A brief Introduction to Entity Linking

Tianxiang Sun (孙天祥)

2019/11/8 @ AWS Shanghai AI Lab

- What is entity linking?
  - Entity Linking (EL) aims to link entity mentions in texts to knowledge bases
  - Also called Named Entity Disambiguation (NED)
  - Non-trivial: entity mentions are usually ambiguous
  - A demo

Napoleon [Napoleon] was the emperor of the First French Empire. He was defeated at **Waterloo** [Battle of Waterloo] by **Wellington** [Arthur Wellesley, 1st Duke of Wellington] and **Blücher** [Gebhard Leberecht von Blücher]. He was banned to Saint **Helena** [Saint Helena], died of stomach cancer, and was buried at **Invalides** [Les Invalides].

- Formulation
  - Input: document $D = \{w_1, \ldots, w_n\}$ (+ $\{m_i\}$ if end-to-end)
  - Output: list of mention-entity pairs $\{(m_i, e_i)\}$
- A EL system typically performs two tasks:
  - NER / Mention Detection (MD)
    - Ent-to-End
    - Disambiguation-only
  - **Entity Disambiguation (ED)**
    - Candidate selection / generation (usually heuristics)
    - **Scoring (Ranking) candidates**
      - local & global

- Outline
  - Models
    - Modules
    - Neural models
    - Symbol-neural hybrid model
  - Related topics
    - Distant learning
    - Entity typing
  - Datasets, metrics, and platform

- Outline
    - Models
        - Modules
        - Neural models
        - Symbol-neural hybrid model
    - Related topics
        - Distant learning
        - Entity typing
    - Datasets, metrics, and platform

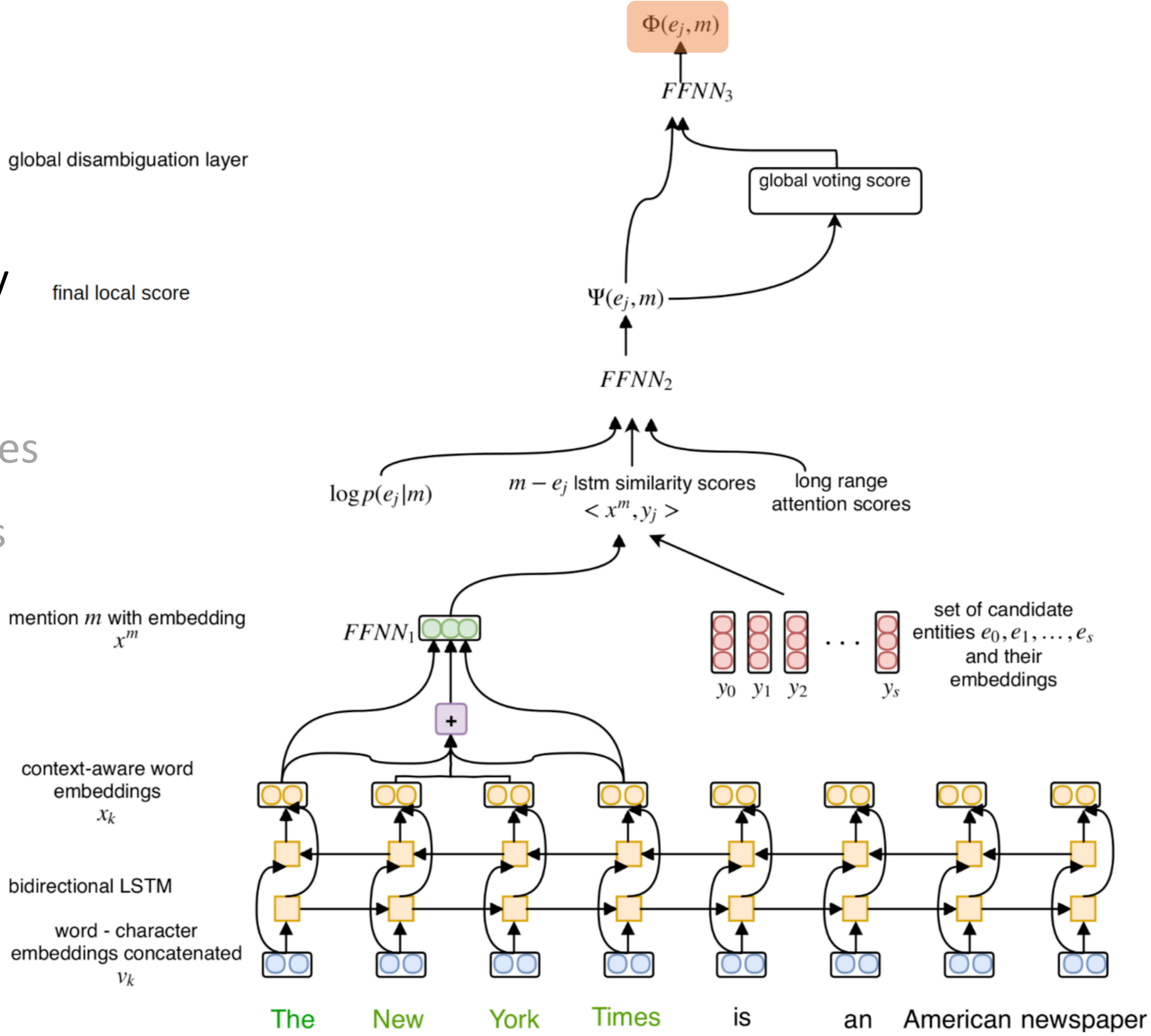
- Modules in pipeline (Disambiguation-Only)
  - Candidate selection
    - Dictionary
    - Anchors statistic
    - Surface matching heuristic
  - Scoring candidates
    - Entity embedding
    - Local compatibility (modeling the selected mention and its context)
    - Global coherence (modeling other mentions and their candidates)

- Candidate selection
  - Dictionary (Hoffart et al., 2011; Yamada et al., 2016; Cao et al., 2017; Cao et al., 2018)
    - Constructed from knowledge bases, e.g., DBpedia, YAGO, etc.
    - Examples:

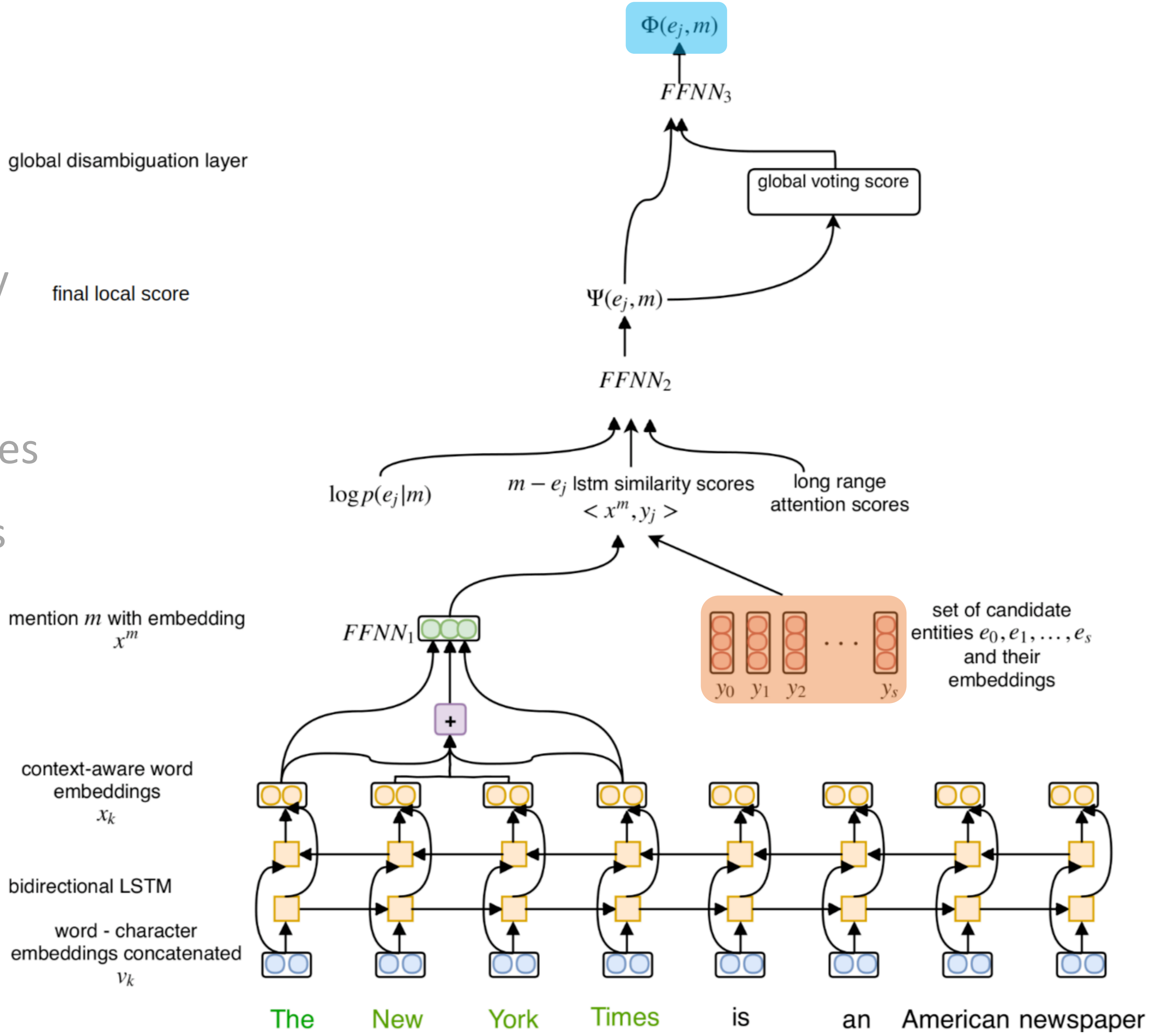
      "Apple" for `Apple Inc.`

      "Big Apple" for `New York City`
  - Anchors statistic (Ganea et al., 2017; Kolitsas et al., 2018)
    - Mention-entity prior: $P(e|m) = |A_{e,m}|/|A_{*,m}|$
    - Computed from mention entity hyperlink count statistic from Wikipedia etc.
    - Also as a feature for disambiguation
  - Surface matching heuristic (Le and Titov, 2019)

