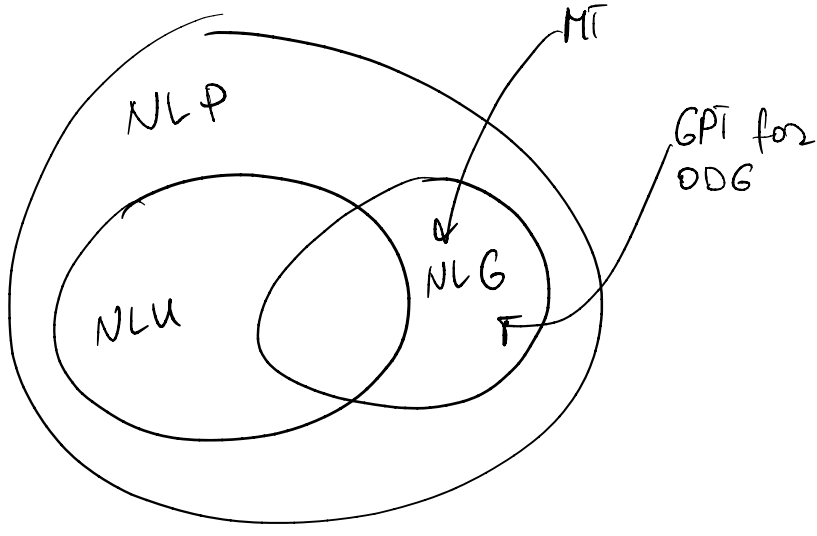
Information extraction

Katya Artemova, DL for NLP



NLU - natural language understanding

NLG - natural language generation

What is the weather like today?

↓

intent : get_weather

slot filling :

17/01/22 def

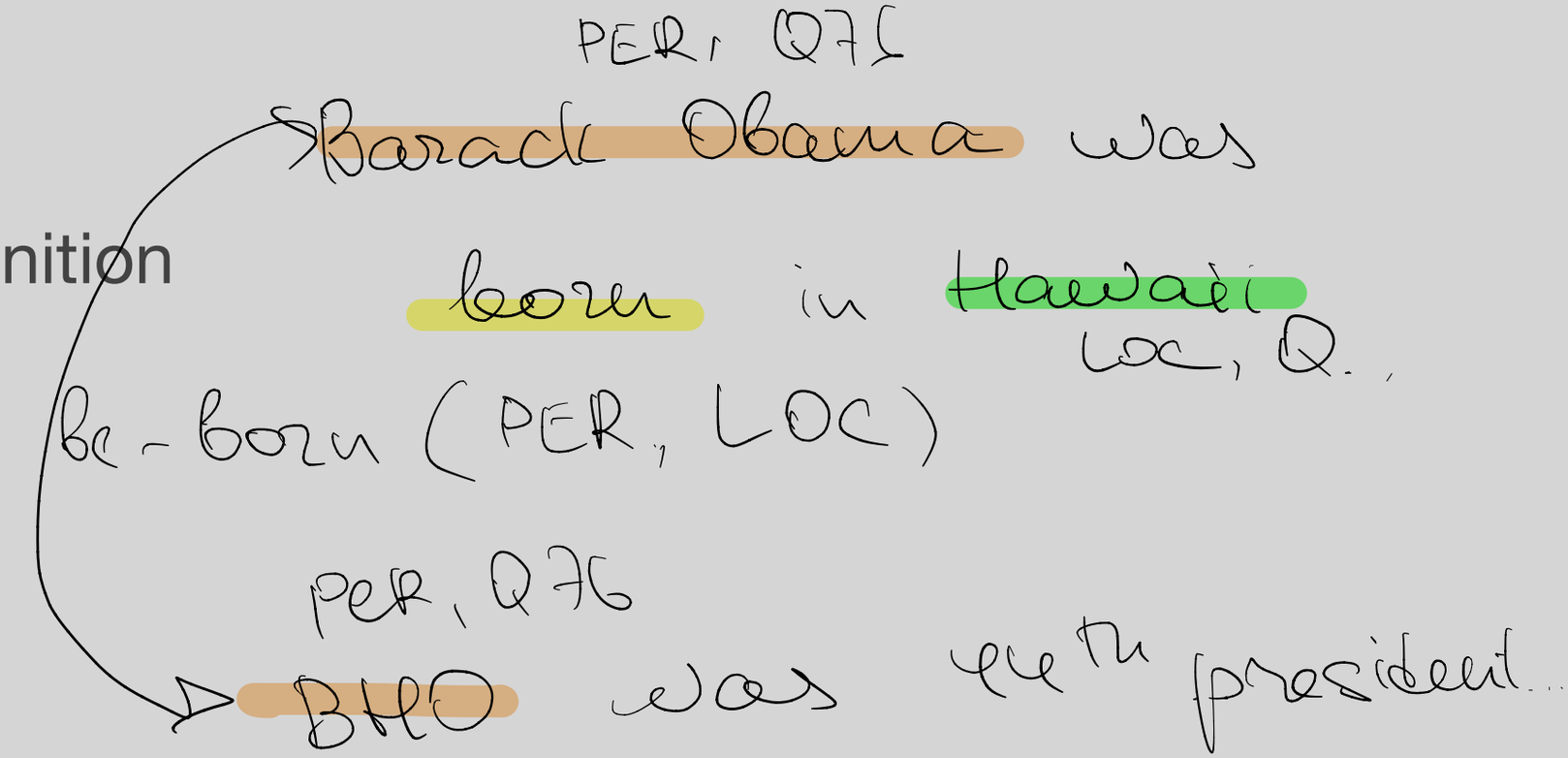
date loc

sequence labelling

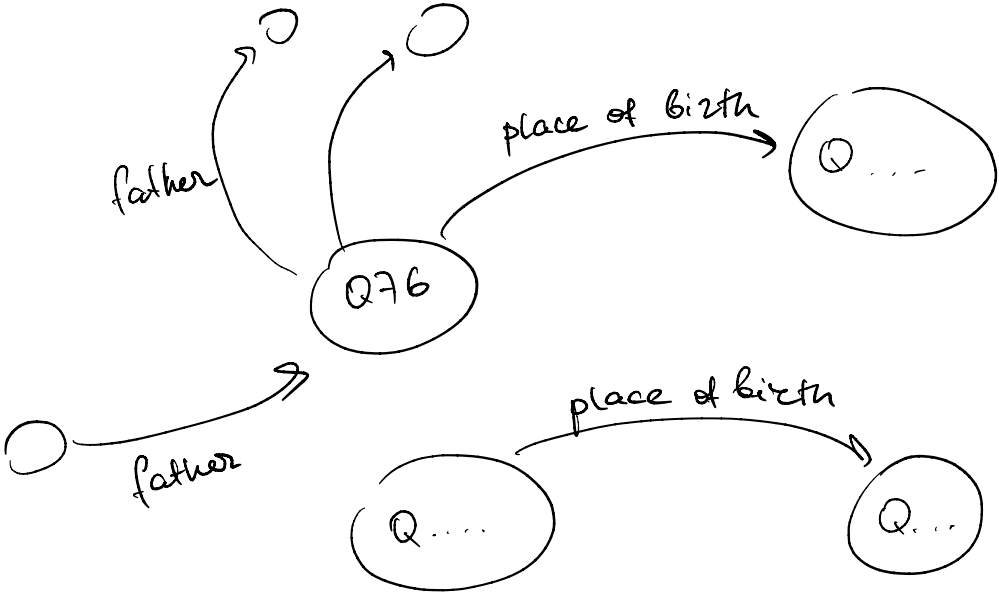
Introduction: landscape of IE tasks

Information extraction (IE)

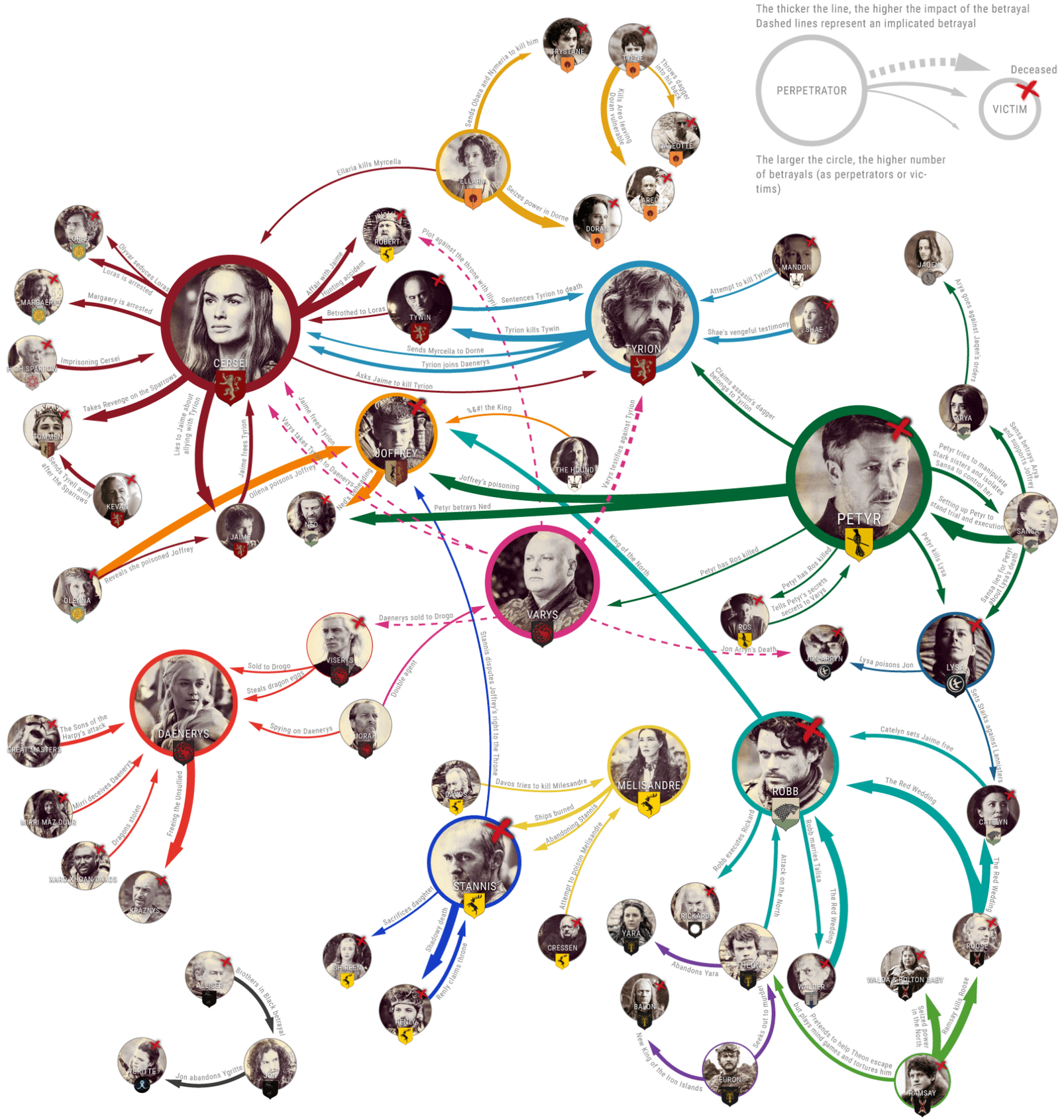
- IE turns the unstructured information expressed in natural language text into a structured representation (Jurafsky and Martin, 2009)
- IE tasks:
  - Named entity recognition
  - Relation extraction
  - Event extraction
  - Entity linking
  - Coreference resolution



