END-TO-END NEURAL ENTITY LINKING

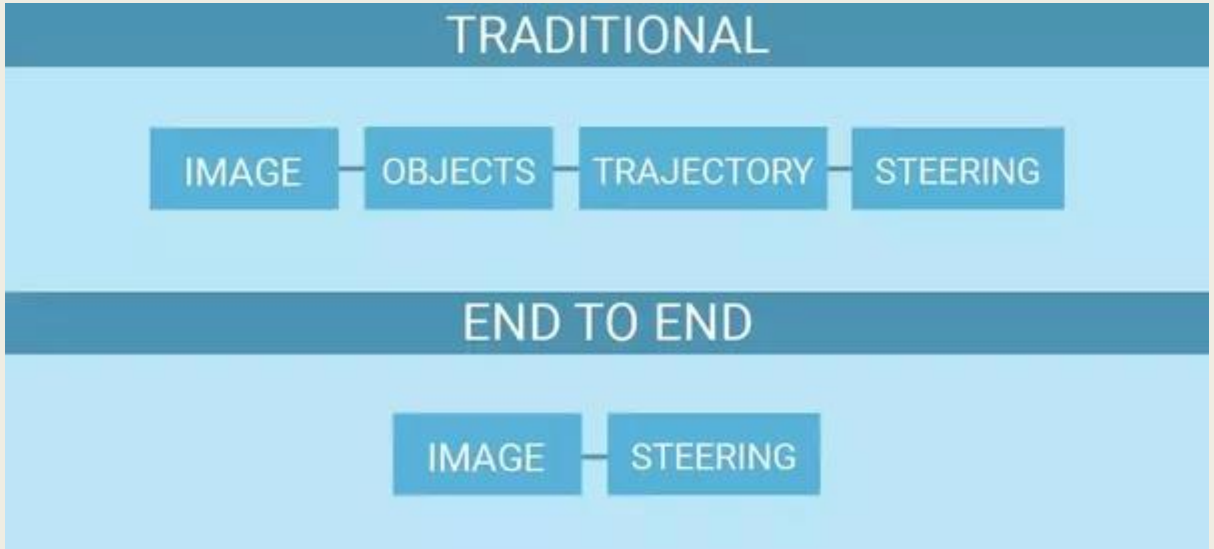
Nikolaos Kolitsas, Octavian-Eugen Ganea, Thomas Hofmann

CoNLL 2018

Mariana Leite
me.leite@campus.fct.unl.pt
NOVA Search Reading Group

End-to-End?

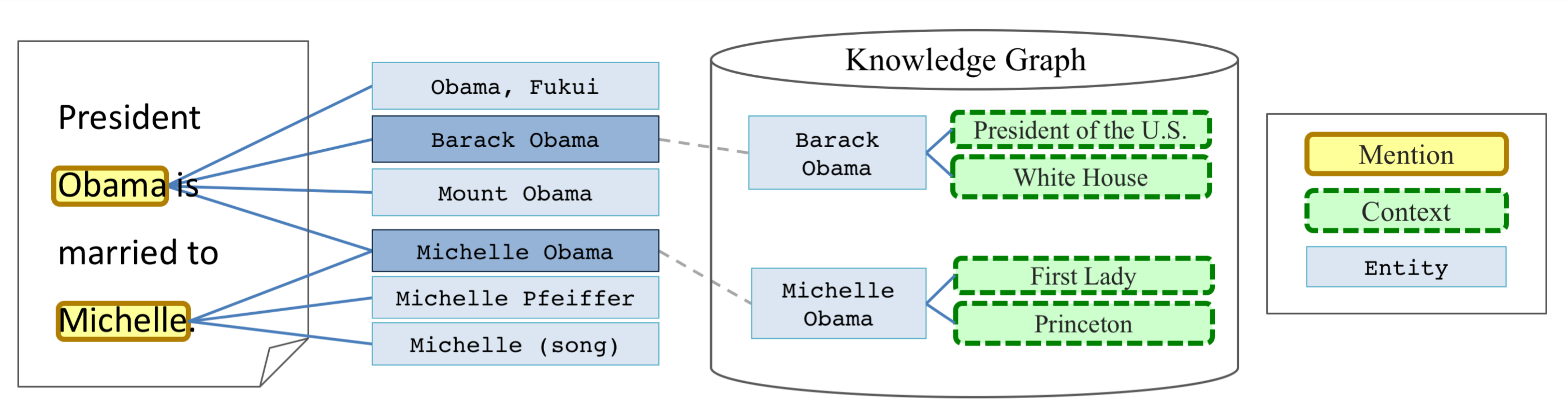
- End-to-end (E2E) learning refers to training a possibly complex learning system represented by a single model (specifically a Deep Neural Network) that represents the complete target system, bypassing the intermediate layers usually present in traditional pipeline designs [1].