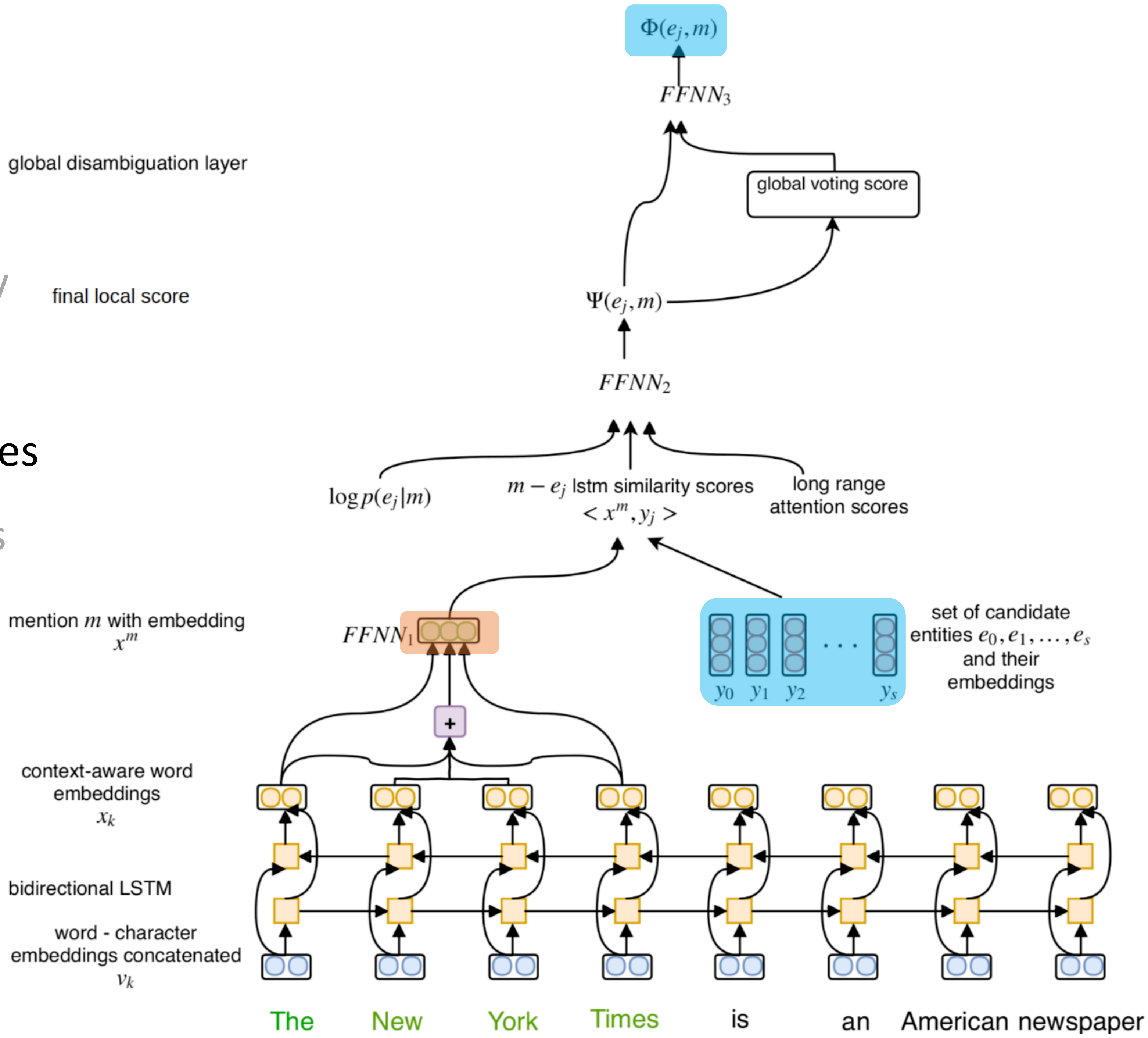
- Scoring candidates
  - ([Kolitsas et al., 2018]())
  - Entity-mention compatibility
  - Entity embedding
  - Context-Independent features
  - Context-Dependent features
  - Mention-entity prior
  - Global features



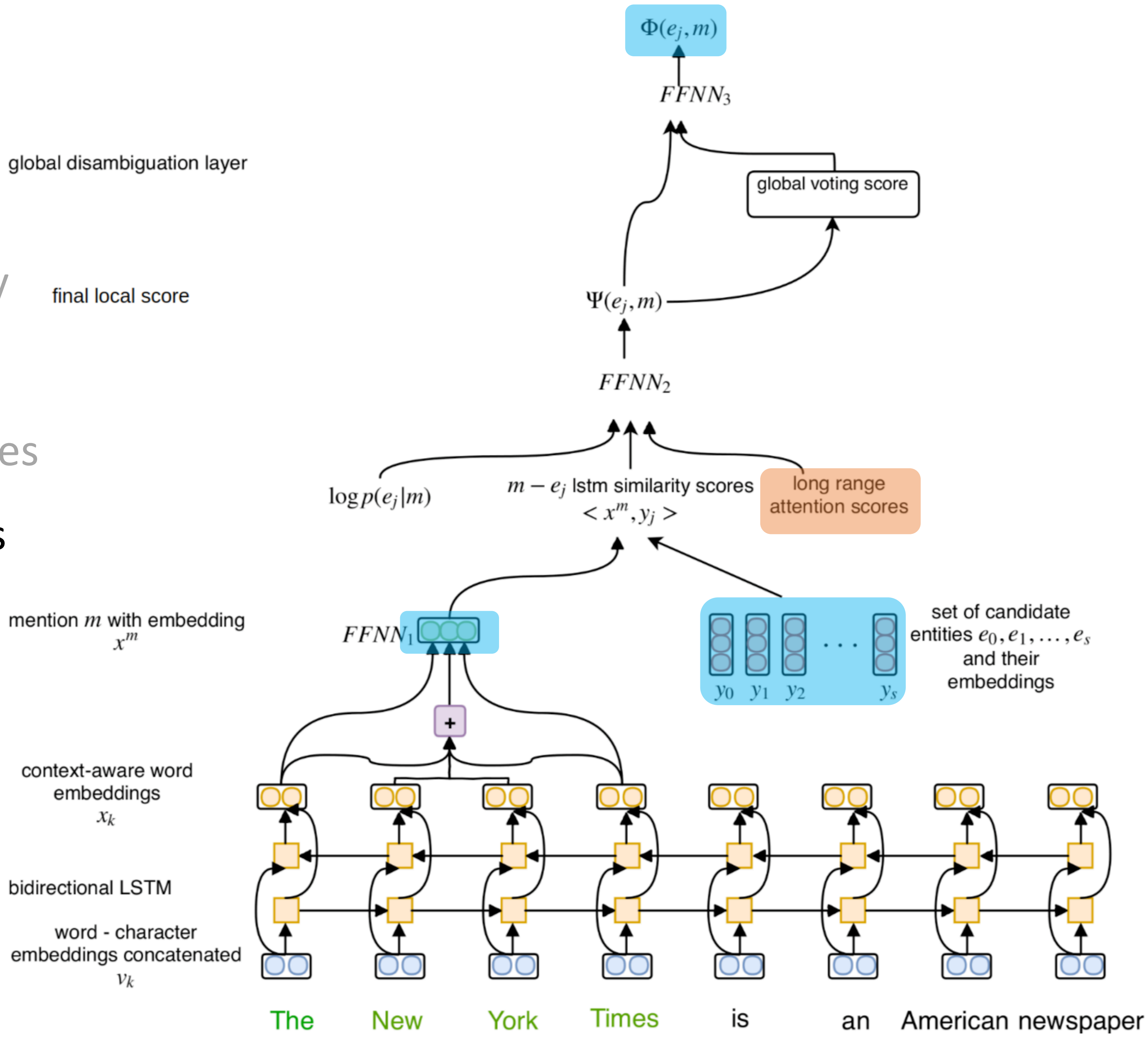
- Scoring candidates
  - ([Kolitsas et al., 2018](#))
  - Entity-mention compatibility
  - Entity embedding
  - Context-Independent features
  - Context-Dependent features
  - Mention-entity prior
  - Global features