Knowledge Graph / Base



Game of Thrones Web of Betrayals



Source: HBO, Game of Thrones. Updated 03/2019

Open IE

Open Information Extraction

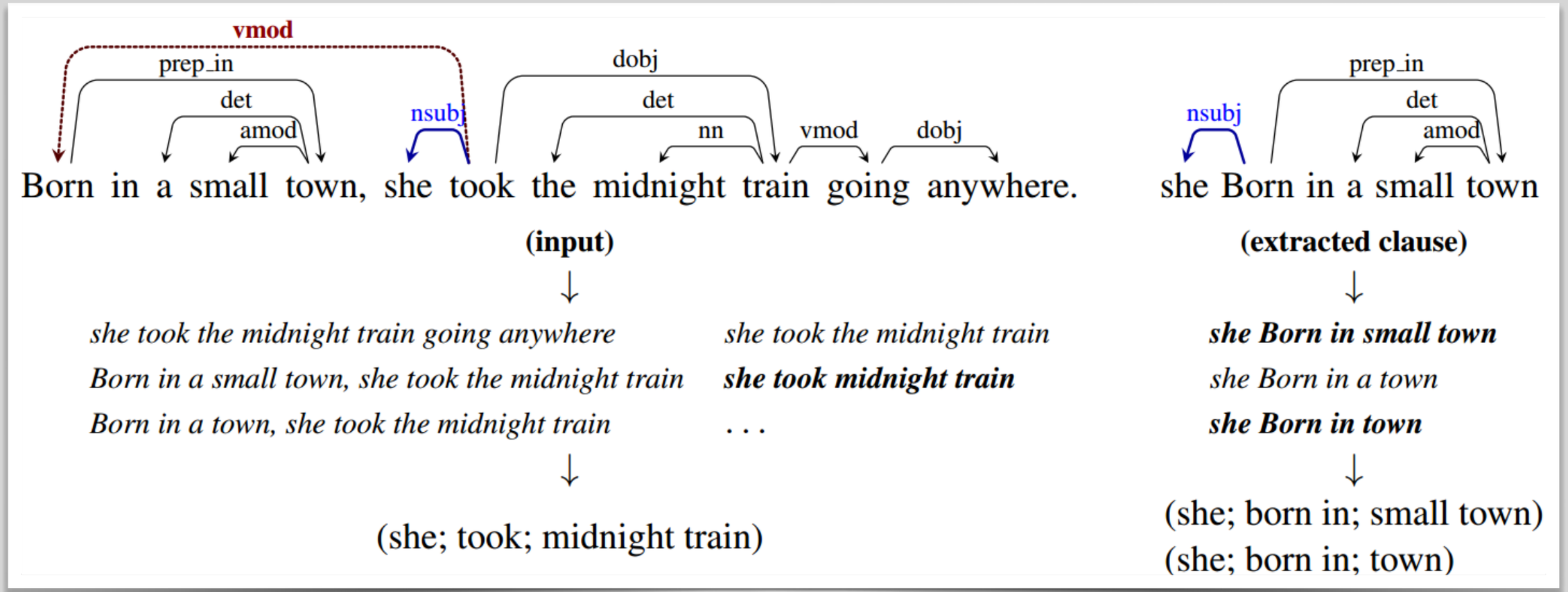
- Open information extraction (open IE) refers to the extraction of relation tuples, typically binary relations, from plain text, such as (Mark Zuckerberg; founded; Facebook).

- The central difference from other information extraction is that the schema for these relations does not need to be specified in advance; typically the relation name is just the text linking two arguments.

- For example, *Barack Obama was born in Hawaii* would create a triple (Barack Obama; was born in; Hawaii), corresponding to the open domain relation `was-born-in(Barack-Obama, Hawaii)`.

https://nlp.stanford.edu/software/openie.html

Stanford OpenIE



https://nlp.stanford.edu/software/openie.html

Stanford OpenIE

- The system first splits each sentence into a set of entailed clauses.
- Each clause is then maximally shortened, producing a set of entailed shorter sentence fragments.
- These fragments are then segmented into OpenIE triples, and output by the system.
- Python3 wrapper for Stanford OpenIE: git.

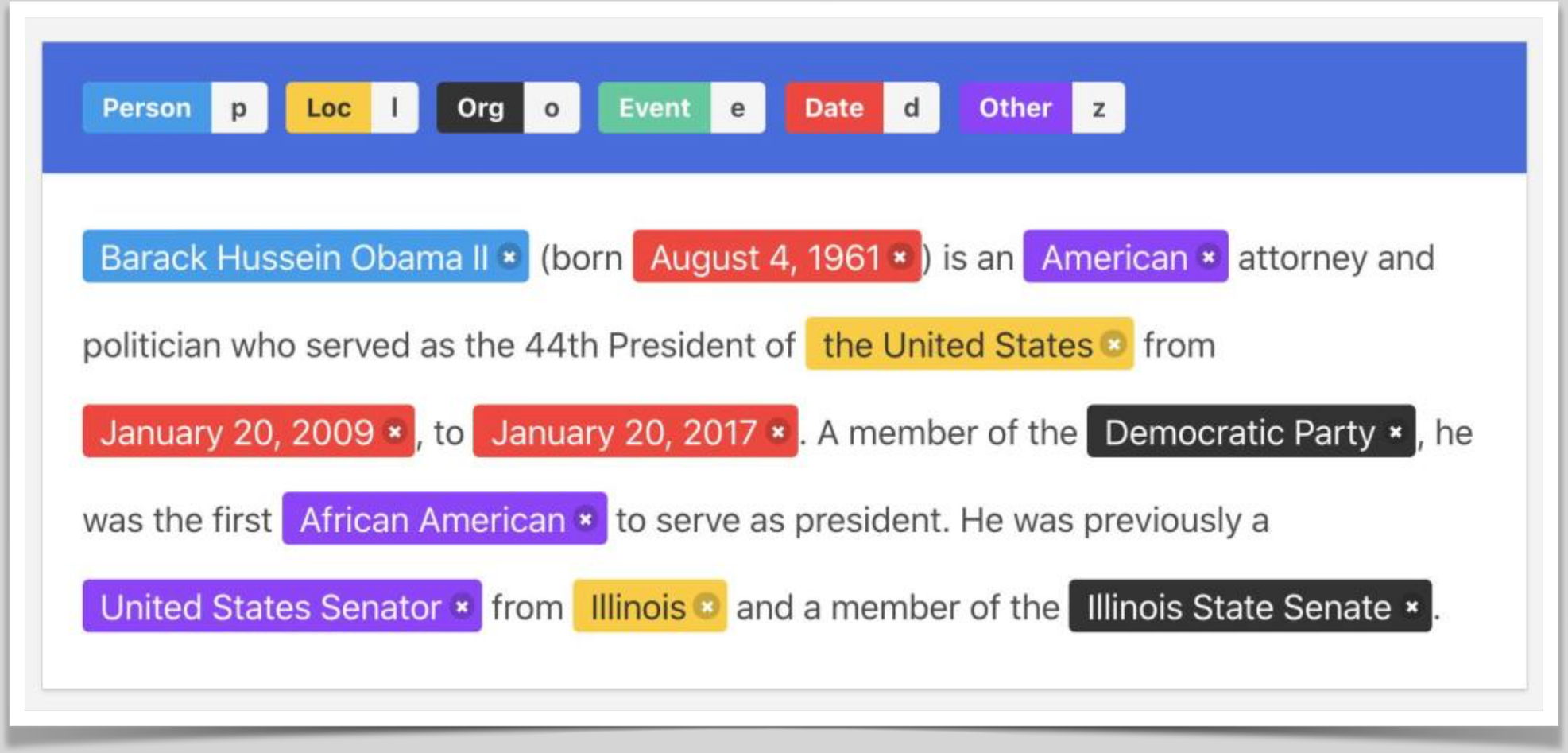
https://nlp.stanford.edu/software/openie.html

Open IE 5.1

- An Open IE system runs over sentences and creates extractions that represent relations in text.
- Open IE 5.1 is the successor to Open IE 4.x.
- Open IE 5.1 improves extractions from noun relations, numerical sentences and conjunctive sentences.
- Open IE standalone: git.

Named entity recognition

Named entity recognition



Weapon
-Transportation

Image source

Off-the-shelf tools

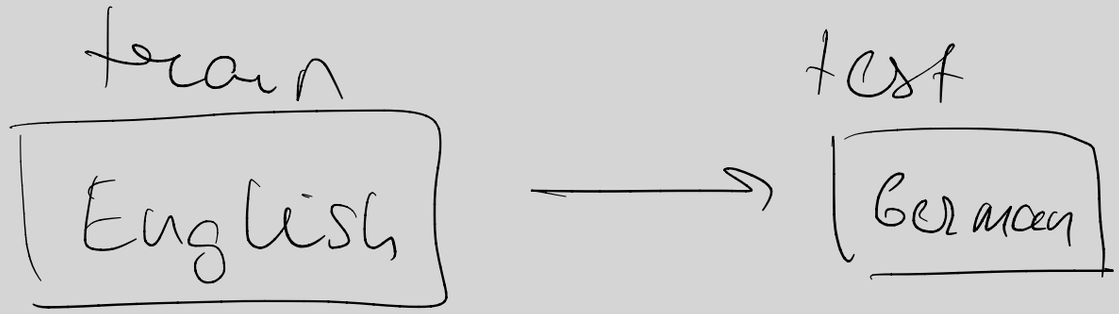
- Natasha (Ru)
- Flair (En, De)
- Stanza (Many languages)
- spaCy (Many languages)
- Pre-trained LMs can be found in HuggingFace Models.