[1] https://towardsdatascience.com/e2e-the-every-purpose-ml-method-5d4f20dafee4

The Goal of Automatic Text Understanding…

- Models are expected to accurately **extract** ambiguous mentions of entities from a textual document and **link** them to a knowledge base.



- EL systems typically perform 2 tasks: Mention Detection (MD) & Entity Disambiguation (ED)

Approaches

- The usual: solve the two tasks independently.
  - ! Important depency between the two steps is ignored;
  - ! Errors in first step will propagate to second step without possibility of recovery.

- This paper’s approach:
  - End-to-End Entity Linking (like humans do!).
  - Emphasizes the **importance of the mutual dependency between MD & ED.**

MD may benefit from ED (and viceversa)

1) MD may split a larger span into two mentions of less informative entities:

B. Obama's wife gave a speech [...]

Federer's coach [...]

2) MD may split a larger span into two mentions of incorrect entities:

Obama Castle was built in 1601 in Japan.

The Kennel Club is UK's official kennel club.

A bird dog is a type of gun dog or hunting dog.

Romeo and Juliet by Shakespeare [...]

Natural killer cells are a type of lymphocyte

Mary and Max, the 2009 movie [...]

MD may benefit from ED (and viceversa)

3) MD may choose a shorter span, referring to an incorrect entity:

The Apple is played again in cinemas.

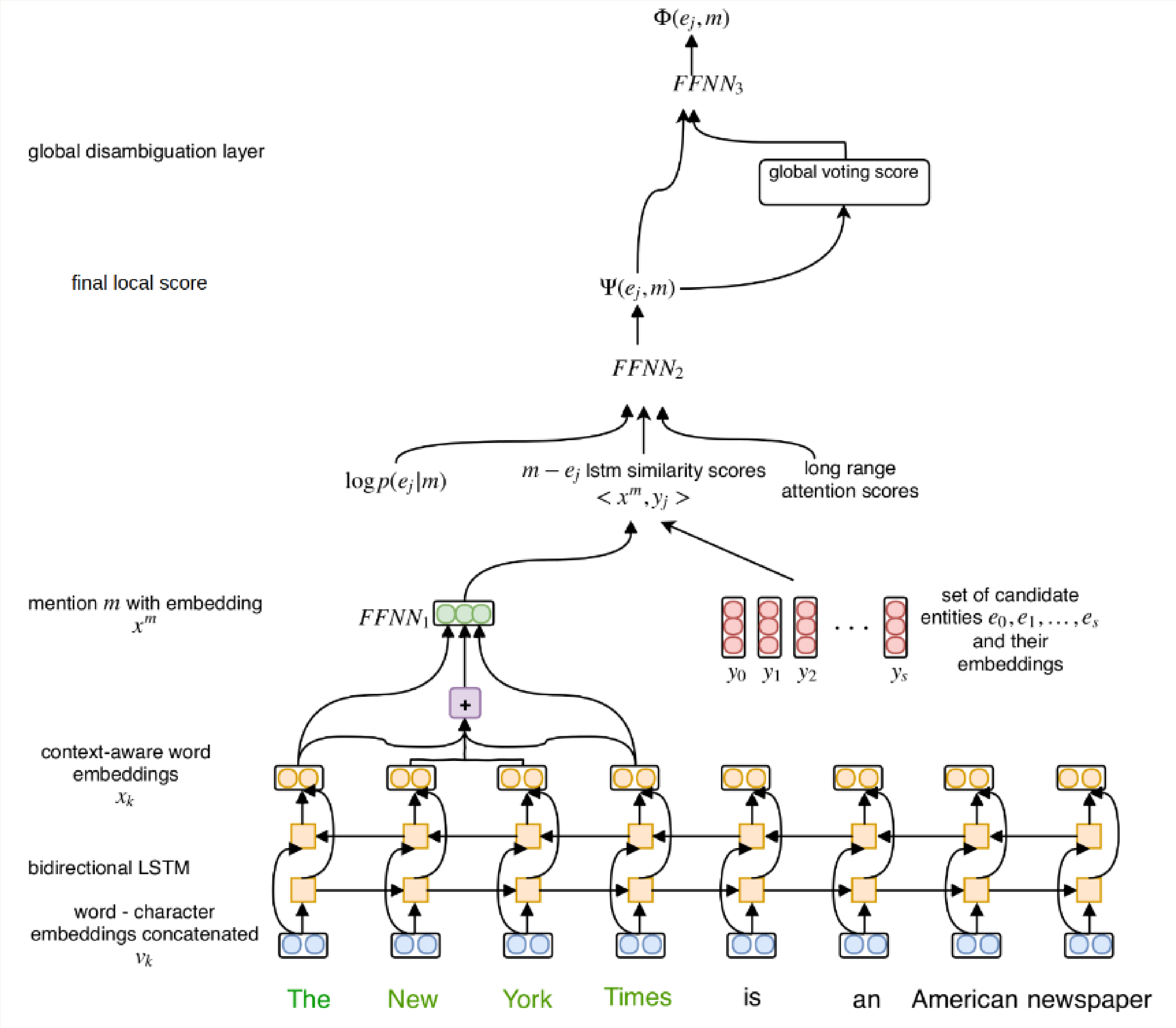
The New York Times is a popular newspaper.

