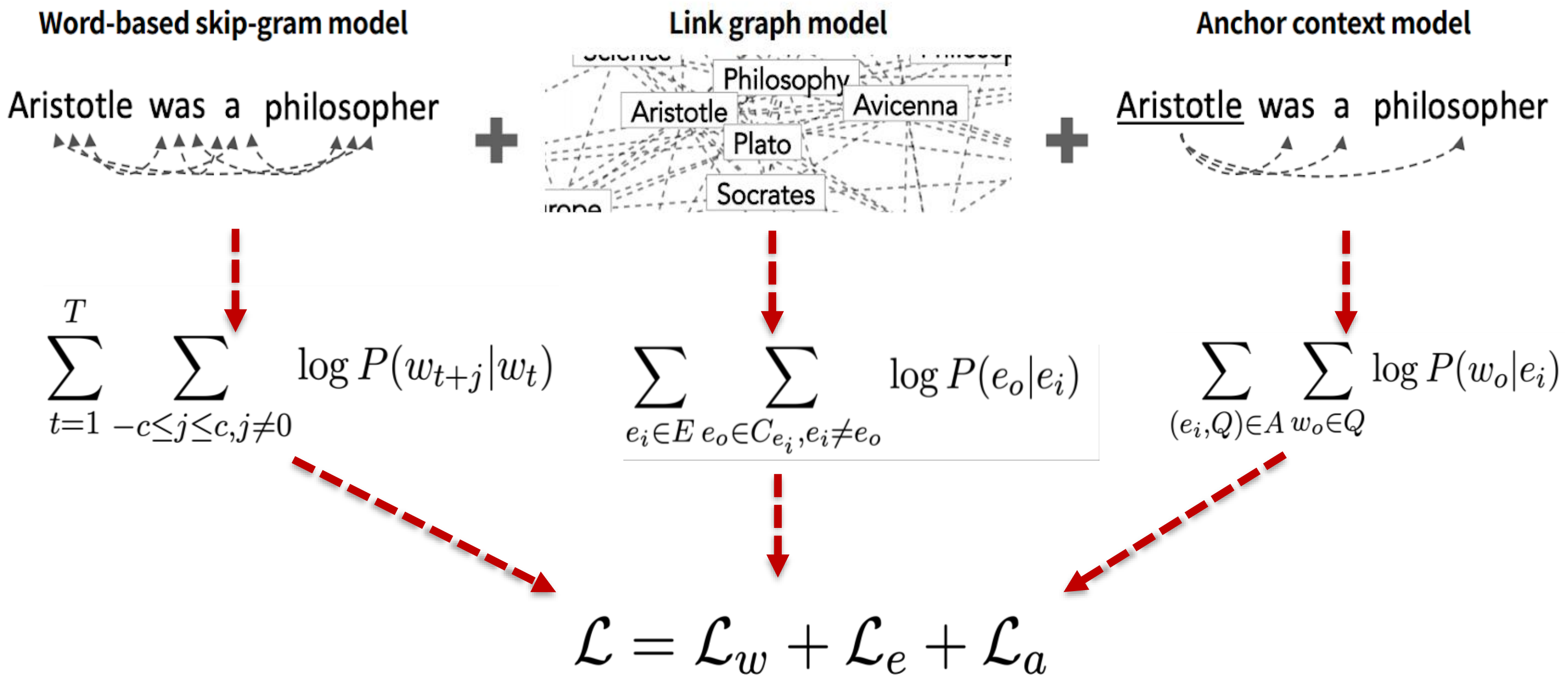
ENTITY LINKING
CANDIDATE SELECTION

- Select up to 7 candidate entities for every mention:
  - 4 entities based on the prior probability $p(e|m) = \frac{n(m,e)}{\sum_{e\prime} n*m,e\prime)}$

    Estimated using the surface dictionary
  - 3 entities based on their similarity to the mention's context (50 tokens)
    Estimated using Wikipedia2Vec embeddings