Inter Model Agreement

main	spacy_small	spacy_big	flair	stanford
spacy_small	100% 2691 /2691	70% 1881 /2691	13% 353 /2691	48% 1297 /2691
spacy_big	71% 1881 /2652	100% 2652 /2652	14% 372 /2652	48% 1261 /2652
flair	25% 353 /1405	26% 372 /1405	100% 1405 /1405	32% 443 /1405
stanford	49% 1297 /2660	47% 1261 /2660	17% 443 /2660	100% 2660 /2660

Image source

NER datasets



- Mono-lingual datasets: English, Chinese, Arabic, German, Russian and many others
- Multi-lingual datasets: OntoNotes, BSNLP, WikiNER, ACE
- Domains: news, wiki, legal, medical, Twitter, literal texts, historical texts
- Tags: PER, LOC, ORG and many others

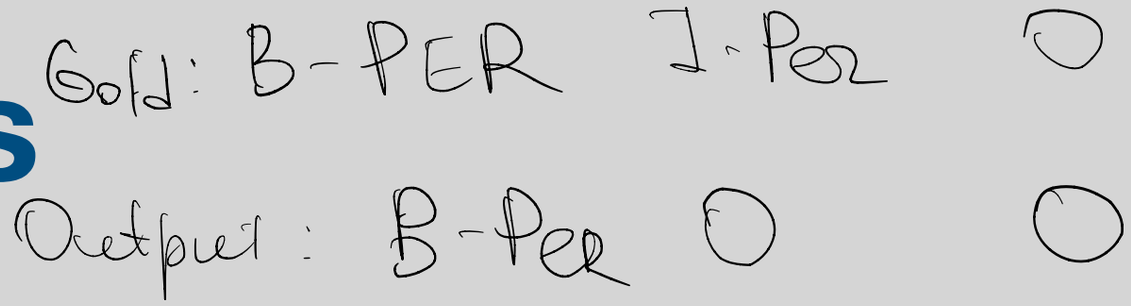
NER methods

Three approaches to monolingual NER

- Sequence labelling (oldie but goodie)
- Machine reading comprehension for NER (~2020)
- Template-based and prompt learning methods (~2021)

NER methods

Measures



- Token-levels vs span-level measures
- Usually NER models are compared according to micro-averaged F1.
- seqeval is a Python framework for sequence labelling evaluation.
- HuggingFace datasets support seqeval.

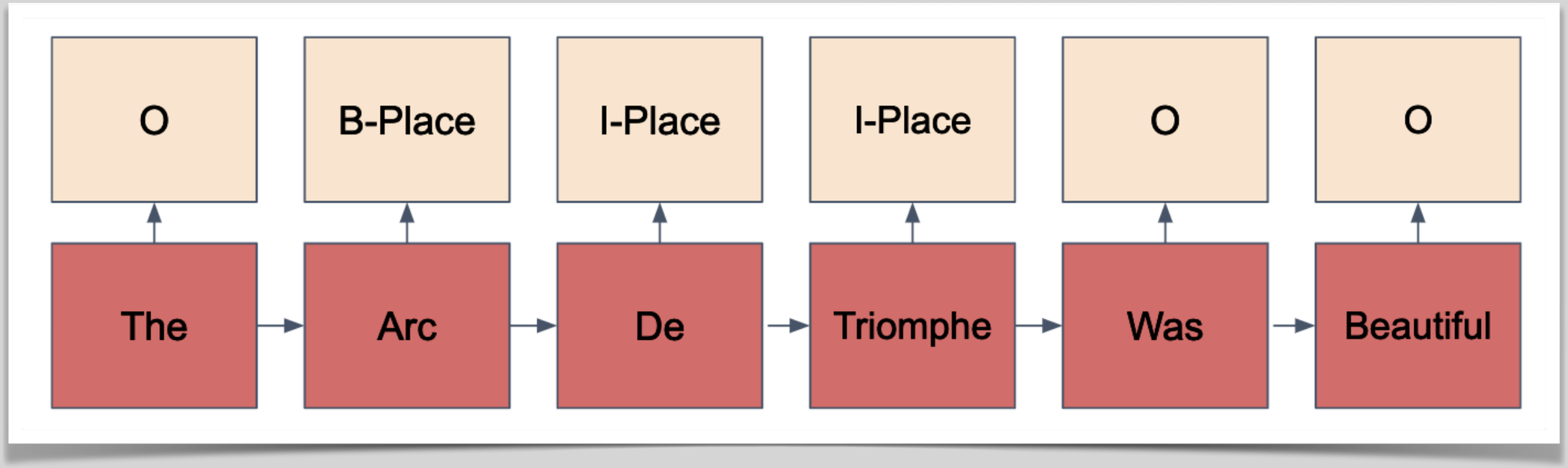
NER methods

I. Sequence labelling

O = outside
B = begin
I = inside

y:

x:



- Begin-Inside-Outside (Ends-Single) encodings
- Make a prediction for each token

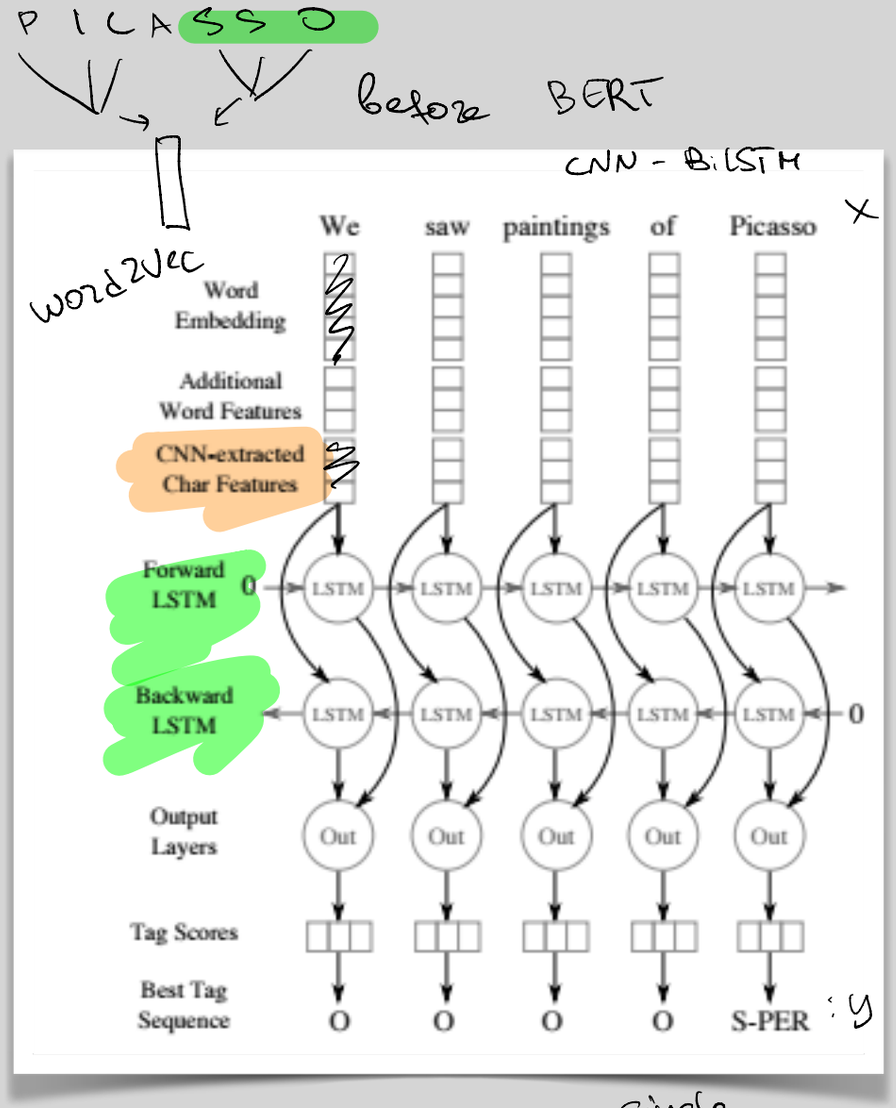
3 types: BIO
tags?
7

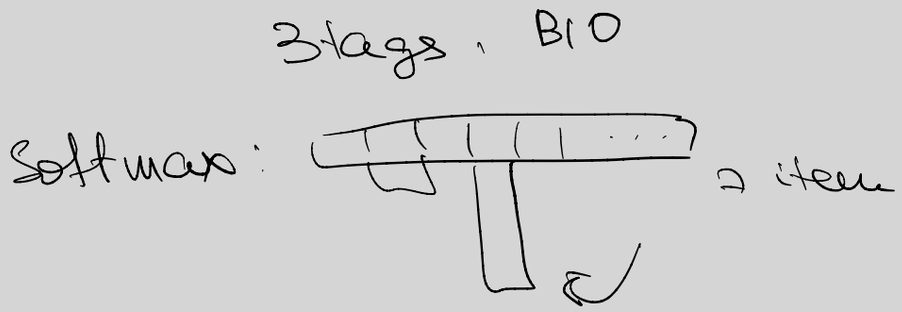
Image source

I. Sequence labelling

Before BERT: CNN-BiLSTM-CRF

- Character level: CNN
- Word level: word embeddings + feature embeddings
- Contextualised layer: BiLSTM
- Outputs
  - softmax scores
  - CRF to score outputs



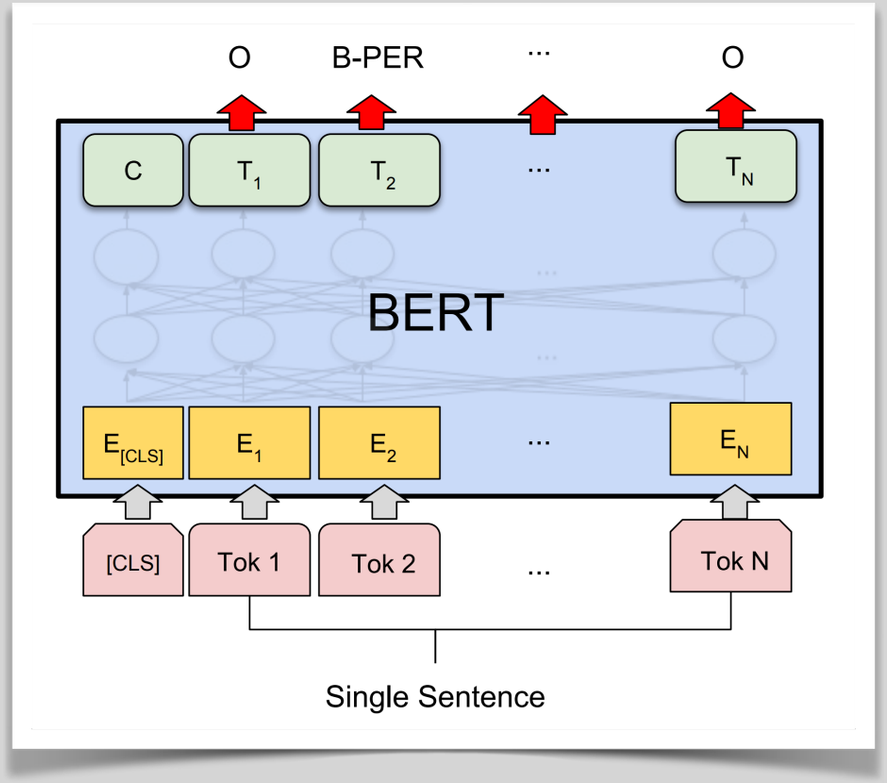


Ma, X., & Hovy, E. (2016, August). End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1064-1074).