4) MD may choose a longer span, referring to an incorrect entity:

Babies Romeo and Juliet were born hours apart.

Neural Joint Mention Detection & Entity Disambiguation

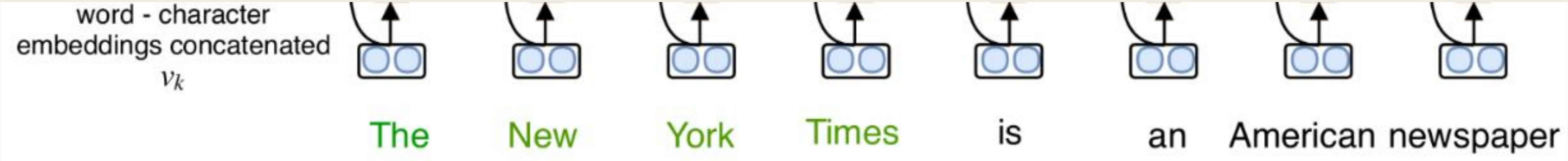
- Formally for EL:
  - **Input:** $D = \{w_1, \dots, w_n\}$, where $w_k \in W$
  - **Output:** list of mention - entity pairs $\{(m_i, e_i)\}_{i \epsilon \overline{1,T}}$
- Formally for ED:
  - **Input:** $D = \{w_1, \dots, w_n\}$, where $w_k \in W$ **+** $\{m_i\}_{i \epsilon \overline{1,T}}$
  - **Output:** $\{e_i\}_{i=\overline{1,T}} \in \varepsilon^T$