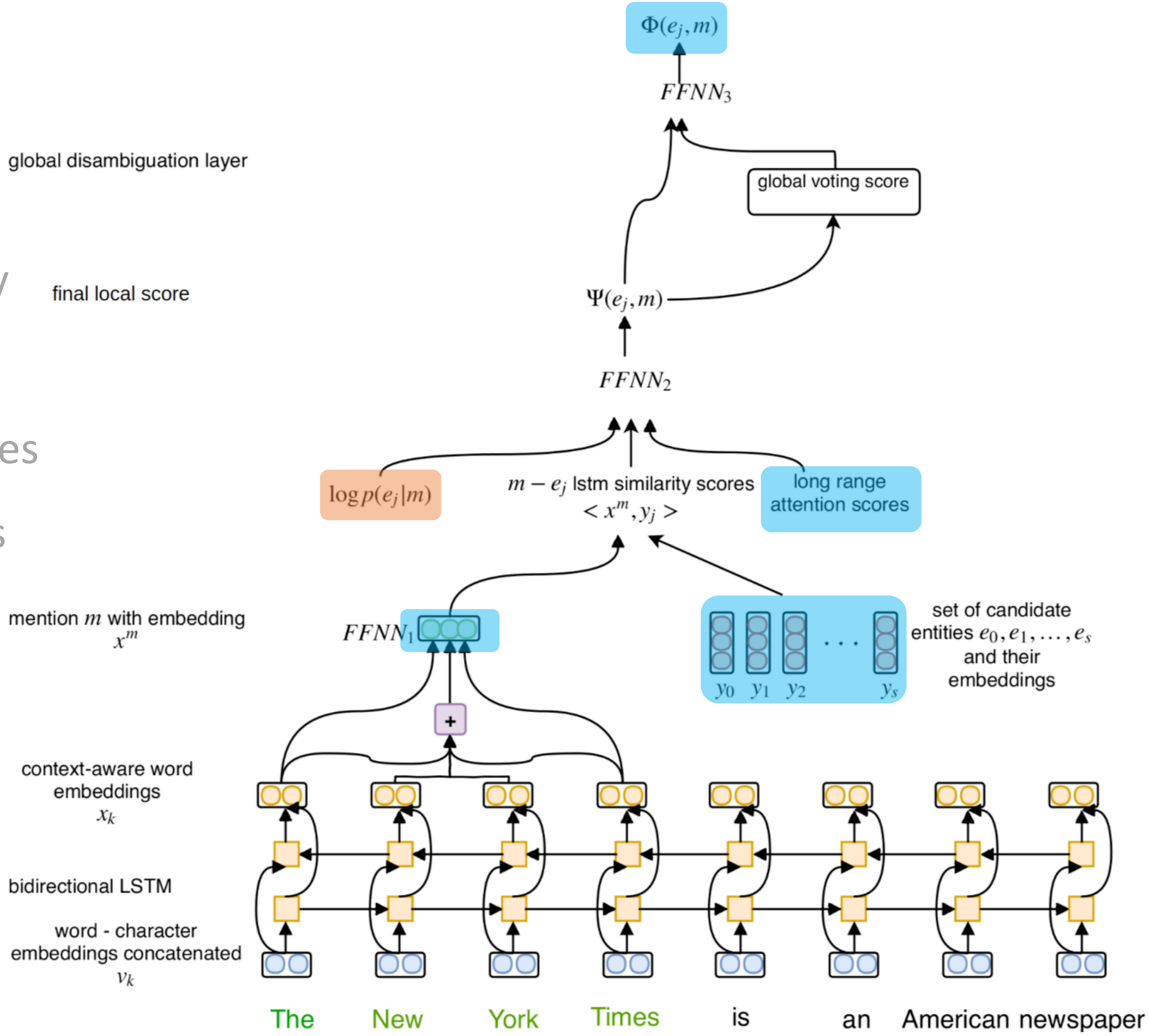
- Scoring candidates
  - (Kolitsas et al., 2018)
  - Entity-mention compatibility
  - Entity embedding
  - Context-Independent features
  - Context-Dependent features
  - Mention-entity prior
  - Global features



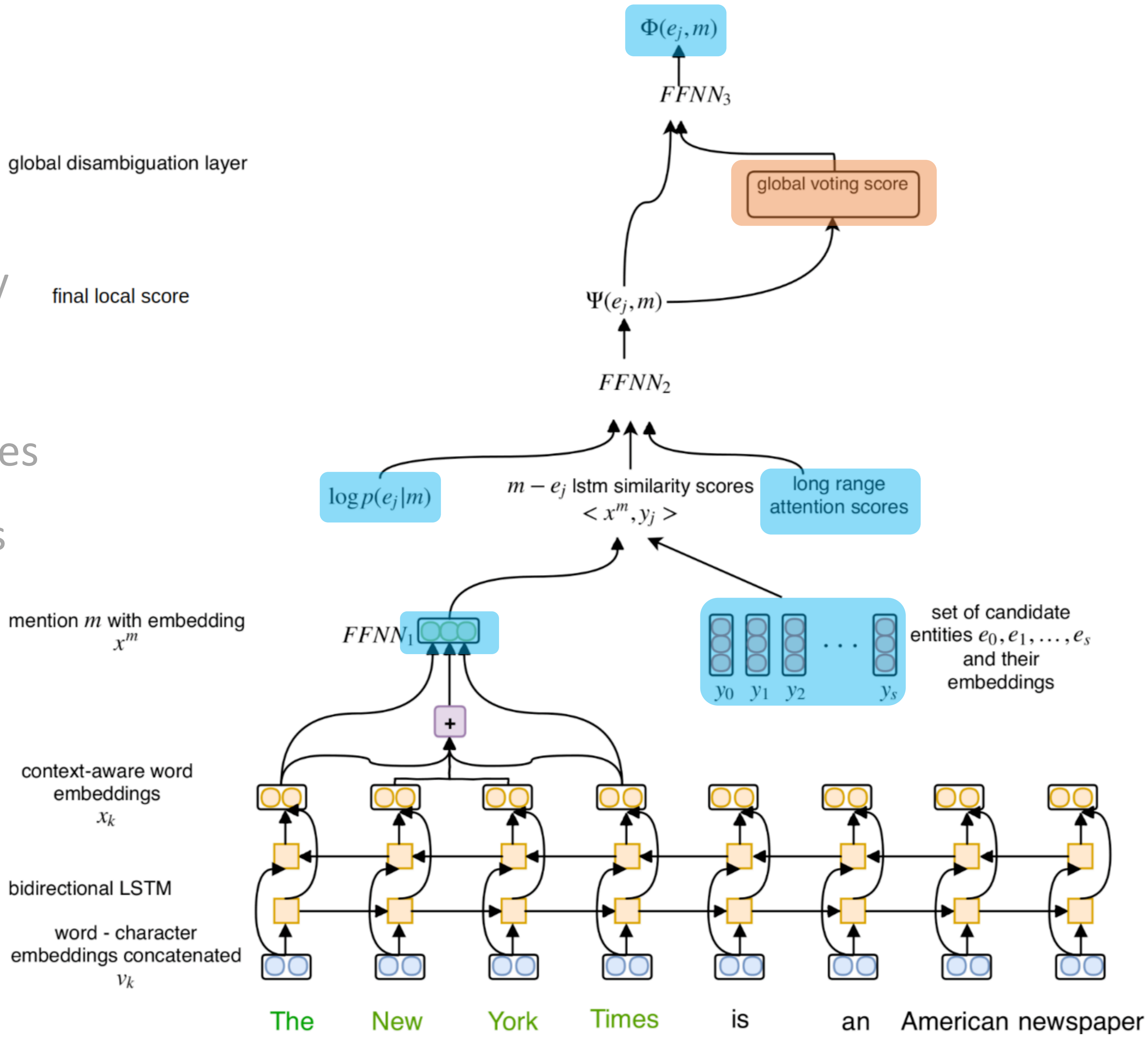
- Scoring candidates
  - (Kolitsas et al., 2018)
  - Entity-mention compatibility
  - Entity embedding
  - Context-Independent features
  - Context-Dependent features
  - Mention-entity prior
  - Global features



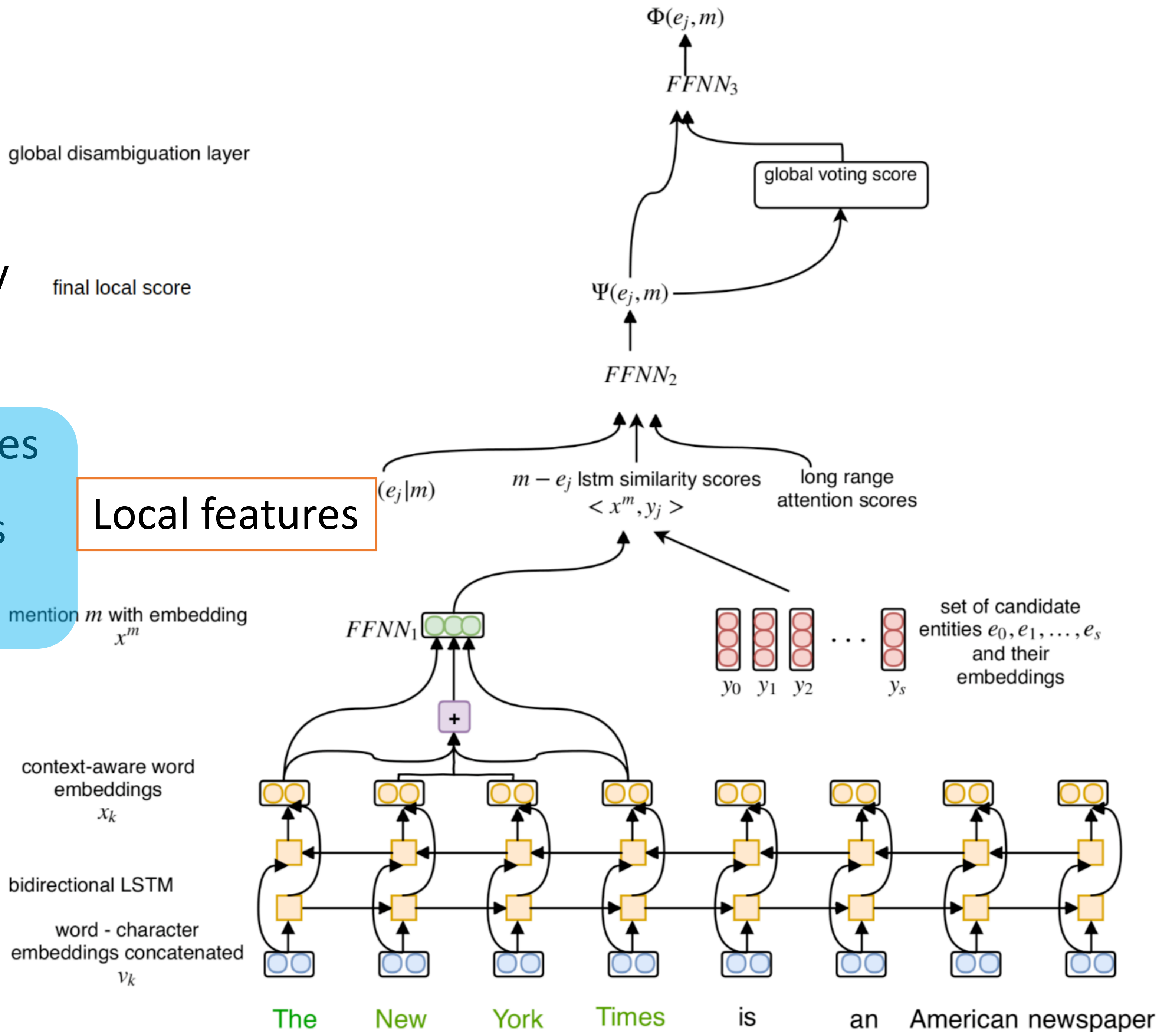
- Scoring candidates
  - (Kolitsas et al., 2018)
  - Entity-mention compatibility
  - Entity embedding
  - Context-Independent features
  - Context-Dependent features
  - Mention-entity prior
  - Global features