I. Sequence labelling

BERToids

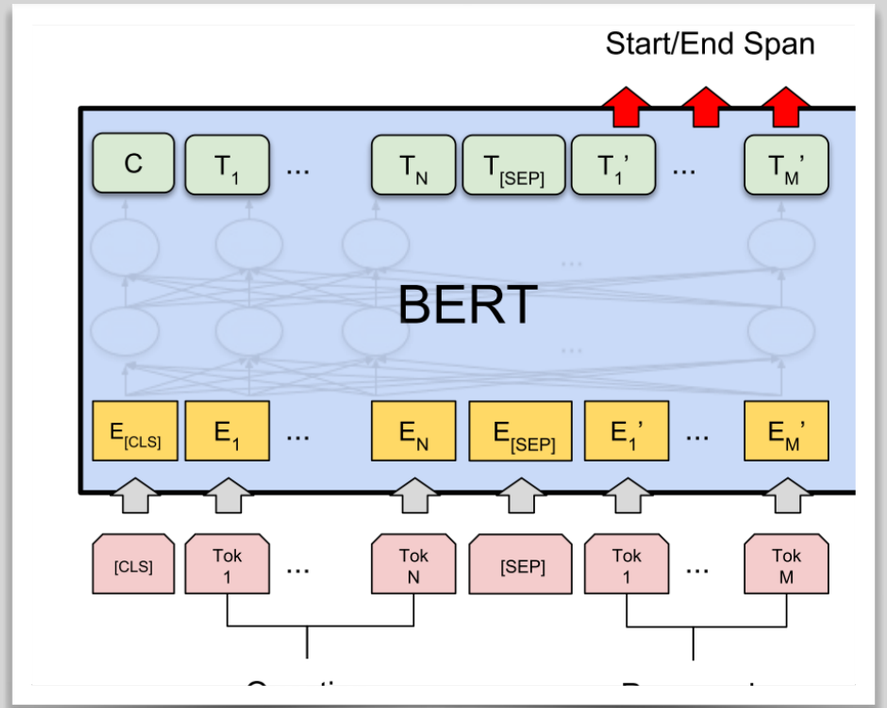
- Input: word sequence
- Output: softmax scores
- CRF layer can be added on top
- Improvements:
  - Use left and right context (FLERT)
  - Entity masking (ERNIE)
  - Entity-aware attention mechanism (LUKE)



Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies(pp. 4171-4186).

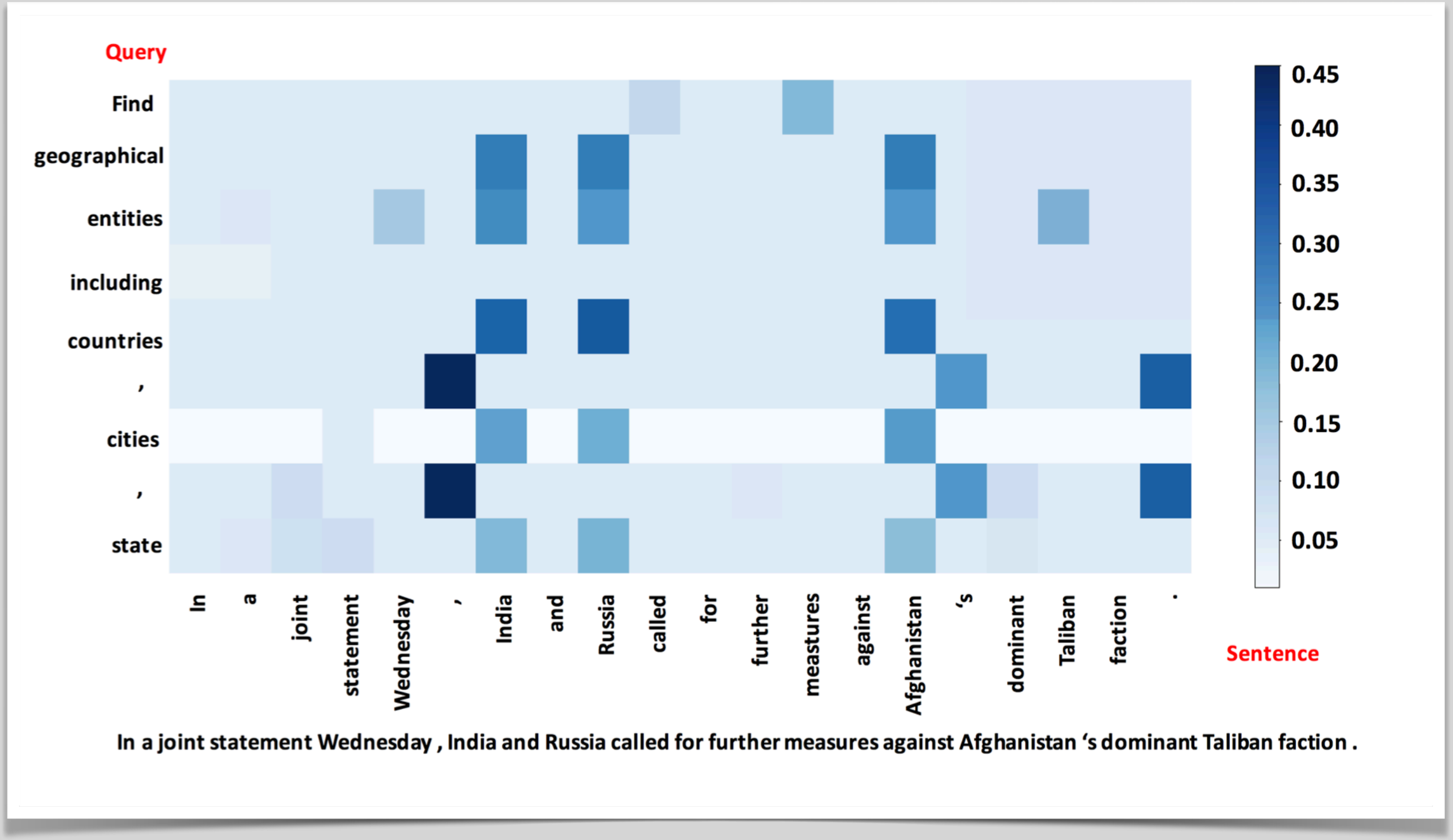
II. Machine reading comprehension

- NER is formulated as a MRC (SQuAD) task
- To extract entities with the Person label extract answer spans to the question “which person is mentioned in the text”
- The MRC framework enables:
  - Extraction of overlapping entities
  - Extraction of nested entities
  - Zero-shot NER



Li, X., Feng, J., Meng, Y., Han, Q., Wu, F., & Li, J. (2020, July). A Unified MRC Framework for Named Entity Recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5849-5859).

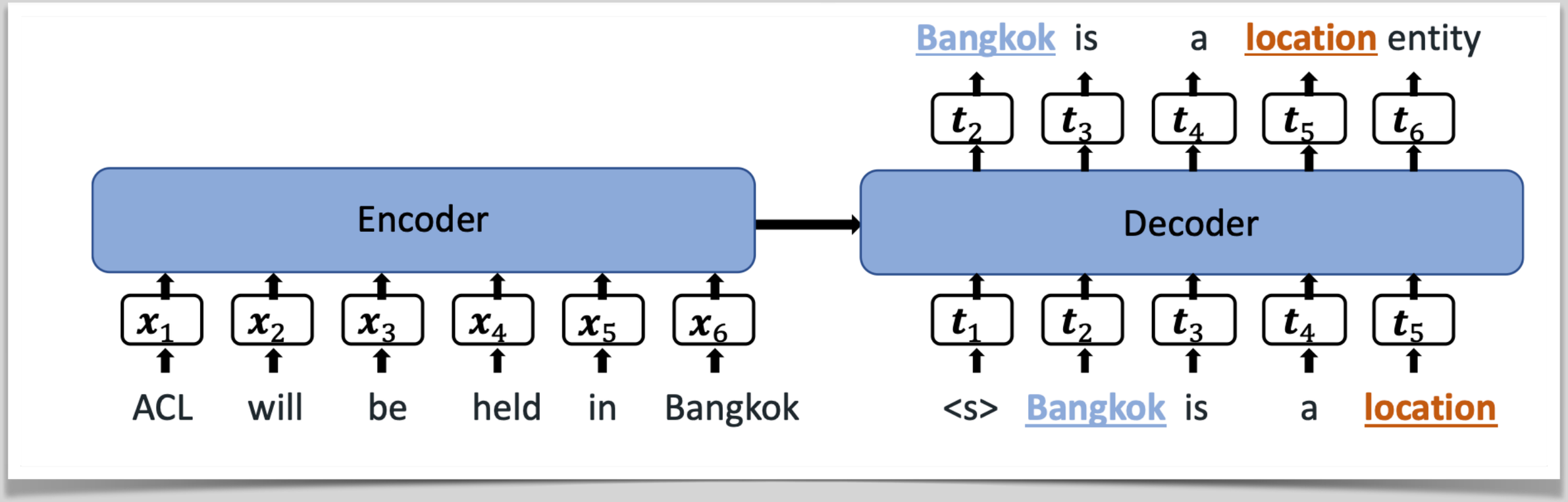
II. Machine reading comprehension



Li, X., Feng, J., Meng, Y., Han, Q., Wu, F., & Li, J. (2020, July). A Unified MRC Framework for Named Entity Recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5849-5859).