The Model Architecture

Step 1 - Word and Char Embeddings

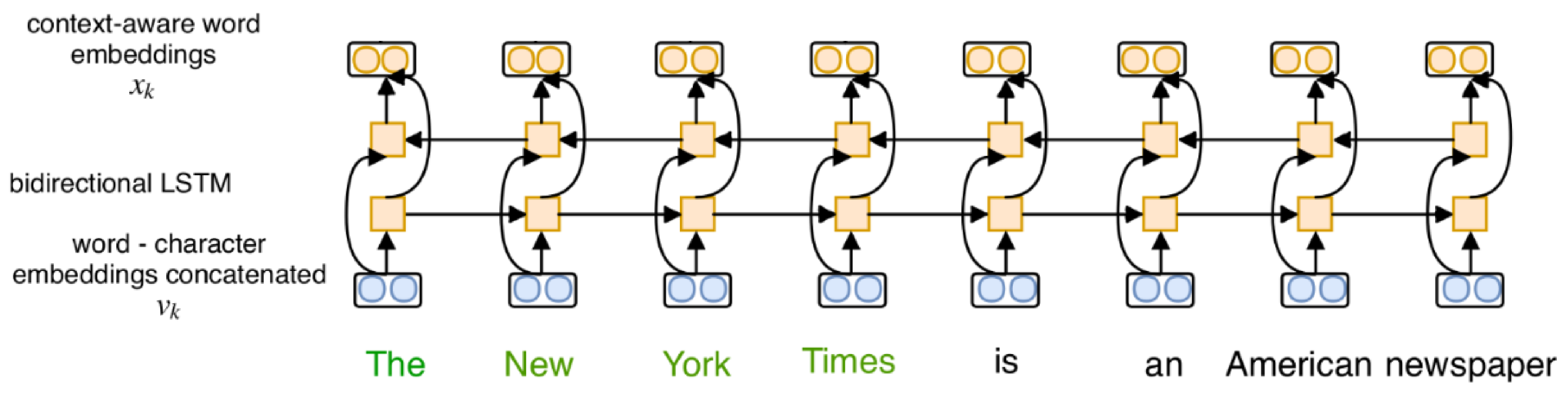
- $\{z_1, \ldots, z_L\}$ - character vectors of word w
- $h_t^f = FWD - LSTM\left(h_{t-1}^f, z_t\right)$
- $h_t^b = BKWD - LSTM\left(h_{t+1}^b, z_t\right)$
- Character embedding of w is $[\boldsymbol{h_L^f}; \boldsymbol{h_1^b}]$
- Character embedding is concatenated with the pre-trained word embedding [2]
  - *Forms the context-independent word-character embedding of w*



[2] Mikolov et al. 2013. Distributed representations of words and phrases and their compositionality.

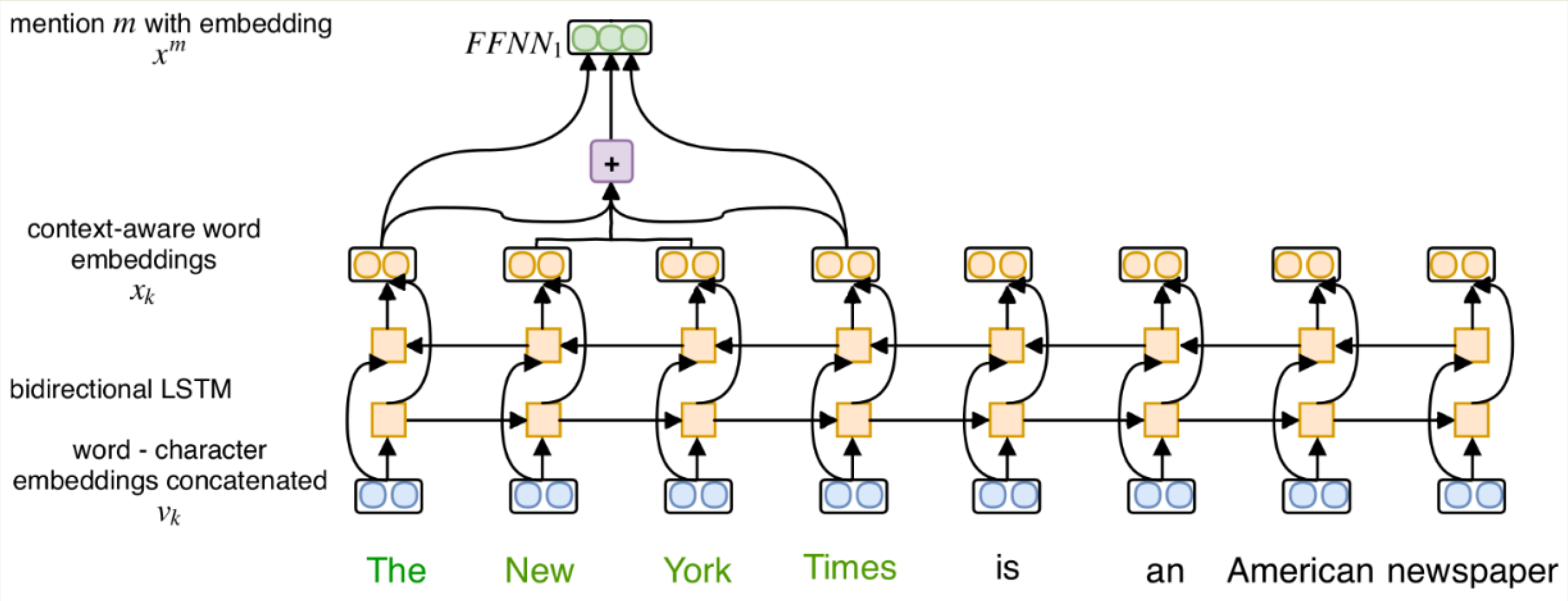
Step 2 – Mention Representation

- To make word embeddings aware of **local context**: bi-LSTM layer!
- Hidden states of the bi-LSTM (corresponding to each word) are concatenated into context-aware word embeddings**.**