- Scoring candidates
  - (Kolitsas et al., 2018)
  - Entity-mention compatibility
  - Entity embedding
  - Context-Independent features
  - Context-Dependent features
  - Mention-entity prior
  - Global features



- Scoring candidates
  - (Kolitsas et al., 2018)
  - Entity-mention compatibility
  - Entity embedding
  - Context-Independent features
  - Context-Dependent features
  - Mention-entity prior
  - Global features



- Scoring candidates – Entity embedding
  - Jointly map words / mentions and entities into the same continuous vector space.
  - (Yamada et al., 2016; Ganea et al., 2017)

1. Skip-gram model (for words)

$$P(w_{t+j}|w_t) = \frac{\exp({\mathbf{V}_{w_t}}^\top \mathbf{U}_{w_{t+j}})}{\sum_{w \in W} \exp({\mathbf{V}_{w_t}}^\top \mathbf{U}_w)}$$

2. KB graph model (extend word embedding matrix V and U for entities)

$$P(e_o|e_i) = \frac{\exp({\mathbf{V}_{e_i}}^\top \mathbf{U}_{e_o})}{\sum_{e \in E} \exp({\mathbf{V}_{e_i}}^\top \mathbf{U}_e)}$$

3. Anchor context model (let words and entities interact with each other via anchors)

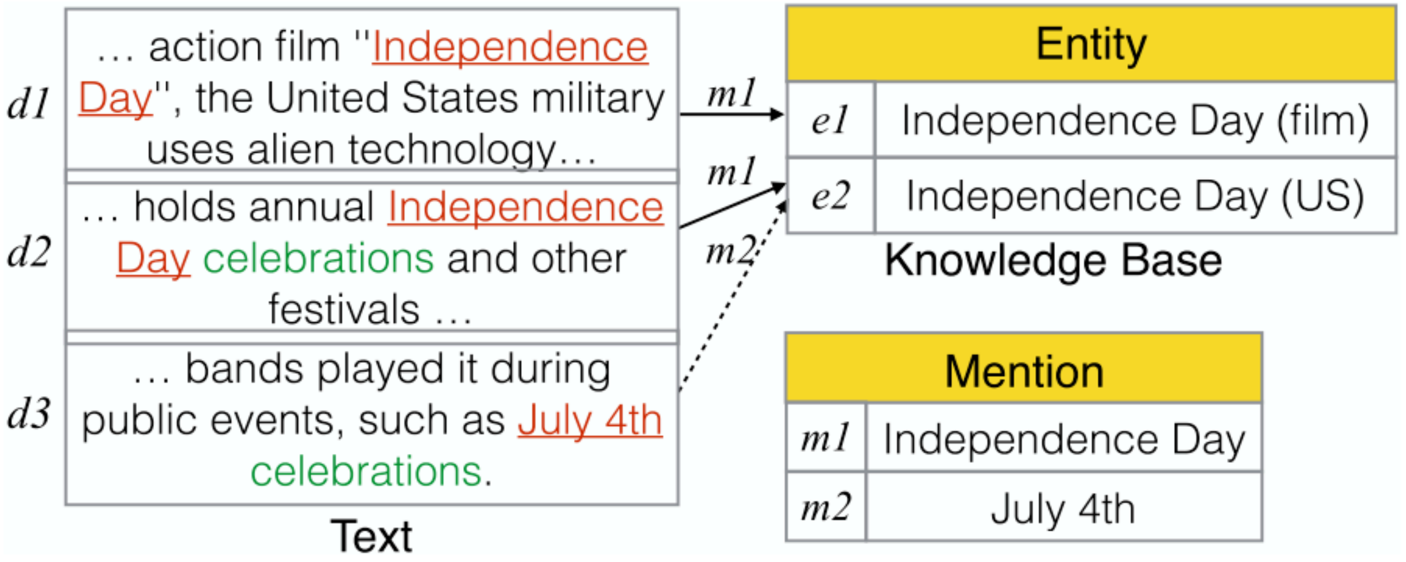
$$P(w_o|e_i) = \frac{\exp({\mathbf{V}_{e_i}}^\top \mathbf{U}_{w_o})}{\sum_{w \in W} \exp({\mathbf{V}_{e_i}}^\top \mathbf{U}_w)}$$

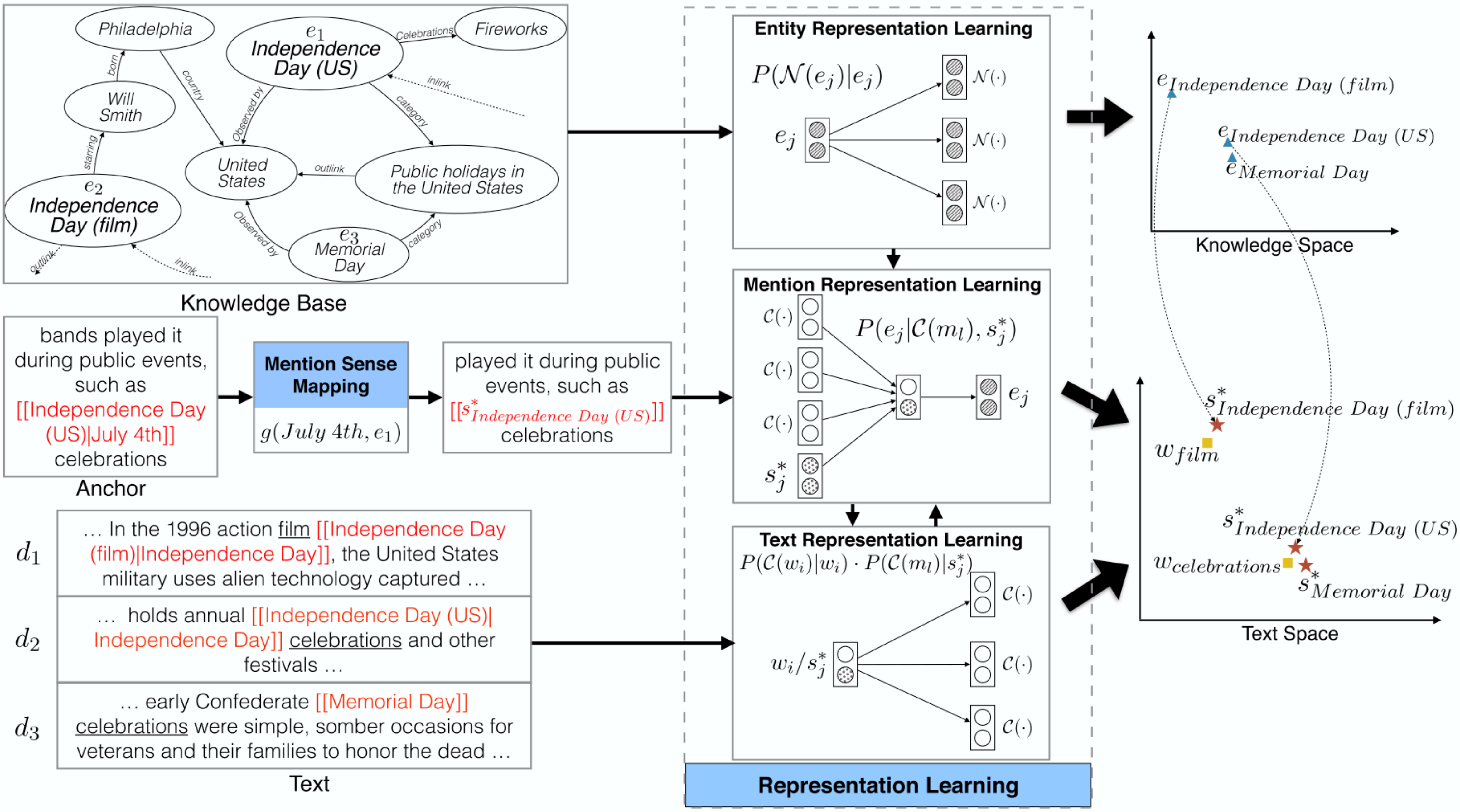
- Scoring candidates – Entity embedding
  - Jointly map words / mentions and entities into the same continuous vector space.
  - (Yamada et al., 2016; Ganea et al., 2017)
  - Based on word2vec pre-trained vectors

$$J(\mathbf{z}; e) := \mathbb{E}_{w^+|e}\, \mathbb{E}_{w^-} \left[h\left(\mathbf{z}; w^+, w^-\right)\right]$$
$$h(\mathbf{z}; w, v) := [\gamma - \langle \mathbf{z}, \mathbf{x}_w - \mathbf{x}_v \rangle]_+$$
$$\mathbf{x}_e := \arg\min_{\mathbf{z}:||\mathbf{z}||=1} J(\mathbf{z}; e)$$

  - where $w^+ \sim \hat{p}(w|e) \propto \#(w, e)$ and $w^- \sim q(w)$
  - Let vectors of positive words are closer to the embedding of entity e.