ENTITY LINKING

WIKIPEDIA2VEC EMBEDDINGS


Word-based skip-gram model
Aristotle was a philosopher
+
Link graph model
Philosophy
Aristotle
Avicenna
Plato
Socrates
+
Anchor context model
Aristotle was a philosopher

$$\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log P(w_{t+j}|w_t)$$

$$\sum_{e_i\in E}\sum_{e_o\in C_{e_i}, e_i\neq e_o}\log P(e_o|e_i)$$

$$\sum_{(e_i,Q)\in A}\sum_{w_o\in Q}\log P(w_o|e_i)$$

$$\mathcal{L}=\mathcal{L}_w+\mathcal{L}_e+\mathcal{L}_a$$

Joint Learning of the Embedding of Words and Entities for Named Entity Disambiguation (Yamada et al., CoNLL 2016)
Wikipedia2Vec: An Efficient Toolkit for Learning and Visualizing the Embeddings of Words and Entities from Wikipedia (Yamada et al., EMNLP 2020)

https://wikipedia2vec.github.io/wikipedia2vec/

ENTITY LINKING

ENTITY DISAMBIGUATION

- Combines **local compatibility** of entity-mention pair $\psi(.)$ with **global compatibility** of all entity linking decisions $\phi(.)$

$$\arg\max_{E \in C_1 \times \ldots \times C_n} \sum_{i=1}^{n} \Psi(e_i, c_i) + \sum_{i \neq j} \Phi(e_i, e_j, D)$$

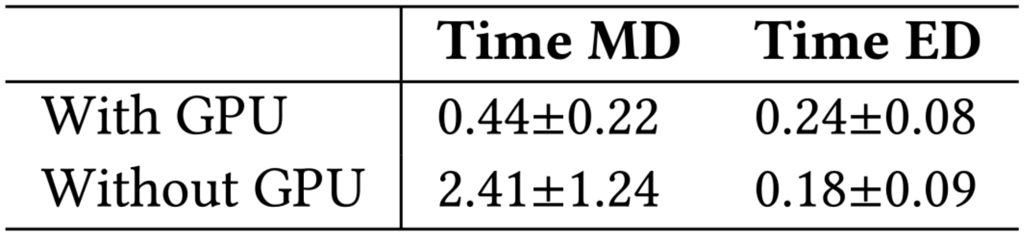
- $\psi(.)$ combines context word vectors, candidate entity priors and embeddings *(but ignores coherence)*
- $\phi(.)$ assumes K latent relations between every two mentions and computes global combability using Loopy Belief Propagation, an approximate inference method based on message passing

Deep Joint Entity Disambiguation with Local Neural Attention (Ganea & Hofmann, EMNLP 2017)
Improving Entity Linking by Modeling Latent Relations between Mentions (Le & Titov, ACL 2018)

ENTITY LINKING

EFFICIENCY

- REL: measured (in seconds per document) for 50 documents with > 200 words from AIDA-B, each containing in average 323 (± 105) words and 42 (± 19) mentions



	Time MD	Time ED
With GPU	0.44±0.22	0.24±0.08
Without GPU	2.41±1.24	0.18±0.09

- Mention detection most expensive, where GPU is really desirable

ENTITY LINKING

TAGGING MS MARCO V2

- MS MARCO v2 is a Web collection often used for deep learning IR research
  - 12 million web documents
  - 33 GB in size (60 compressed JSONL files)

- Linking using default REL installation way slower than expected ☹
  - Longer documents (5x) and more mentions (2x)

- Modifications to REL for dataset-at-a-time tagging (REBL):
  - Separate Mention Detection completely from Candidate Selection and Entity Disambiguation, so the expensive mention detection can be run on our GPU cluster – store intermediate representations
  - Batch sentences for processing by FLAIR
  - Memory requirements for internal data representations could be improved by a factor of 5
  - Explicit GPU memory management (clear embeddings after every document)

ENTITY LINKING

MMEAD

Advertorial!

- MMEAD is a specification for entity links for MSMARCO
  - JSON specification for sharing and using entity links
  - Pretrained Wikipedia2Vec embeddings
  - Python library to use both resources (entity links and embeddings) easily
- MS Marco V1 and V2, passage and doc, tagged with entity annotations
  - RE(B)L
  - BLINK *(only V1 passage completed right now)*
- Ready to use!