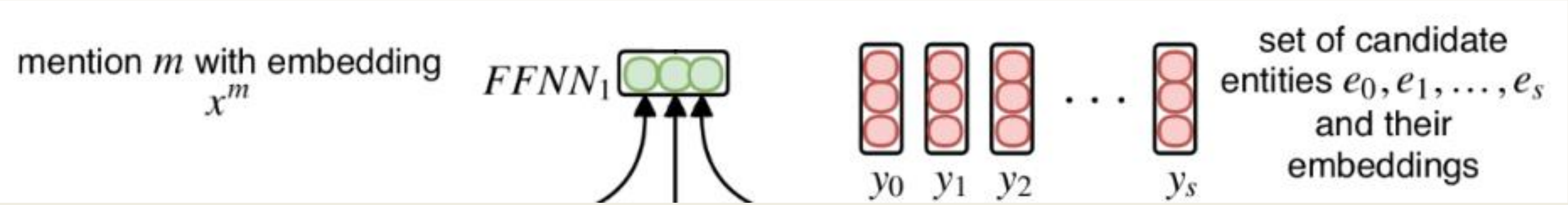
Step 2 – Mention Representation

- For each possible mention $m = w_q, \dots, w_r$, concatenate first, last and "soft head" words
  - *fixed size representation* $g^m = [x_q; x_r; \hat{x}^m]$
- To learn non-linear interactions between the component word vectors:
  - *project* g^m *to a final mention representation* $x^m = FFNN_1(g^m)$



Step 3 - Entity Embeddings
Step 4 - Candidate Selection

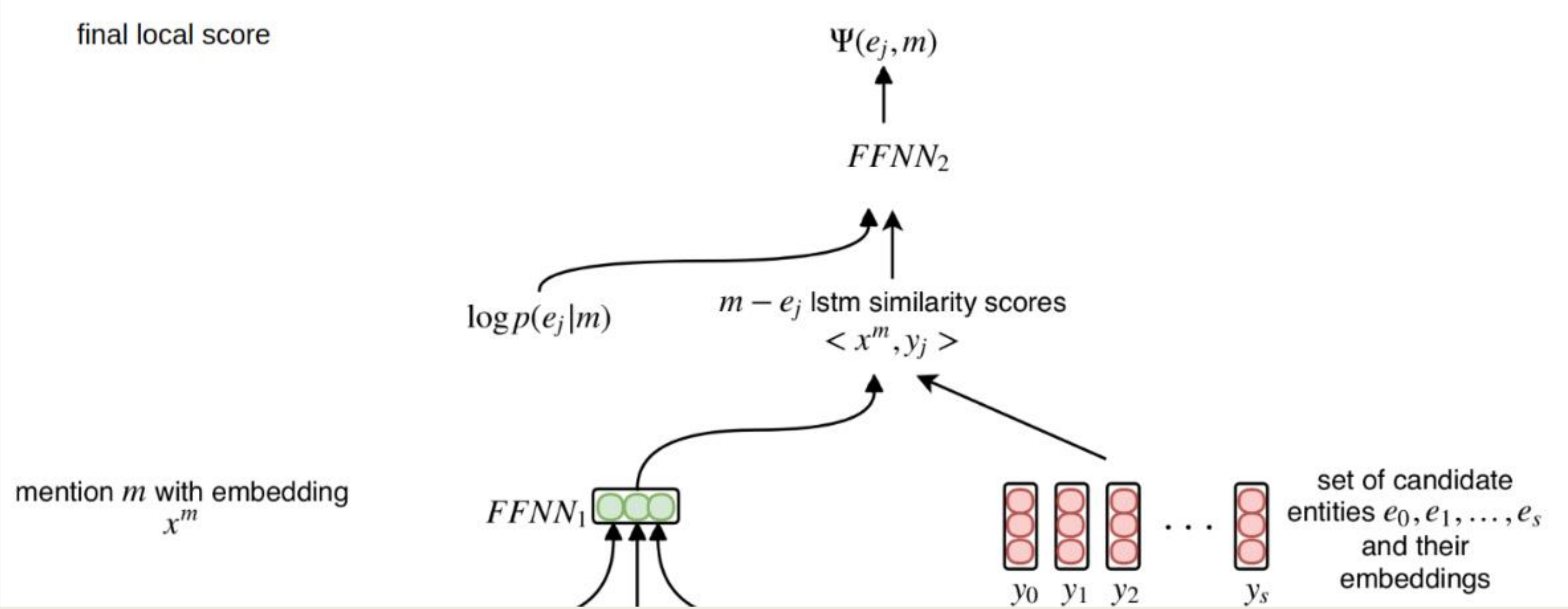
- Use the fixed continuous **pre-trained entity representations** of Ganea and Hofmann [3].
- For each span *m* select the top *s* entity candidates.
  - *Candidate set $C(m)$*
- Top is based on empirical probabilistic entity – map $p(\mathrm{e}|m)$ [3].
  - *From Wikipedia, Crosswikis and YAGO dictionaries*
- $C(m)$ is used at training and test time.