- Scoring candidates – Entity embedding
  - Map words / mentions and entities into different vector space.
  - (Cao et al., 2017)
  - Based on Skip-gram and CBOW
  - Learn representations for words, entities, and mention senses.



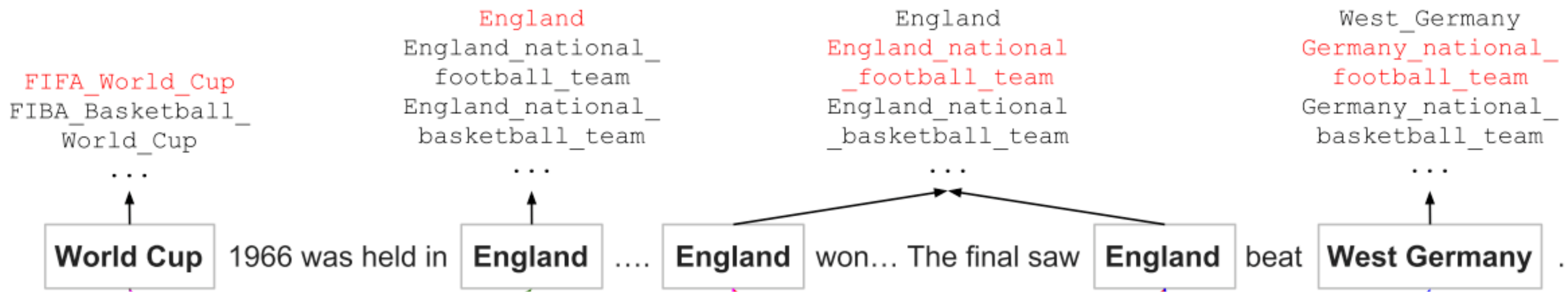


Philadelphia
Fireworks
e_1 Independence Day (US)
Celebrations
born
country
inlink
Will Smith
Observed by
category
starring
United States
outlink
Public holidays in the United States
e_2 Independence Day (film)
Observed by
e_3 Memorial Day
category
outlink
inlink
Knowledge Base
bands played it during public events, such as [[Independence Day (US)|July 4th]] celebrations
Anchor
Mention Sense Mapping
$g(July\ 4th, e_1)$
played it during public events, such as [[$s^*_{Independence\ Day\ (US)}$]] celebrations
d_1
… In the 1996 action film [[Independence Day (film)|Independence Day]], the United States military uses alien technology captured …
d_2
… holds annual [[Independence Day (US)| Independence Day]] celebrations and other festivals …
d_3
… early Confederate [[Memorial Day]] celebrations were simple, somber occasions for veterans and their families to honor the dead …
Text
Entity Representation Learning
$P(\mathcal{N}(e_j)|e_j)$
e_j
$\mathcal{N}(\cdot)$
Mention Representation Learning
$P(e_j|\mathcal{C}(m_l), s^*_j)$
$\mathcal{C}(\cdot)$
s^*_j
e_j
Text Representation Learning
$P(\mathcal{C}(w_i)|w_i) \cdot P(\mathcal{C}(m_l)|s^*_j)$
w_i/s^*_j
$\mathcal{C}(\cdot)$
Representation Learning
$e_{Independence\ Day\ (film)}$
$e_{Independence\ Day\ (US)}$
$e_{Memorial\ Day}$
Knowledge Space
$s^*_{Independence\ Day\ (film)}$
w_{film}
$s^*_{Independence\ Day\ (US)}$
$w_{celebrations}$
$s^*_{Memorial\ Day}$
Text Space

- Scoring candidates – Local feature (modeling mentions, contexts, and entities)
  - Mention-entity prior: $P(e|m) = |A_{e,m}|/|A_{*,m}|$
  - Context-Independent feature
    - String similarity (Cao et al., 2018)
    - Char BiLSTM (Kolitsas et al., 2018)
  - Context-Dependent feature
    - Average over context words (Yamada et al., 2016; Cao et al., 2017)
    - BiLSTM (Kolitsas et al., 2018 ; Le and Titov, 2019)
    - Attention (Ganea et al., 2017; Kolitsas et al, 2018; Cao et al., 2018)

- Scoring candidates – Global feature (modeling other mentions and their candidates)
  - Hand-crafted feature like number of shared incoming links… (Hoffart et al., 2011)
  - Bag-of-Words (Yamada et al., 2016)
  - Voting-based (Kolitsas et al, 2018)
  - Markov chain (Delpeuch et al., 2019)
  - CRF (Ganea et al., 2017)
  - GCN (Cao et al., 2018)

All mentions in a document shall be on the same topic!



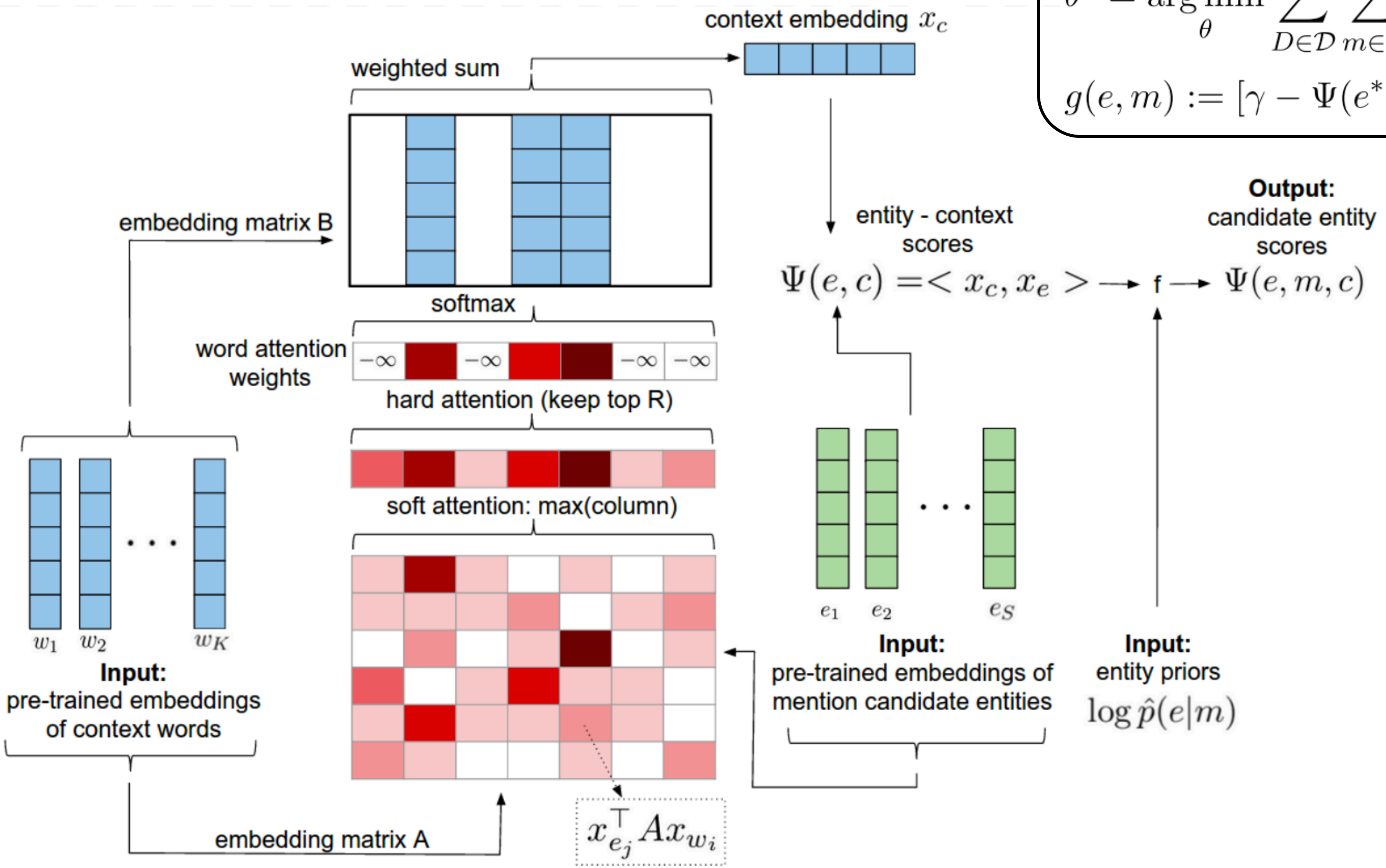
- Outline
  - Models
    - Modules
    - Neural models
    - Symbol-neural hybrid model
  - Related topics
    - Distant learning
    - Entity typing
  - Datasets, metrics, and platform