DuckDB

Chris Kamphuis, Aileen Lin, Siwen Yang, Jimmy Lin, Arjen de Vries and Faegheh Hasibi.
MMEAD: MS MARCO Entity Annotations and Disambiguations. In SIGIR 2023 (!).
https://github.com/informagi/MMEAD

ENTITY LINKING

EXAMPLE: MMEAD IN QLEVER

- Take MMEAD, reformat the entity links into RDF, and ingest the results into QLever
- Combining MMEAD with RDF data from Wikidata and OpenStreetMap, we can issue SPARQL queries like “show me all passages in MS MARCO about France”

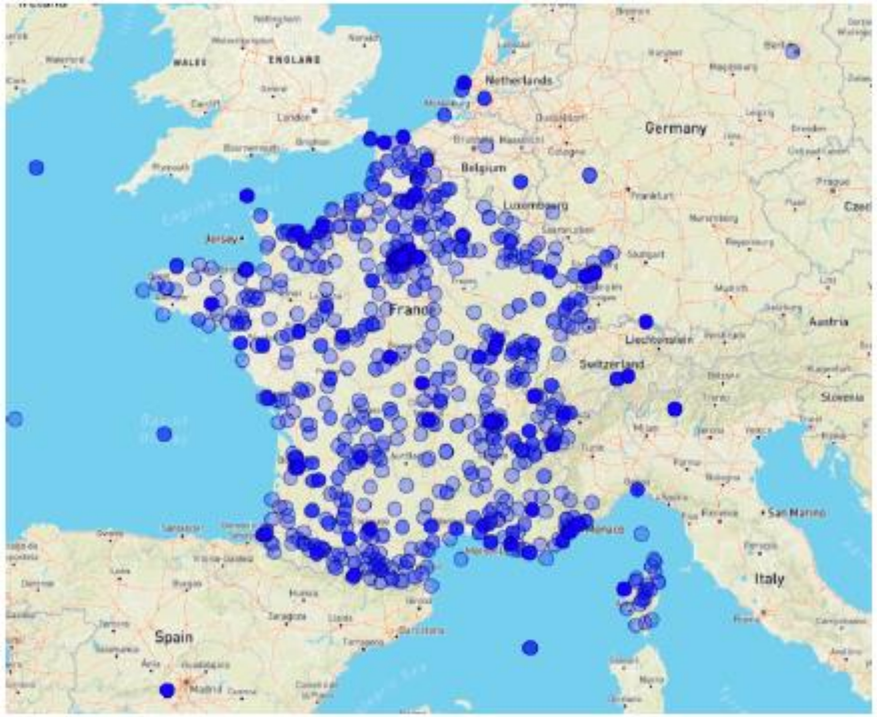
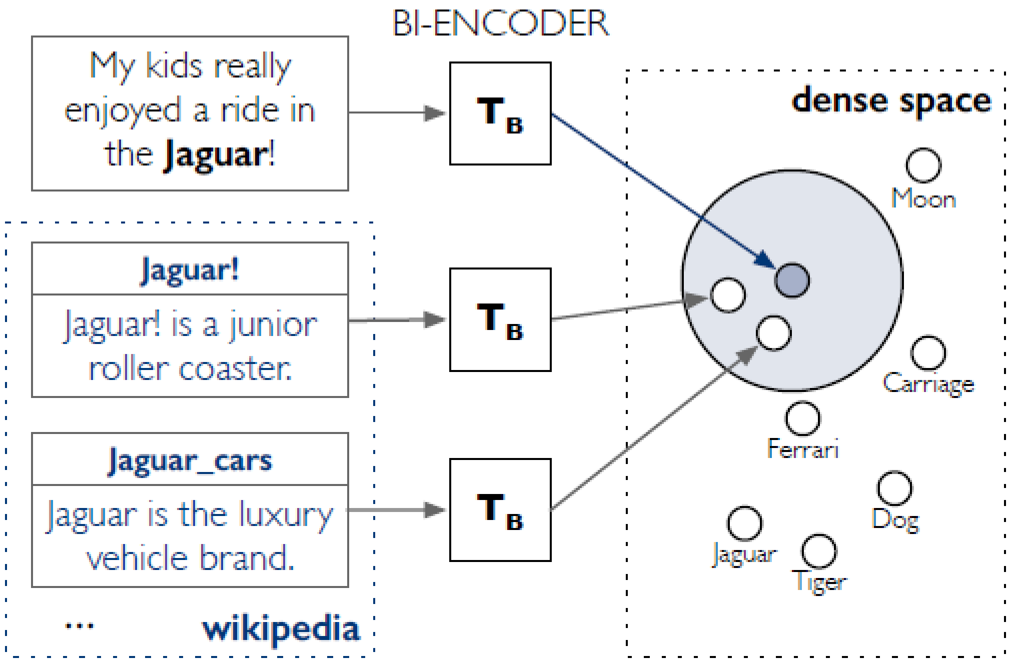