III. Template-Based NER

Training

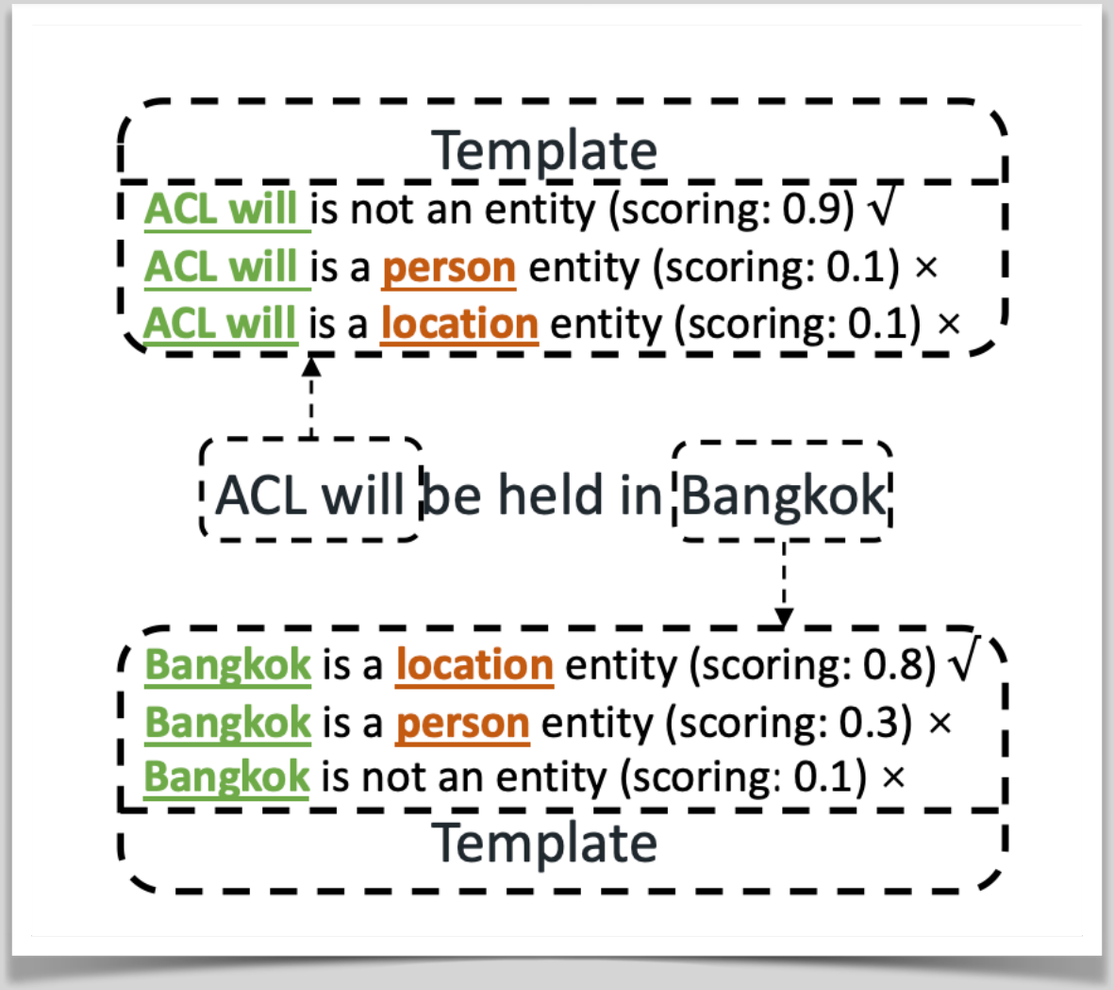


Template used: “x is a y”, x is an entity, y is an entity type

Cui, L., Wu, Y., Liu, J., Yang, S., & Zhang, Y. (2021). Template-Based Named Entity Recognition Using BART. *arXiv preprint arXiv:2106.01760*

III. Template-Based NER

Inference



Template used: “x is a y”, x is an entity, y is an entity type

Cui, L., Wu, Y., Liu, J., Yang, S., & Zhang, Y. (2021). Template-Based Named Entity Recognition Using BART. *arXiv preprint arXiv:2106.01760*

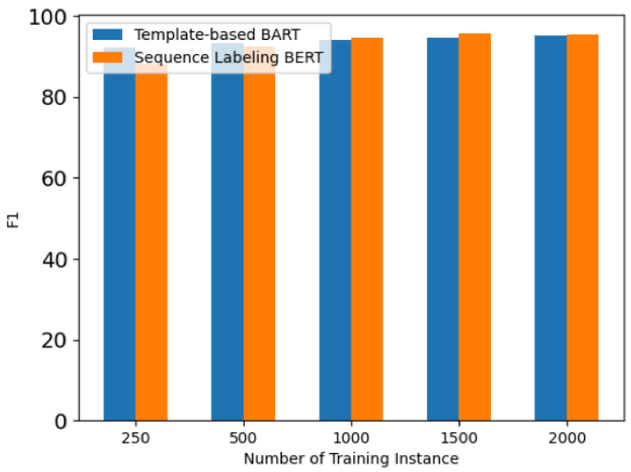
III. Template-Based NER

- The template is created manually.
- A non-entity template is used, too: “x is not a named entity”.
- At inference time, all n-grams ($n \in [1,8]$) are considered.
- Loop all over entity labels, score each template with log likelihood.
- Assign the entity type with the highest score.

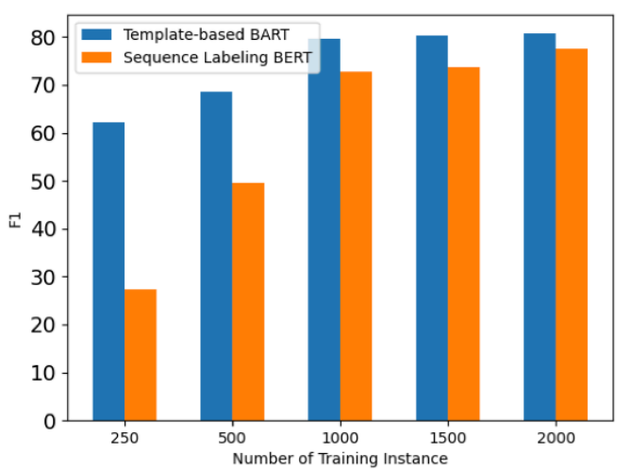
Cui, L., Wu, Y., Liu, J., Yang, S., & Zhang, Y. (2021). Template-Based Named Entity Recognition Using BART. *arXiv preprint arXiv:2106.01760*

III. Template-Based NER

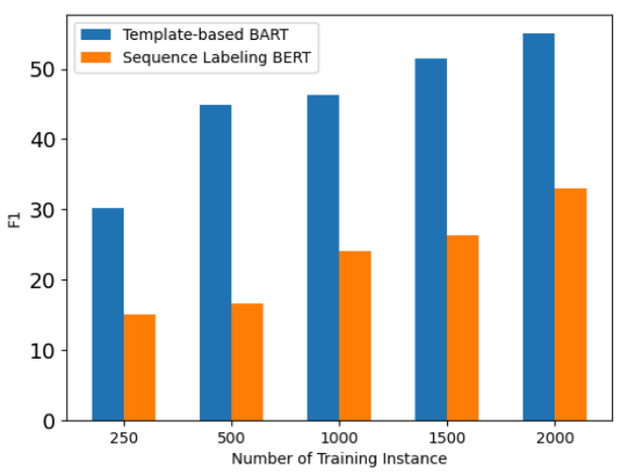
works well for low frequency entity types



(a) High Frequency.



(b) Mid Frequency.



(c) Low Frequency.

Cui, L., Wu, Y., Liu, J., Yang, S., & Zhang, Y. (2021). Template-Based Named Entity Recognition Using BART. *arXiv preprint arXiv:2106.01760*

Relation extraction

Relation extraction

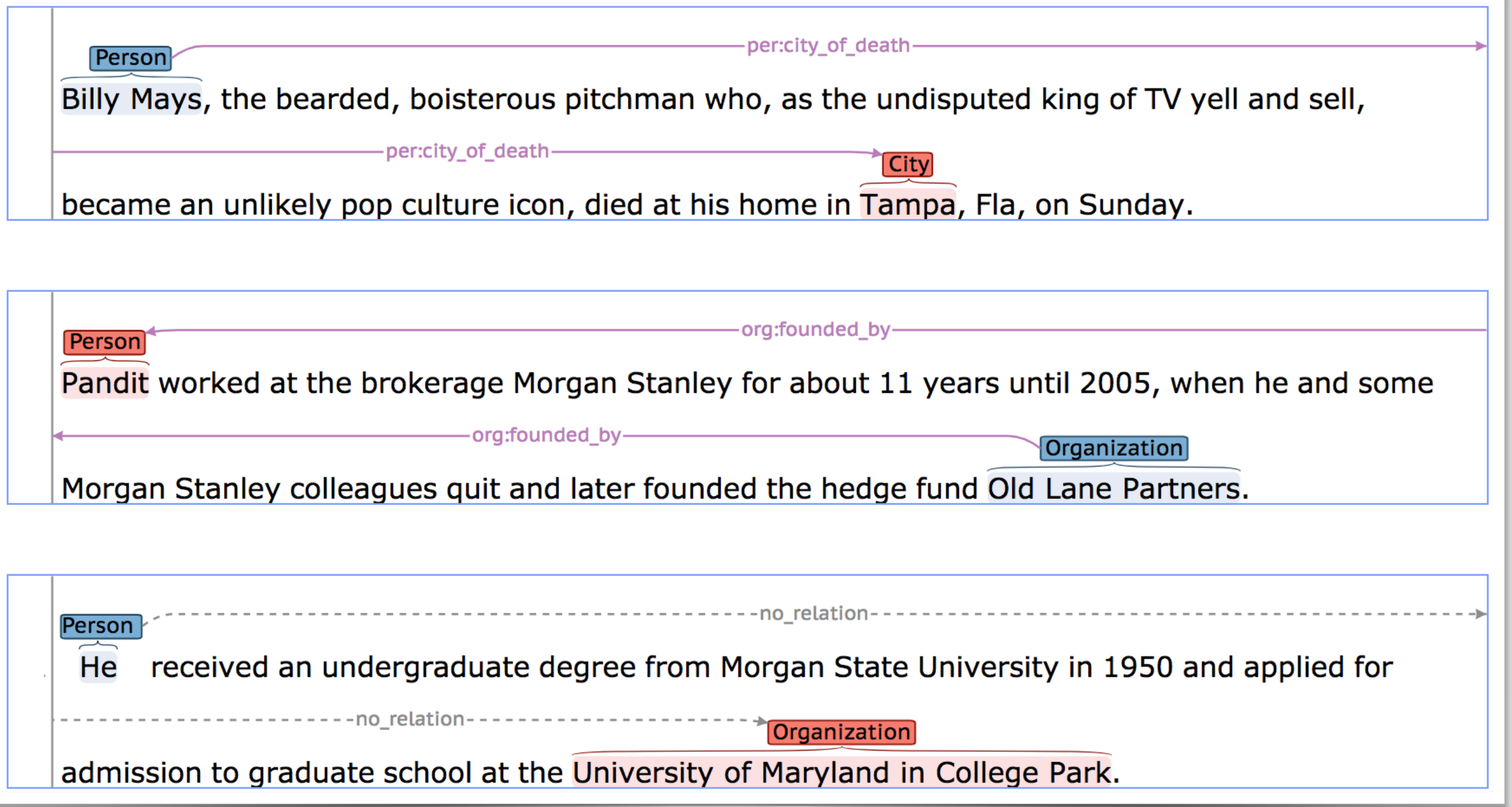


Image source

RE datasets

- TACRED (news, social media, English)
  - 23 entity types, 41 relation types
- NEREL (news, Russian, under development)
  - 29 entity types, 43 relation types
- Medical and legal datasets