- A local model (Ganea et al., 2017)

Training objective (max-margin loss)

$$\theta^* = \arg\min_{\theta} \sum_{D \in \mathcal{D}} \sum_{m \in D} \sum_{e \in \Gamma(m)} g(e, m),$$

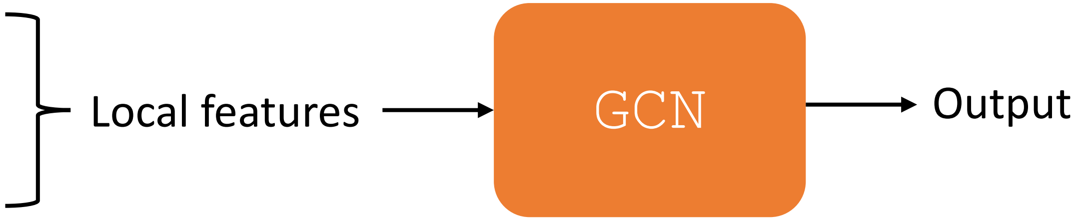
$$g(e, m) := [\gamma - \Psi(e^*, m, c) + \Psi(e, m, c)]_+$$

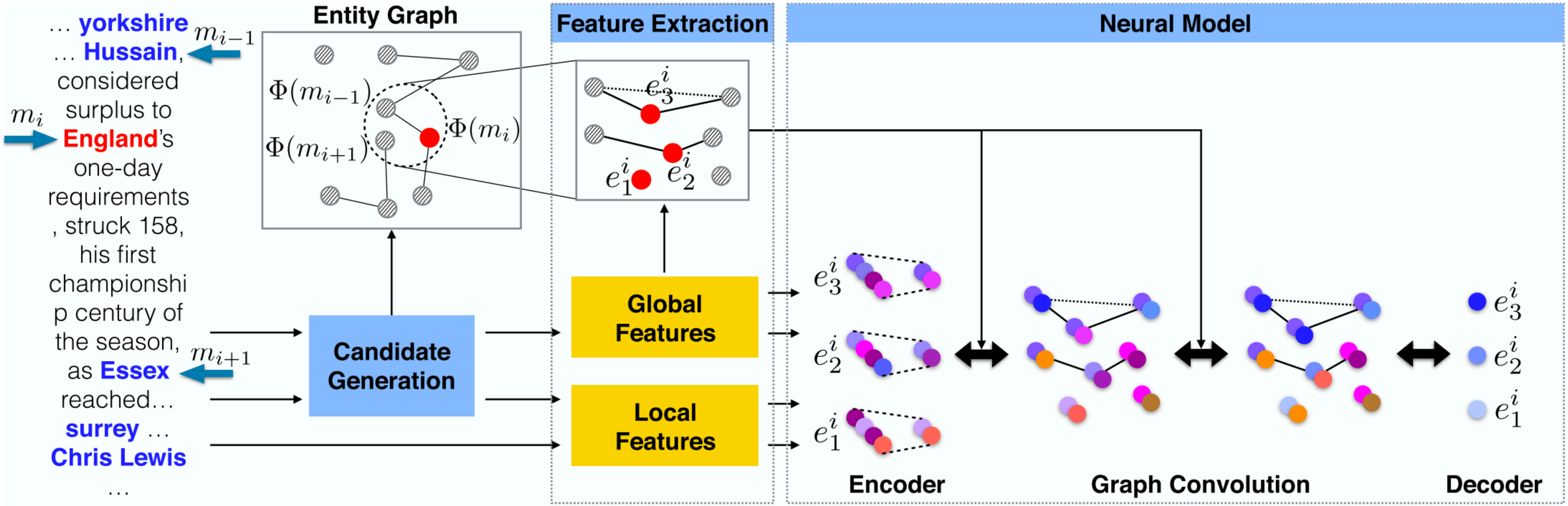


- A global model (Cao et al., 2018)

$$\mathbf{c}_{m_i,e_j} = \sum_{w_k \in \mathcal{C}(m_i)} \alpha_{kj} \mathbf{w}_k$$

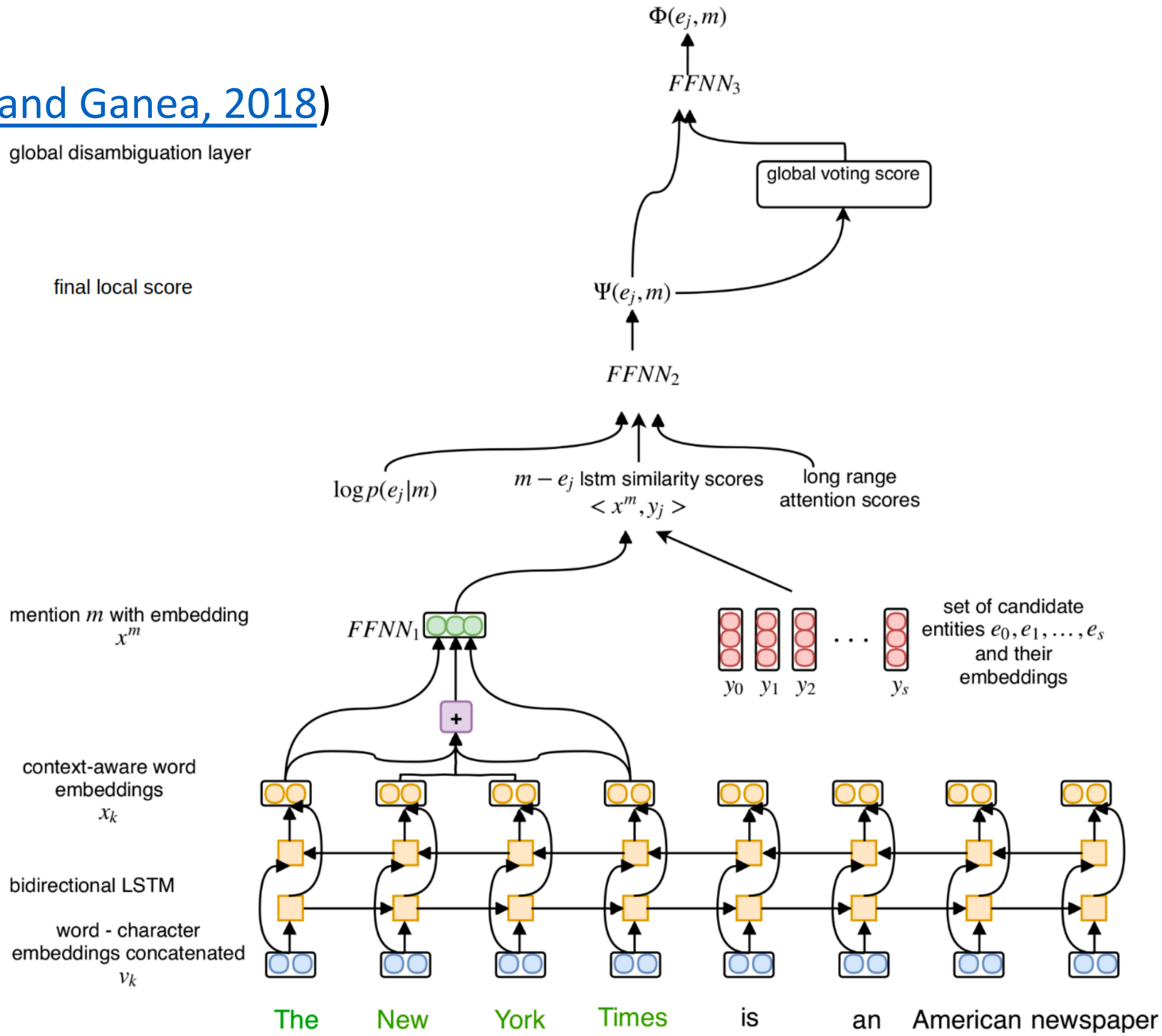
$$\{sim(\mathbf{e}_j, \mathbf{m}_i) | m_i \in \mathcal{N}(m_i)\}$$





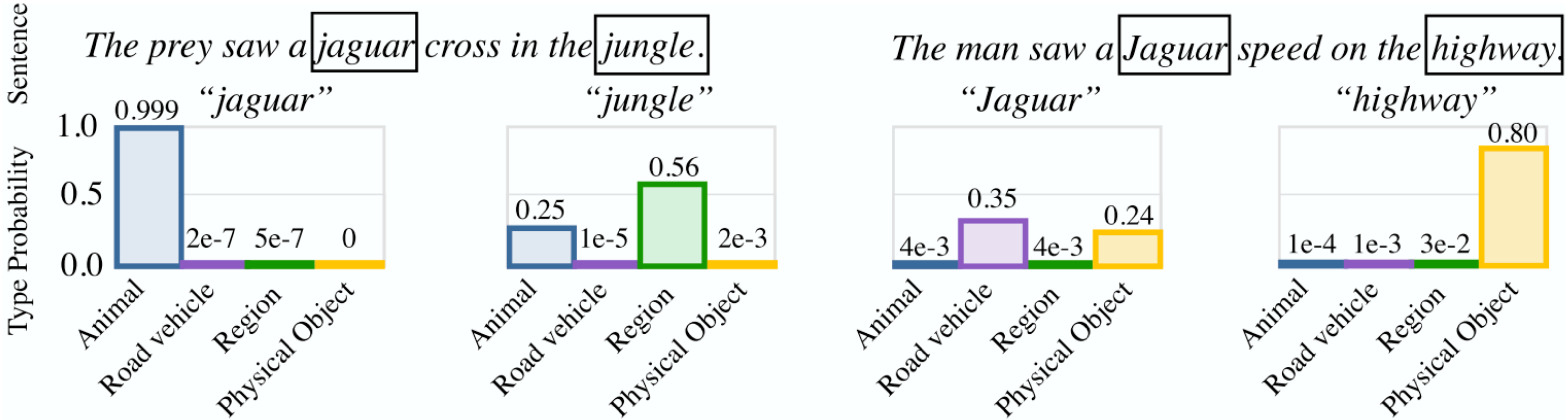
- An end-to-end Model ([Kolitsas and Ganea, 2018](#))

"At training time, for each input document we collect the set M of all (potentially overlapping) token spans m for which |C(m)| ≥ 1."