Figure 11: First 100 entities found in that are connected to France. Entities are represented with a blue dot on the map.

ENTITY LINKING
META'S BLINK ENTITY LINKER

- Mention Detection: Flair
- Candidate Selection:
  - What entities in the KB are close to the query entity?
  - Using bi-encoders with FAISS / HNSW for efficiency


BI-ENCODER
My kids really enjoyed a ride in the **Jaguar**!
T_B
T_B
T_B
Jaguar!
Jaguar! is a junior roller coaster.
Jaguar_cars
Jaguar is the luxury vehicle brand.
...
wikipedia
dense space
Moon
Carriage
Ferrari
Dog
Jaguar
Tiger

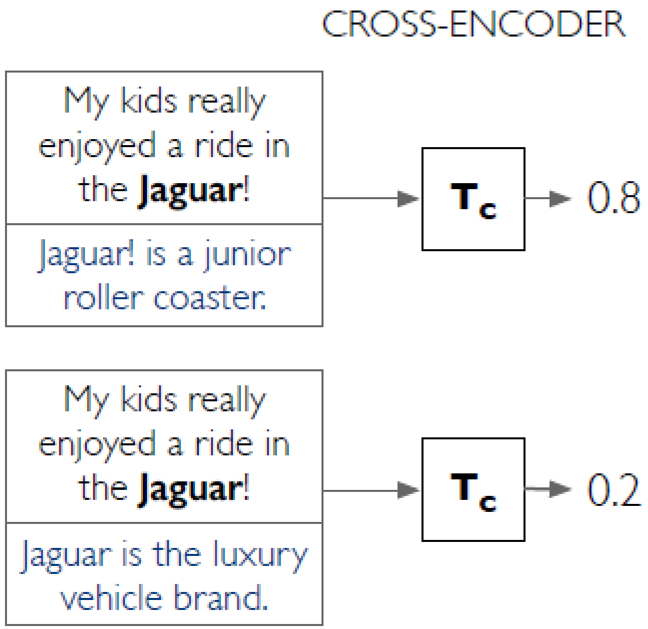
Scalable Zero-shot Entity Linking with Dense Entity Retrieval (Wu et al., EMNLP 2020)
https://github.com/facebookresearch/BLINK

ENTITY LINKING

META'S BLINK ENTITY LINKER

- Mention Detection: Flair
- Candidate Selection
  - What entities in the KB are close to the query entity?
  - Using bi-encoders with FAISS / HNSW for efficiency

- Entity Disambiguation
  - Attend to the input and each of the candidates' entity descriptions to create a probability distribution over the candidates
  - Using cross-encoder for effectiveness


CROSS-ENCODER
My kids really enjoyed a ride in the **Jaguar**!
Jaguar! is a junior roller coaster.
T_c
0.8
My kids really enjoyed a ride in the **Jaguar**!
Jaguar is the luxury vehicle brand.
T_c
0.2

Scalable Zero-shot Entity Linking with Dense Entity Retrieval (Wu et al., EMNLP 2020)
https://github.com/facebookresearch/BLINK

ENTITY LINKING

META'S BLINK/ELQ: ENTITY LINKING FOR QUESTIONS

- Approach to perform mention detection as part of the inference process:
  - Consider all spans [i; j] (i-th to j-th tokens of q) in the text up to length L
  - Estimate a probability for span [i; j] to be a mention, by adding scores for every token in the span to be a member of an entity mention, and two extra scores for s(i) and s(j) to be a start of end token, respectively
  - Learned using binary cross-entropy loss across all mention candidates

Efficient One-Pass End-to-End Entity Linking for Questions (Li et al., EMNLP 2020)
https://github.com/facebookresearch/BLINK/tree/main/elq

OPEN TOPICS

ENTITY LINKING

BUT THE WORLD NEVER STOPS!