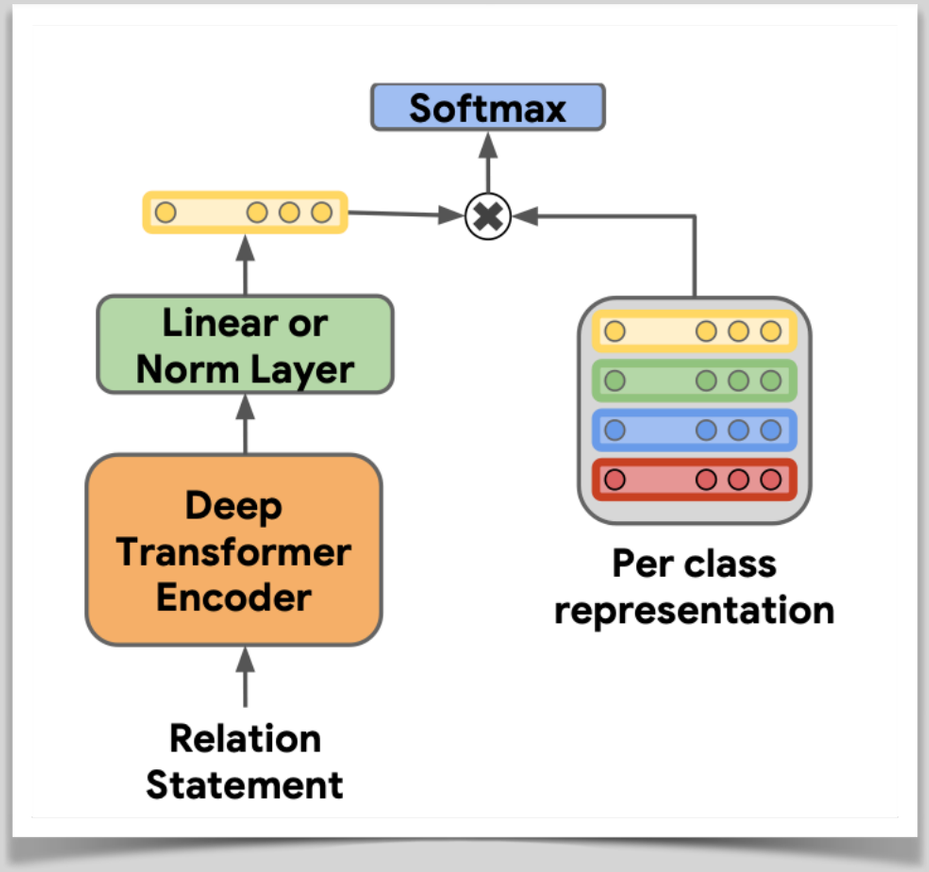
Relation extraction

Methods and metrics

- Methods
  - Relation extractions is often treated as a classification task.
  - Entailment-based approaches work well for text classification tasks!
  - Joint NER and RE is a promising research direction.
- Performance
  - Per class precision, recall, F1
  - Micro-averaged F1

Relation extraction

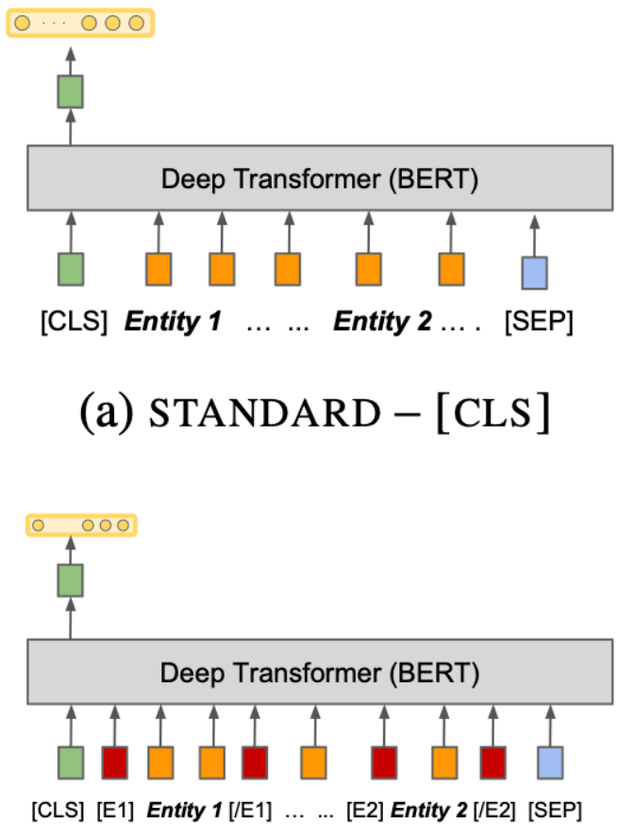
I. Relation learning. Loss function



Soares, L. B., FitzGerald, N., Ling, J., & Kwiatkowski, T. (2019, July). Matching the Blanks: Distributional Similarity for Relation Learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 2895-2905).

Relation extraction

I. Relation learning. Relation representations



Soares, L. B., FitzGerald, N., Ling, J., & Kwiatkowski, T. (2019, July). Matching the Blanks: Distributional Similarity for Relation Learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 2895-2905).

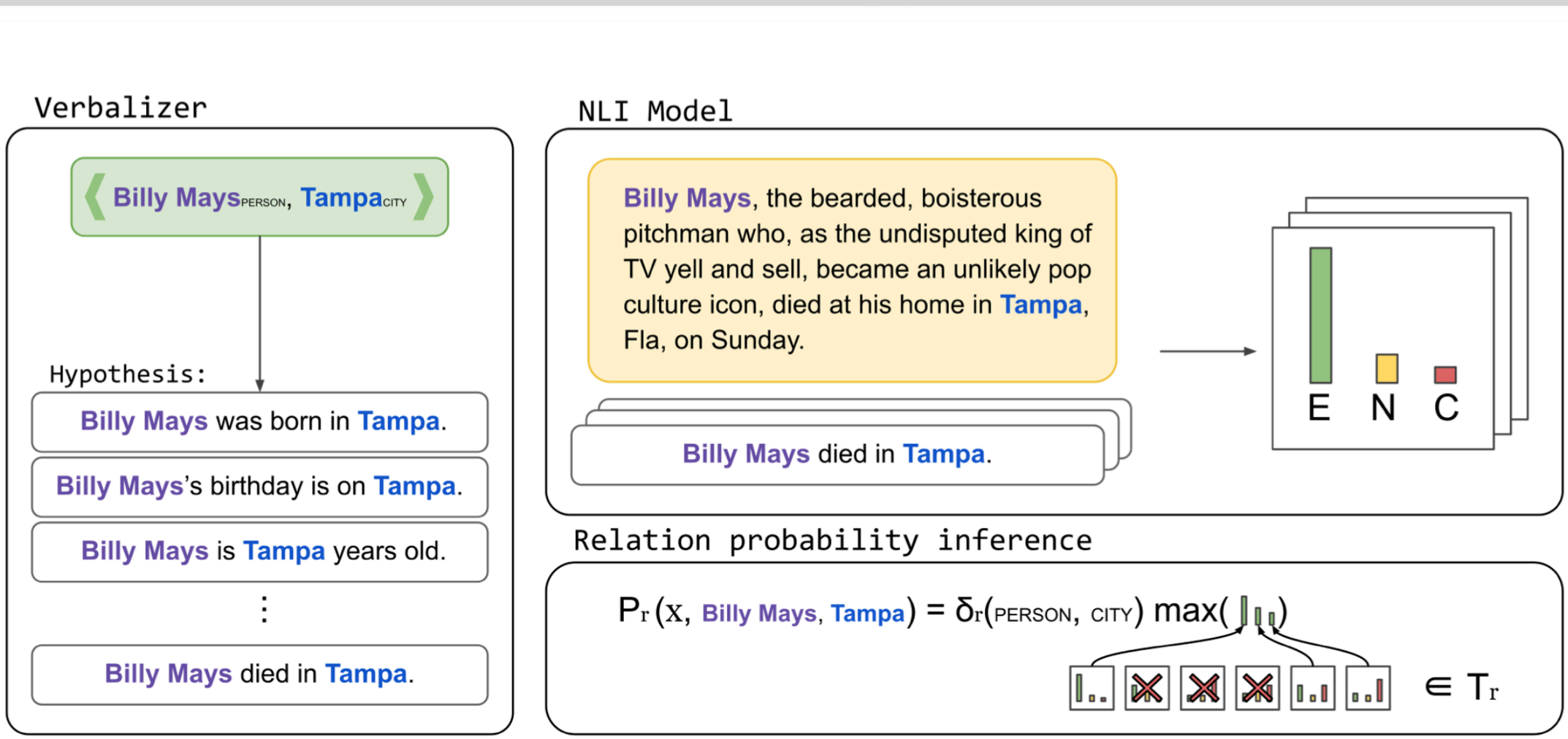
Relation extraction

I. Relation learning. Relation representations

- The model classifies over a predefined dictionary of relation types.
- Different variants of architectures for extracting relation representations from deep Transformers network are considered.