- Outline
  - Models
    - Modules
    - Neural models
    - Symbol-neural hybrid model
  - Related topics
    - Distant learning
    - Entity typing
  - Datasets, metrics, and platform

- DeepType ([Raiman and Raiman, 2018](#))
  - Associate with each entity a series of types (e.g. *Person*, *Place*, etc.) that if known, would rule out invalid answers, and therefore ease linking.



Entity	**jaguar**	**Jaguar**	**jungle**	**jungle**	**jaguar**	**Jaguar**	**highway**	**Highway**
Type	Animal	Road vehicle	Region	Music	Animal	Road vehicle	Physical Object	Film
only link Prob.	0.29	**0.60**	**0.35**	0.17	0.29	**0.60**	**0.85**	0.04
Prob. w/. types	**1.0**	0.0	**1.0**	0.0	0.0	**1.0**	**1.0**	0.0

- DeepType (Raiman and Raiman, 2018)
  - Terminology
    - Relation (e.g. `instance of`)
    - Type

      A label defined by a relation, e.g., the type applied to all children of `Human` connected by `instance of` is `IsHuman`.
    - Type Axis: a set of mutually exclusive types
    - Type System: type axes + type labelling function

- DeepType (Raiman and Raiman, 2018)
  - Type System
    - A: the assignment for the boolean discrete variables that define the type system.

      $A_i = 1$ if the i-th parent-child relation gets included in the type system.

      $$A = \{0, 1, 0, 1, 1, \ldots\}$$
    - Optimize: heuristic search / stochastic optimization (mixed integer problem)
  - Type Classifier
    - θ: continuous variables that parameterize the classifier to fit to the type system.
    - Optimize: gradient descent
  - Objective: solve A and θ

$$\max_{\mathcal{A}} \max_{\theta} S_{\text{model}}(\mathcal{A}, \theta) = \frac{\sum_{(m, e_{\text{GT}}, \mathcal{E}_m) \in M} \mathbb{1}_{e_{\text{GT}}}(e^*)}{|M|}.$$

- DeepType (Raiman and Raiman, 2018)
  - Discrete optimization of the type system
    - Define an objective to measure how good a solution is
    - There is a trade-off
      - Disambiguation power

        Measure the improvement of entity linking accuracy of the solution.
      - Learnability

        Measure how learnable the type axes in the selected solution.
    - Regularization

$$J(\mathcal{A}) = (S_{\text{oracle}} - S_{\text{greedy}}) \cdot \text{Learnability}(\mathcal{A}) + S_{\text{greedy}} - |\mathcal{A}| \cdot \lambda.$$

- DeepType (Raiman and Raiman, 2018)
  - Objective of type system

$$J(\mathcal{A}) = (S_{\text{oracle}} - S_{\text{greedy}}) \cdot \text{Learnability}(\mathcal{A}) + S_{\text{greedy}} - |\mathcal{A}| \cdot \lambda.$$

  - Mention-entity prior: $\mathbb{P}_{\text{Link}}(e|m) = \frac{\text{LinkCount}(m,e)}{\sum_{j \in \mathcal{E}_m} \text{LinkCount}(m,j)}$
  - Greedy: predicts only according to the mention-entity prior.
  - Oracle: prunes candidate set to only contain entities whose types match those of e_i^{GT}

$$\text{Oracle}(m) = \underset{e \in \mathcal{E}_{m,\text{oracle}}}{\text{argmax}} \; \mathbb{P}_{\text{entity}}(e|m, \text{types}(x)).$$

$$S_{\text{oracle}} = \frac{\sum_{(m, e_{\text{GT}}, \mathcal{E}_m) \in M} \mathbb{1}_{e_{\text{GT}}}(\text{Oracle}(m))}{|M|}.$$

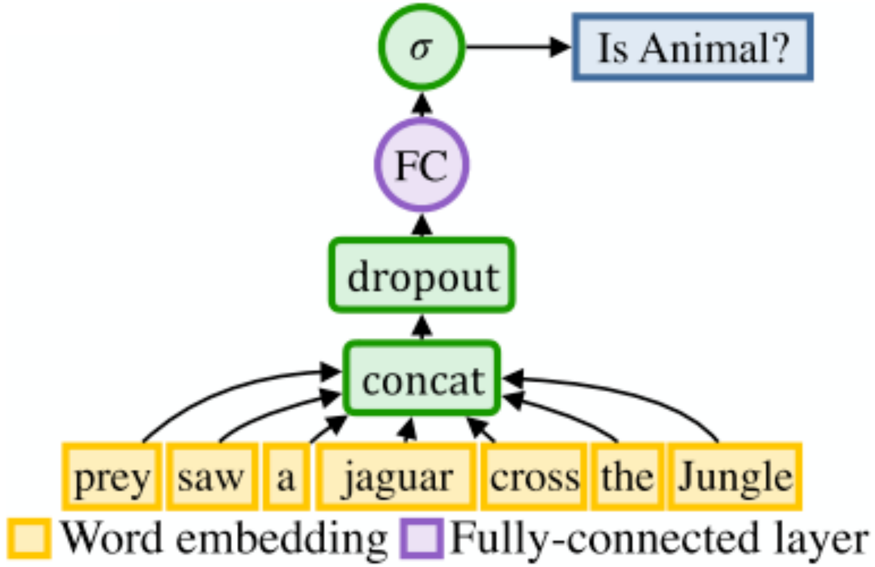
- DeepType (Raiman and Raiman, 2018)
  - Objective of type system

$$J(\mathcal{A}) = (S_{\text{oracle}} - S_{\text{greedy}}) \cdot \text{Learnability}(\mathcal{A}) + S_{\text{greedy}} - |\mathcal{A}| \cdot \lambda.$$

  - Learnability

$$\text{Learnability}(\mathcal{A}) = \frac{\sum_{t \in \mathcal{A}} \text{AUC}(t)}{|\mathcal{A}|}$$

  - λ: per type axis penalty term

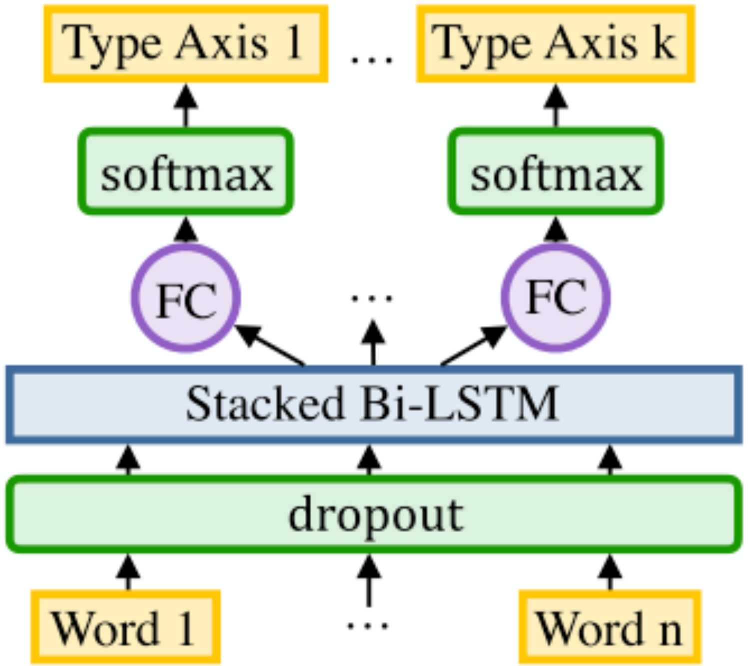


- DeepType (Raiman and Raiman, 2018)
  - Objective of type system

$$
\begin{aligned}
J(\mathcal{A}) = &(S_{\text{oracle}} - S_{\text{greedy}}) \cdot \text{Learnability}(\mathcal{A}) + \\
&S_{\text{greedy}} - |\mathcal{A}| \cdot \lambda.
\end{aligned}
$$

  - Search methodologies
    - Beam search and greedy selection
    - Cross-entropy method
    - Genetic algorithm
    - …

- DeepType (Raiman and Raiman, 2018)
  - Discrete optimization of the type system
  - Type classifier
    - Classify per-token type



- DeepType (Raiman and Raiman, 2018)
  - Discrete optimization of the type system
  - Type classifier
  - Inference
    - Given Input words $w_0, \ldots, w_L$ and mention *m* covering words $w_x, \ldots, w_y$
    - Through type classifier, we obtain the type conditional probability for all type axes i: $\{\mathbb{P}_i(\cdot|w_x, D), \ldots, \mathbb{P}_i(\cdot|w_y, D)\}$
    - Aggregate using max-over-time and obtain $\mathbb{P}_{i,*}(\cdot|m, D)$
    - Take the prior into consideration, we get the final entity score

$$s_{e,m,D,\mathcal{A},\theta} = \mathbb{P}_{\text{Link}}(e|m) \cdot \left(1 - \beta + \beta \cdot \left\{\prod_{i=1}^{k}(1 - \alpha_i + \alpha_i \cdot \mathbb{P}_{i,*}(t_i|m, D))\right\}\right).$$