- Entity Linkers like REL and BLINK are not updated continuously
  - Yes… that is our plan… but it is more complicated than anticipated!

- Dependencies on data include:
  - Wikipedia dump – what year/month?
  - Associated Wikipedia2Vec embeddings
  - Surface dictionary with prior probability estimates, drawn from Wikipedia, Web data and Yagoo
  - GloVe embeddings

- Some of these dependencies may also interact with mention detection (i.e., Flair)

- *Can we get this under control, without too much of a hazzle?*

ENTITY LINKING

LINKING TO OTHER KNOWLEDGE GRAPHS THAN WIKIPEDIA

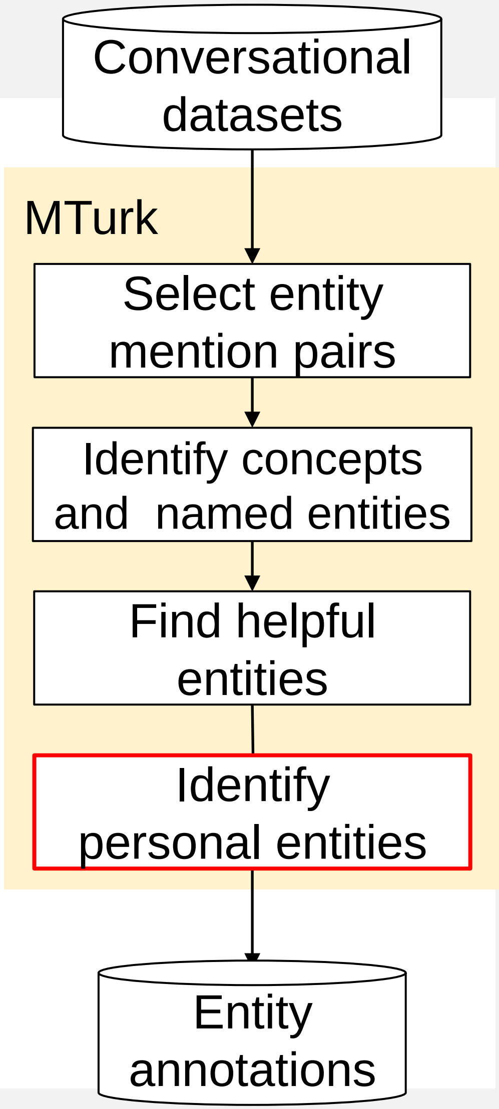
- Target knowledge graph is usually Wikipedia / WikiData, but…
- UMLS for medical?
- Product KG for e-commerce?

- *Resources like surface dictionary and graph embeddings have to be constructed?!*
  "Hyperparameters" such as the number of mentions considered as candidates?

ENTITY LINKING

WHAT ABOUT CONVERSATIONS INSTEAD OF DOCUMENTS?

- Conversations are informal
  - References to entities are made by their pronouns
  - E.g., “my city”, “my guitar”, “its population”
- Based on an annotation process, we found:
  - 33% of dialogues in social chat contain **personal entities** (not in KG)
  - 57% of the entities *marked as helpful* by crowd workers are **concepts** E.g., “I'm looking for a hotel with a 3 star rating”



Conversational datasets
MTurk
Select entity mention pairs
Identify concepts and named entities
Find helpful entities
Identify personal entities
Entity annotations

Hideaki Joko, Faegheh Hasibi, Krisztian Balog, and Arjen P. de Vries. “Conversational Entity Linking: Problem Definition and Datasets”. SIGIR '21