Relation extraction

II. Entailment for RE



Sainz, O., de Lacalle, O. L., Labaka, G., Barrena, A., & Agirre, E. (2021). Label Verbalization and Entailment for Effective Zero-and Few-Shot Relation Extraction. *arXiv preprint arXiv:2109.03659*.

Relation extraction

II. Entailment for RE

- RE as an entailment task: given the input text containing the two entity mentions as the premise and the verbalized description of a relation as hypothesis, the task is to infer if the premise entails the hypothesis according to the NLI model

- The hypotheses are automatically generated using a set of templates

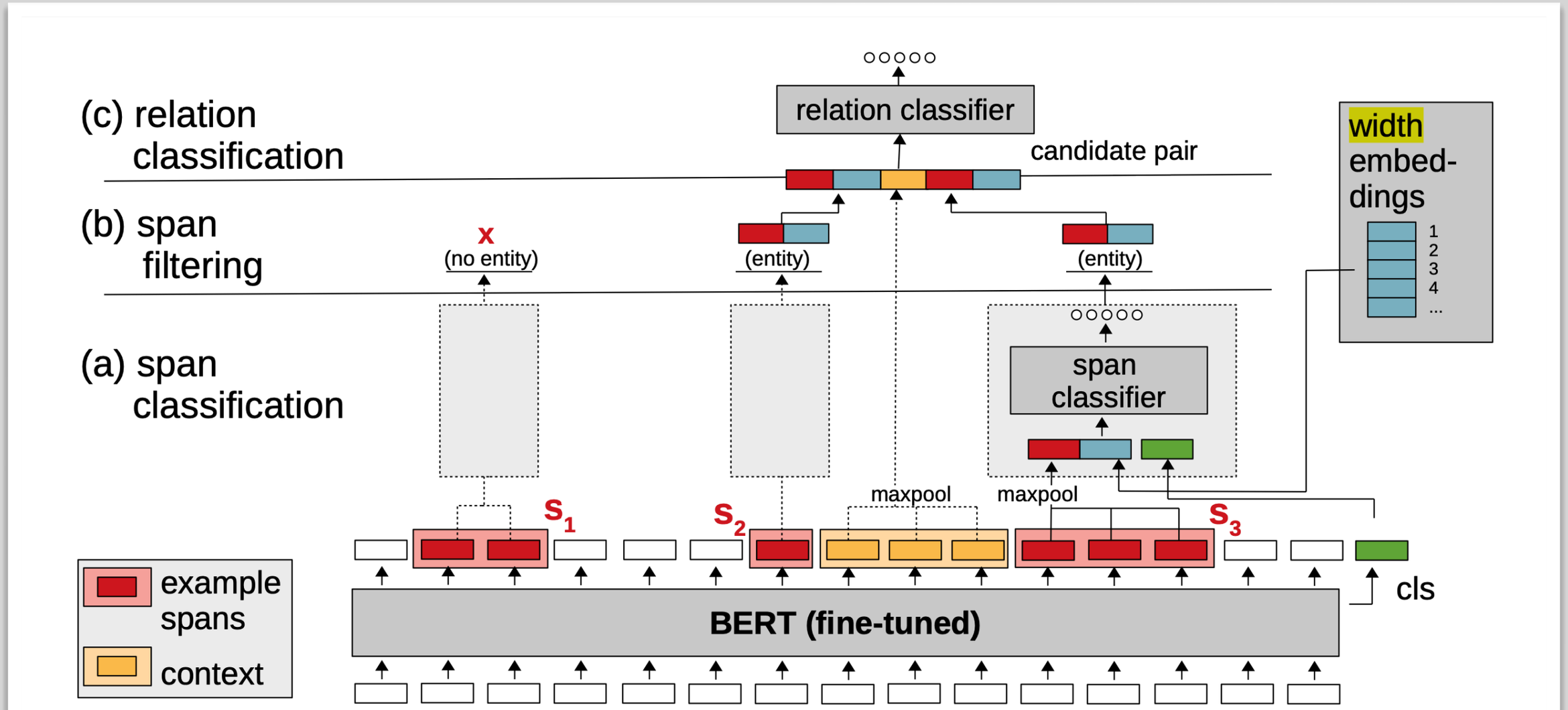
- The function δ checks entity coherence between the template and the current relation label

- Choose the relation R with the highest entailment probability

Sainz, O., de Lacalle, O. L., Labaka, G., Barrena, A., & Agirre, E. (2021). Label Verbalization and Entailment for Effective Zero-and Few-Shot Relation Extraction. *arXiv preprint arXiv:2109.03659*.

Relation extraction

III. SpERT: joint NER + RE



Eberts, M., & Ulges, A. (2019). Span-based joint entity and relation extraction with transformer pre-training. *arXiv preprint arXiv:1909.07755*.

Relation extraction

III. SpERT: joint NER + RE

- SpERT first passes a token sequence through BERT.
- All spans within the sentence are classified into entity types.
- Spans classified as non-entites are filtered.
- All pairs of remaining entities are combined with their context (the span between the entities) and classified into relations.

Eberts, M., & Ulges, A. (2019). Span-based joint entity and relation extraction with transformer pre-training. *arXiv preprint arXiv:1909.07755*.

Entity linking

Entity linking

is the task to link entity mentions in text with their corresponding entities in a knowledge base

We know ‘Sebastian Thrun’ is a person
but do we know which person exactly?

When Sebastian Thrun PERSON started at Google ORG in 2007 DATE , few people outside of the company took him seriously. “I can tell you very senior CEOs of major American NORP car companies would shake my hand and turn away because I wasn’t worth talking to,” said Thrun PERSON , now the co-founder and CEO of online higher education startup Udacity, in an interview with Recode ORG earlier this week DATE .

A little less than a decade later DATE , dozens of self-driving startups have cropped up while automakers around the world clamor, wallet in hand, to secure their place in the fast-moving world of fully automated transportation.

About: Sebastian Thrun

An Entity of Type : scientist, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org

Sebastian Thrun (born May 14, 1967) is an innovator, entrepreneur educator, and computer scientist from Germany. He was CEO and cofounder of Udacity. Before that, he was a Google VP and Fellow, and a Professor of Computer Science at Stanford University. At Google, he founded Google X. He is currently also an Adjunct Professor at Stanford University and at Georgia Tech.

Property	Value
dbo:abstract	• Sebastian Thrun (born May 14, 1967) is an innovator, entrepreneur educator, and computer scientist from Germany. He was CEO and cofounder of Udacity. Before that, he was a Google VP and Fellow, and a Professor of Computer Science at Stanford University. At Google, he founded Google X. He is currently also an Adjunct Professor at Stanford University and at Georgia Tech. Thrun led development of the robotic vehicle Stanley which won the 2005 DARPA Grand Challenge, and which has since been placed on exhibit in the Smithsonian Institution's National Museum of American History. His team also developed a vehicle called Junior, which placed second at the DARPA Grand Challenge (2007). Thrun led the development of the Google self-driving car. Thrun is also known for his work on probabilistic algorithms for robotics with applications including robotic mapping. In recognition of his contributions, and at age 39, Thrun was elected into the National Academy of Engineering and also into the Academy of Sciences Leopoldina in 2007. In 2011, Thrun received the Max-Planck-Research Award, and the inaugural AAAI Ed Feigenbaum Prize.Fast Company selected Thrun as the fifth most creative person in the business world. The Guardian recognized Thrun as one of 20 "fighters for internet freedom". (en)

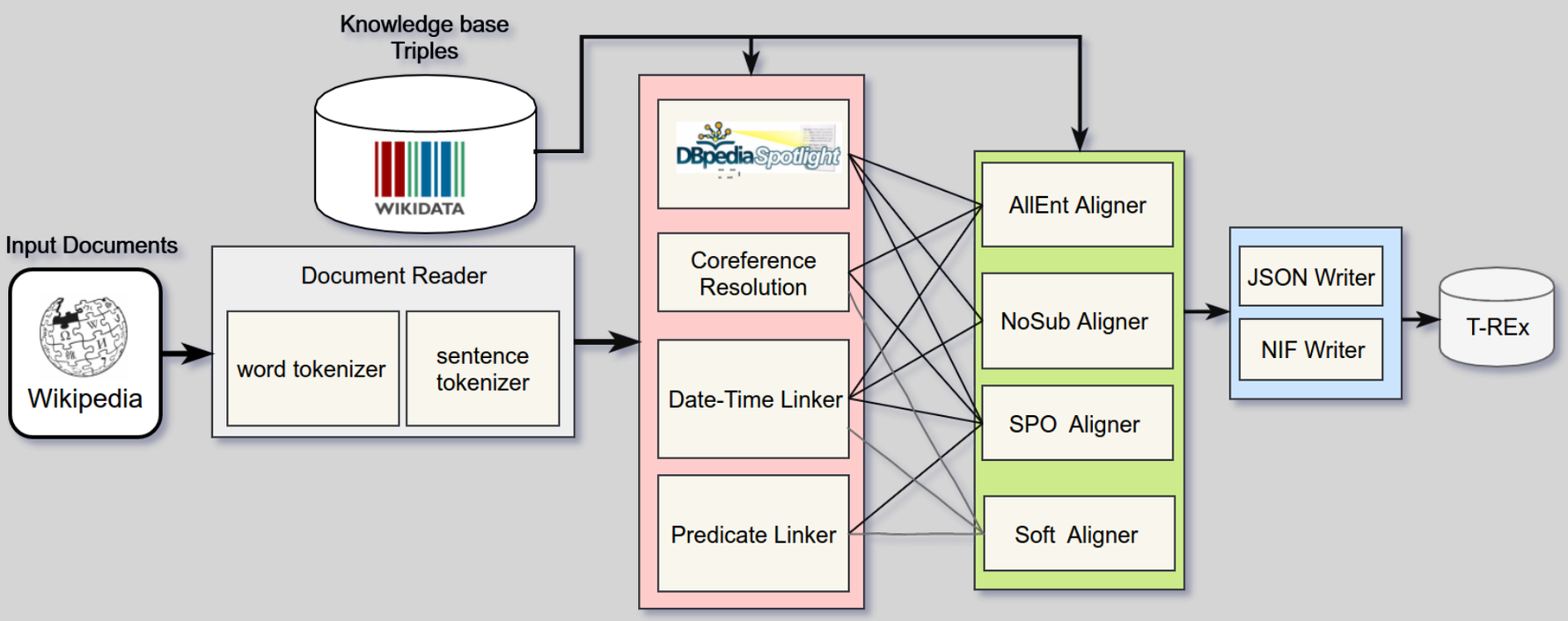
http://dbpedia.org/page/Sebastian_Thrun

- *End-to-End* processes a piece of text to extract the entities and then disambiguate these extracted entities to the correct entry in a given knowledge base.
- *Disambiguation-Only* takes gold standard named entities as input and only disambiguates them to the correct entry in a given knowledge base.