- Outline
  - Models
    - Modules
    - Neural models
    - Symbol-neural hybrid model
  - Related topics
    - Distant learning
    - Entity typing
  - Datasets, metrics, and platform

- Distant learning
  - Distant supervision (also referred to weak supervision) assumption:

    If two entities participate in a relation, all sentences that mention these two entities express that relation.
  - An example:

    Elevation Partners*, the $ 1.9 billion private equity group that was* <u>*founded*</u> *by* ***Roger McNamee***
  - However, the assumption can be violated:

    Roger McNamee*, a managing director at* ***Elevation Partners****, ...*

- Distant learning
  - When aligning Freebase to Wikipedia and New York Times…

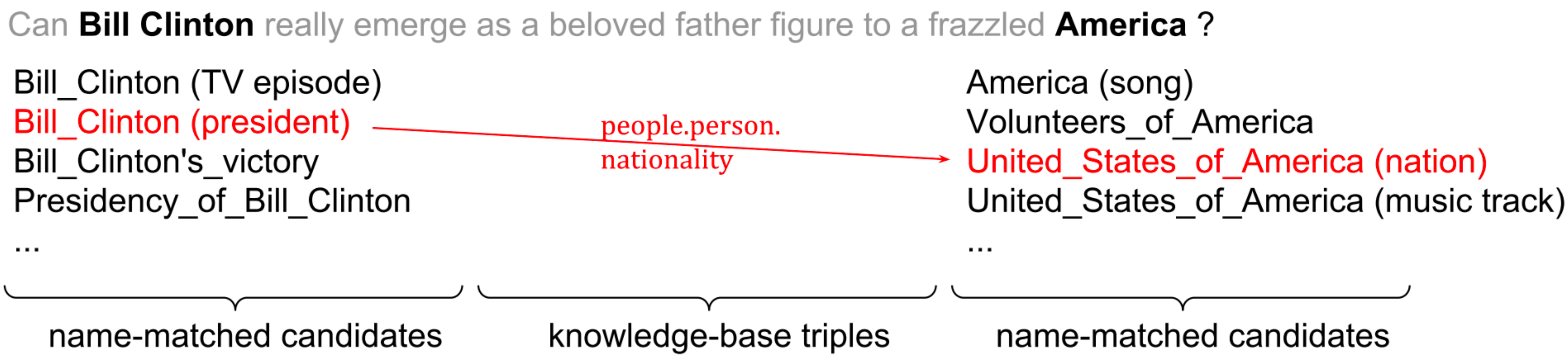
Table 1. Percentage of times a related pair of entities is mentioned in the same sentence, but where the sentence does not express the corresponding relation

Relation Type	New York Times	Wikipedia
nationality	38%	20%
place_of_birth	35%	20%
contains	20%	10%

  - (Riedel et al., 2010) proposed a relaxed assumption:

    *If two entities participate in a relation, **at least one sentence** that mentions these two entities might express that relation.*

- Distant learning in entity linking (Le and Titov, 2019)
  - Construct distant supervision: surface matching heuristics (measure overlap)
  - Positive lists: top candidates from the matching heuristics
  - Negative lists: randomly sampled sets of entities
  - Multi-Instance Learning (MIL): find the entity should be linked



- Distant learning in entity linking (Le and Titov, 2019)
  - During training, we have $\langle m, c, E^+, E^- \rangle$, in testing, $E^- = \emptyset$.
  - MIL: we want to train the model to score at least one candidate in E^+ higher than any candidate in E^-. To achieve this, we employ a max-margin loss

$$l(m,c) = [\max_{e \in E^-} g(e,m,c) + \delta - \max_{e \in E^+} g(e,m,c)]_+$$

$$L_1 = \sum_{(m,c) \in D} l(m,c)$$

  - Recall that many data points are noisy. E^+ may not contain the correct entity.

- Distant learning in entity linking (Le and Titov, 2019)
  - Representation for E^+
    - Use attention $\mathbf{e}_{E^+} = \sum_{e \in E^+} \alpha_e \mathbf{e}$
  - Noise detection
    - Use a binary classifier

$$p_N(1|m, c, E^+) = \sigma\left(\frac{\mathrm{FFN}_f([\mathbf{e}_{E^+}, \mathbf{f}_{h-1}, \mathbf{b}_{h-1}, \mathbf{f}_k, \mathbf{b}_k])}{T}\right)$$

  - Training
    - Down-weight potentially noisy data points. New loss:

$$L_2 = \sum_{(m,c)\in D} p_N(0|m, c, E^+) l(m, c) + \eta \times \mathbf{KL}\left(\frac{\sum_{(m,c)\in D} p_N(\cdot|m, c, E^+)}{|D|} \middle| p_N^*\right)$$

  - Testing: with / without noise detector

- Outline
  - Models
    - Modules
    - Neural models
    - Symbol-neural hybrid model
  - Related topics
    - Distant learning
    - Entity typing
  - Datasets, metrics, and platform

- Entity Typing
  - FIGER (Ling and Weld, 2012)
  - Fine-grained NER task
  - Hierarchical labels

```
person/
person/actor
location/
location/city
…
```

person	doctor
actor	engineer
architect	monarch
artist	musician
athlete	politician
author	religious_leader
coach	soldier
director	terrorist

organization	terrorist_organization
airline	government_agency
company	government
educational_institution	political_party
fraternity_sorority	educational_department
sports_league	military
sports_team	news_agency

location	body_of_water
city	island
country	mountain
county	glacier
province	astral_body
railway	cemetery
road	park
bridge	

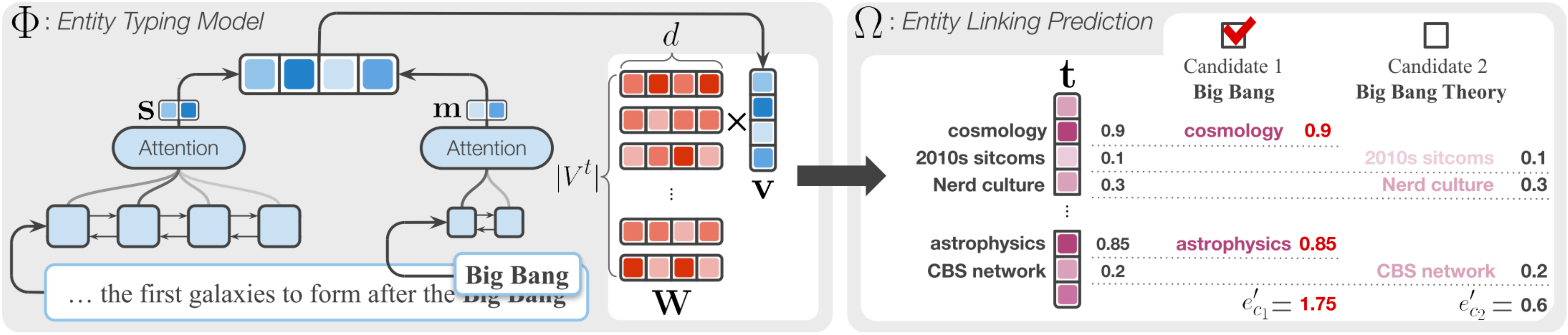
product	camera
engine	mobile_phone
airplane	computer
car	software
ship	game
spacecraft	instrument
train	weapon

art	written_work
film	newspaper
play	music

event	military_conflict
attack	natural_disaster
election	sports_event
protest	terrorist_attack

building	time	chemical_thing	website
airport	color	biological_thing	broadcast_network
dam	award	medical_treatment	broadcast_program
hospital	educational_degree	disease	tv_channel
hotel	title	symptom	currency
library	law	drug	stock_exchange
power_station	ethnicity	body_part	algorithm
restaurant	language	living_thing	programming_language
sports_facility	religion	animal	transit_system
theater	god	food	transit_line

- Entity Typing for Entity Linking (ET4EL) ([Onoe and Durrett, 2019])
  - Alleviate overfitting
  - Construct entity typing dataset using hyperlinks and Wiki categories
  - Two parts:
    - Entity typing: $\Phi : (m, s) \rightarrow T.$
    - Entity linking: $e = \Omega(\Phi(m, s), C).$



- Entity Typing for Entity Linking (ET4EL) (Onoe and Durrett, 2019)
  - Entity linking prediction (heuristic, untrained)
    - Ω is defined as the sum of probabilities for each type

$$e'_c = \sum_i t_i \cdot \mathbb{1}_{T_c}\left(V_i^t\right)$$

$$e = \arg\max_e \left(e'_1, \ldots, e'_{|C|}\right)$$

    - No need to access the labeled entity linking data.

- Outline
  - Models
    - Modules
    - Neural models
    - Symbol-neural hybrid model
  - Related topics
    - Distant learning
    - Entity typing
  - Datasets, metrics, and platform

- Datasets
  - AIDA-CoNLL (Hoffart et al., 2011)
    - Text data: CoNLL 2003 NER task
    - Knowledge base: YAGO
  - TAC 2010 (Ji et al., 2010)
    - Text data: news articles from various agencies and Web log data
  - WikiDisamb30 (Raiman and Raiman, 2018)
- Platform
  - GERBIL

- Metrics
  - Disambiguation-only
    - Micro accuracy
    - Macro accuracy
  - End-to-End
    - Micro F1
    - Macro F1
  - InKB v.s. NIL ("unlinkable")

Q & A

Paper list