github.com/informagi/conversational-entity-linking

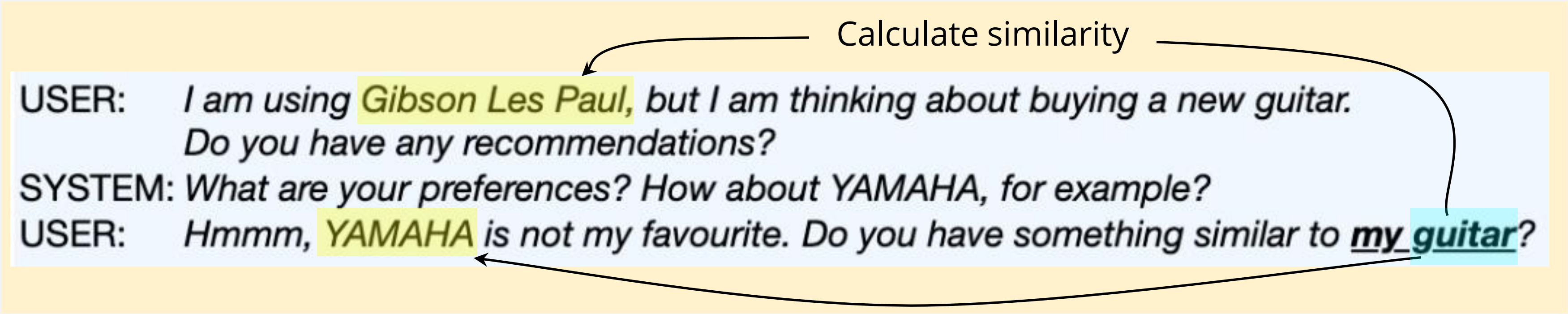
Hideaki Joko and Faegheh Hasibi. Personal Entity, Concept, and Named Entity Linking in Conversations. In CIKM '22. 10.1145/3511808.3557667

ENTITY LINKING

Personal Entities

Does a simple extension of traditional EL work for personal entities?

Step 1: **Calculate cosine similarity of the embeddings between the reference of personal entity and all entities in a dialogue history**



USER: *I am using Gibson Les Paul, but I am thinking about buying a new guitar. Do you have any recommendations?*
SYSTEM: *What are your preferences? How about YAMAHA, for example?*
USER: *Hmmm, YAMAHA is not my favourite. Do you have something similar to **my guitar**?*

Step 2: Select the entity which has the highest similarity
(Entities which do not exceed the threshold are ignored)

ENTITY DISAMBIGUATION

CO-REFERENCE RESOLUTION

- Coreference resolution = cluster multiple mentions of the same entity within a given text
- S-2-E: Start-to-End Coreference Resolution
  - Using a smart trick from dependency parsing literature (“deep biaffine attention”) to determine efficiently how likely the span’s start/end token qs/qe is a beginning/ending of an entity mention, and whether those tokens are the boundary points of the same entity mention.

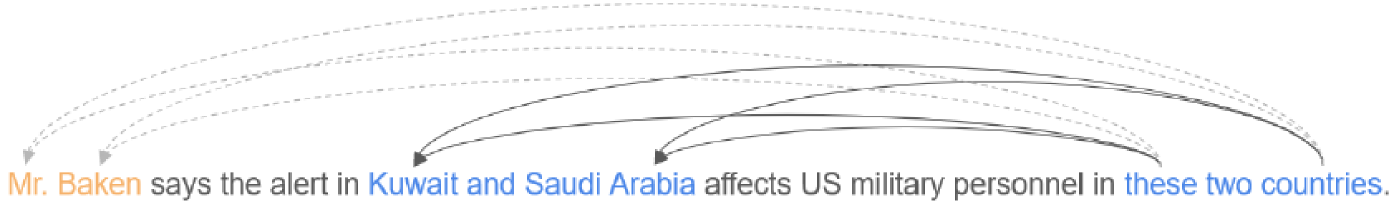

Mr. Baken says the alert in Kuwait and Saudi Arabia affects US military personnel in these two countries.