Image source

EL datasets

- CoNLL-AIDA maps entities from CONLL2003 NER dataset to Wikipedia
- T-REx utilises WikiData.
- Other domains.

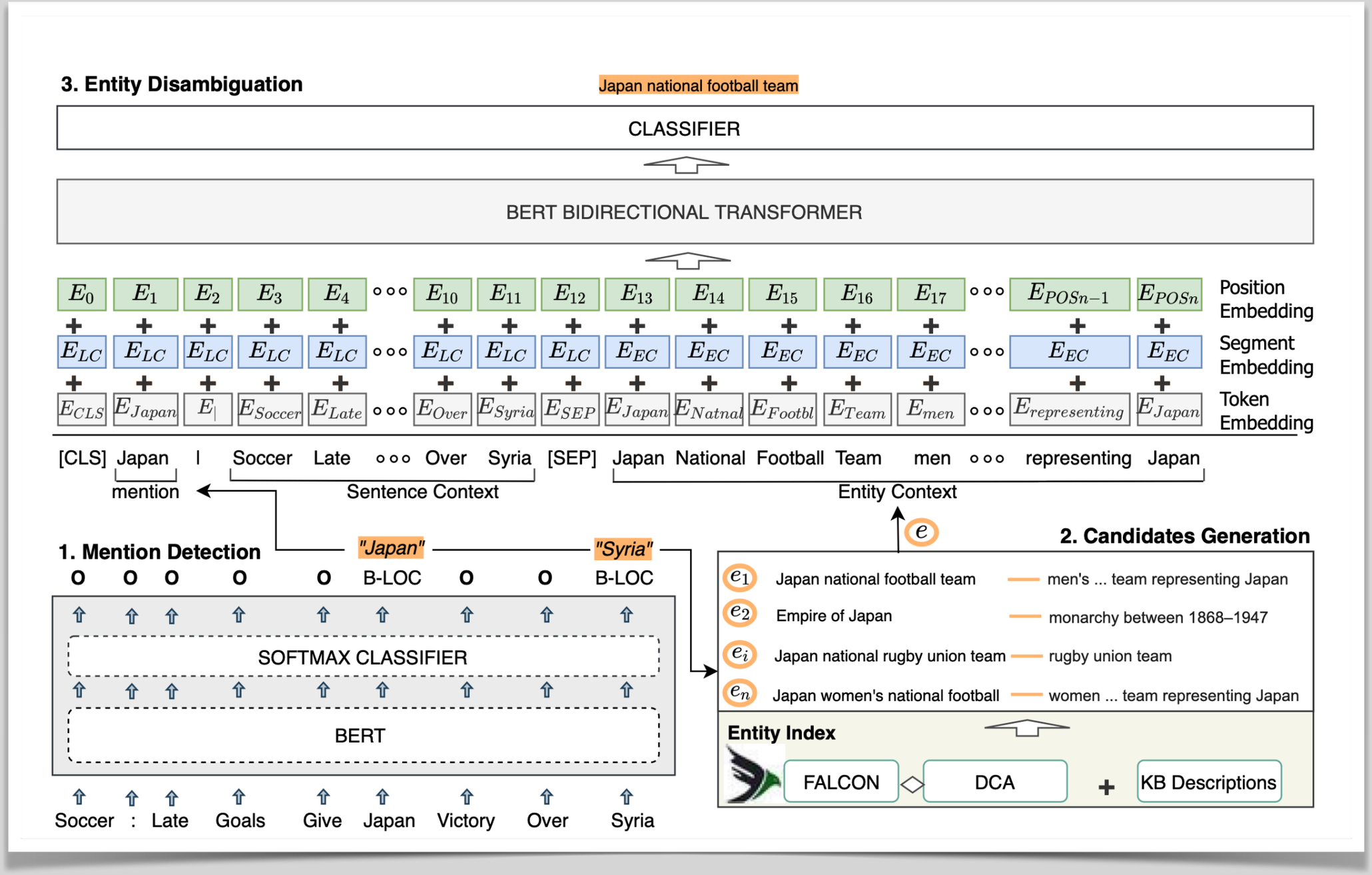


EL subtasks and metrics

- Mention detection
  - NER metrics
- Candidate generation
  - IR metrics
- Entity disambiguation
  - micro-F1

EL methods

CHOLAN: a modular approach for neural EL



Ravi, M. P. K., Singh, K., Mulang, I. O., Shekarpour, S., Hoffart, J., & Lehmann, J. (2021, January). CHOLAN: A Modular Approach for Neural Entity Linking on Wikipedia and Wikidata. In *EACL*.

EL methods

CHOLAN: a modular approach for neural EL

- Mention detection: predict entity mentions in BIO labelling
- Candidate generation: reuse candidate lists, retrieve candidates with BM25
- Entity disambiguation: use cross-encoder
  - Input: `[CLS]` mention `|` sentence context `[SEP]` entity context
  - Binary classification task: does mention relate to the entity?

Ravi, M. P. K., Singh, K., Mulang, I. O., Shekarpour, S., Hoffart, J., & Lehmann, J. (2021, January). CHOLAN: A Modular Approach for Neural Entity Linking on Wikipedia and Wikidata. In *EACL*.

Other topics

Other IE tasks

- **Coreference resolution** is the task of finding all expressions that refer to the same entity in a text.

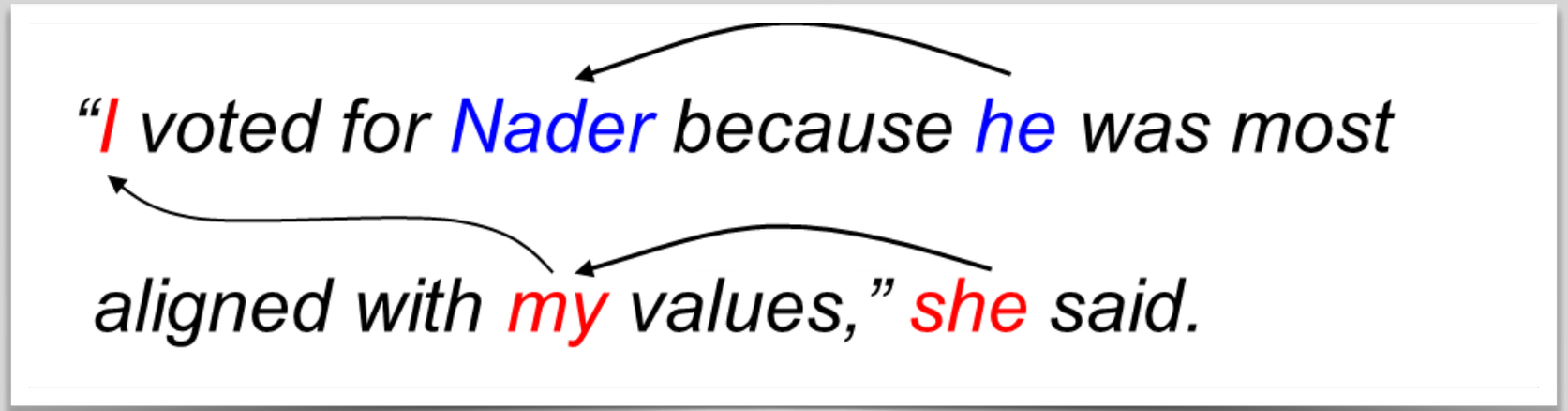


Image source

Other IE tasks

- **Semantic parsing** is the task of translating natural language into a formal meaning representation on which a machine can act.

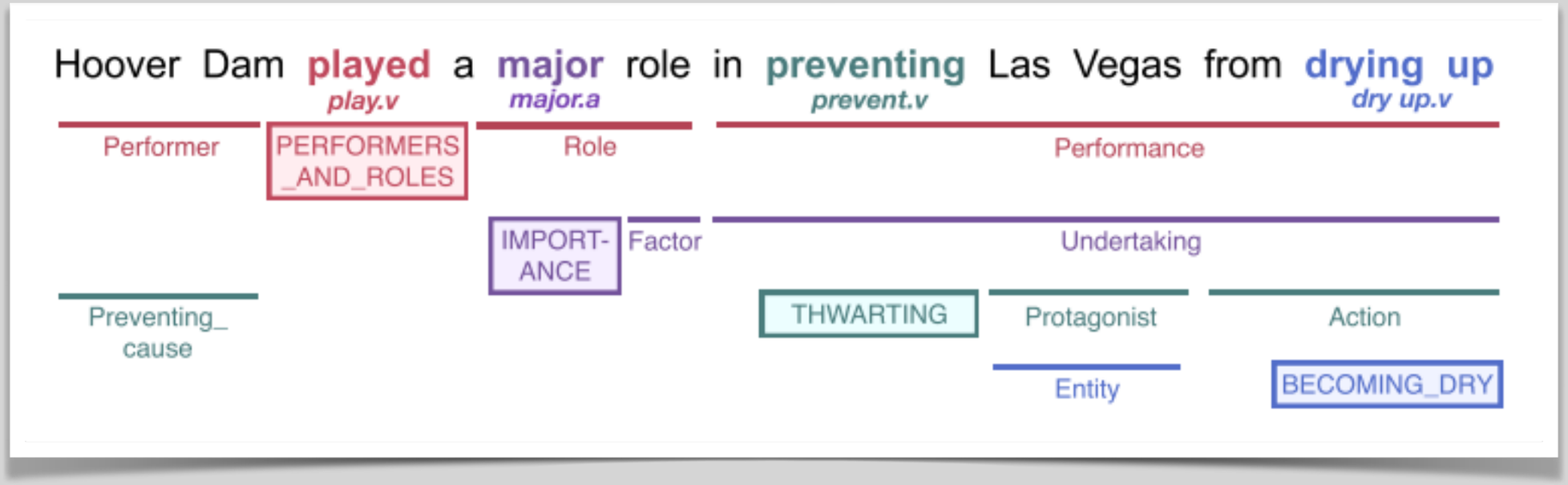


Image source

Conclusion

- Older approaches to IE utilise common setups:
  - Sequence labelling for NER
  - Sequence classification for RE
- Novel approaches formulate IE tasks as QA and NLI problems
  - i.e. leverage upon **textual description** of the task
- Open problems:
  - Zero- and few-shot IE
  - Cross-domain and cross-lingual IE