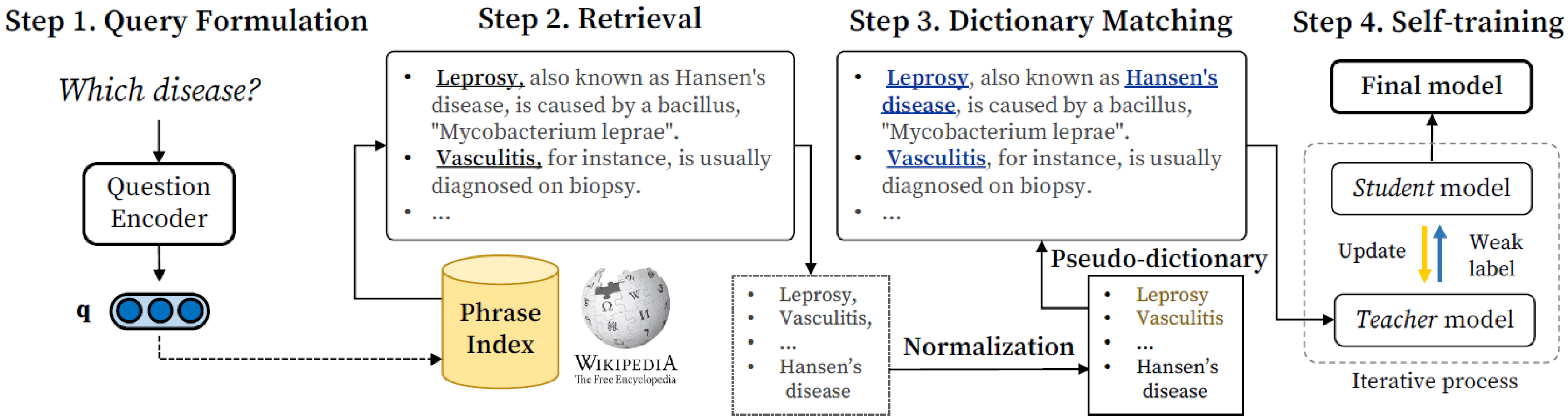
Figure 1: The antecedent score $f_a(c, q)$ of a query mention $q = (q_s, q_e)$ and a candidate antecedent $c = (c_s, c_e)$ is defined via bilinear functions over the representations of their endpoints c_s, c_e, q_s, q_e. Solid lines reflect factors participating in positive examples (coreferring mentions), and dashed lines correspond to negative examples.

Coreference Resolution without Span Representations (Kirstain et al., ACL-IJCNLP 2021)
https://github.com/yuvalkirstain/s2e-coref/

ENTITY LINKING

GENERATE ENTITY LINKERS ON-THE-FLY?

- Highly flexible “ask-to-generate” approach:


Step 1. Query Formulation
Step 2. Retrieval
Step 3. Dictionary Matching
Step 4. Self-training
Which disease?
Question Encoder
q
Phrase Index
WIKIPEDIA
The Free Encyclopedia
Leprosy, also known as Hansen's disease, is caused by a bacillus, "Mycobacterium leprae".
Vasculitis, for instance, is usually diagnosed on biopsy.
...
Leprosy, also known as Hansen's disease, is caused by a bacillus, "Mycobacterium leprae".
Vasculitis, for instance, is usually diagnosed on biopsy.
...
Leprosy,
Vasculitis,
...
Hansen’s disease
Normalization
Pseudo-dictionary
Leprosy
Vasculitis
...
Hansen’s disease
Final model
Student model
Update
Weak label
Teacher model
Iterative process

Simple Questions Generate Named Entity Recognition Datasets (Kim et al., EMNLP 2022)
Jinhyuk Lee, Mujeen Sung, Jaewoo Kang, Danqi Chen. Learning dense representations of phrases at scale. ACL-IJCNLP 2021

github.com/dmis-lab/GeNER/
github.com/princeton-nlp/DensePhrases

TUTORIAL OUTLINE

NEUROSYMBOLIC REPRESENTATIONS FOR INFORMATION RETRIEVAL

- Part 1: Symbolic AI representations and tasks
  - Welcome/Purpose of this tutorial
  - (Sub)symbolic AI, and representations
  - Question Answering on Knowledge Graphs
- Part 2: Text-to-symbols and Ranking
  - Neural Text Representations
  - Text-Symbol Alignment and Semantic Annotations
  - Entity Representations and Entity Ranking ← we are here ☺
- Part 3: Neuro-symbolic representations for Reasoning
  - Reasoning about Relevance
  - Neuro Pseudo-Relevance Feedback with Explainability
- Part 4: Applications for Neuro-symbolic approaches
  - Use Case: Knowledge Discovery
  - Use Case: Task-based Assistance
  - Use Case: Generating relevant (long-form) Articles
- Panel & Discussion

github.com/laura-dietz/neurosymbolic-representations-for-IR/