[3] Octavian-Eugen Ganea and Thomas Hofmann. 2017. Deep joint entity disambiguation with local neural attention.

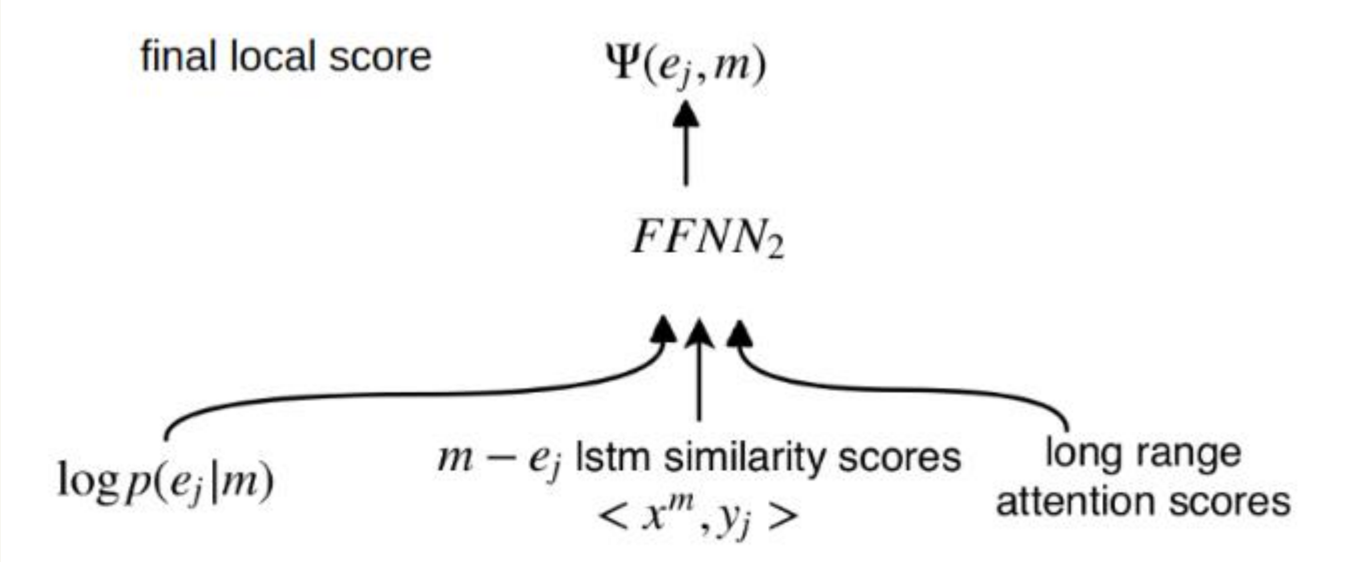
Step 5 – Final Local Score

- Compute **similarity score** using embedding dot-product for each candidate e
  - *When* $|C(m)| \geq 1$
- Combine it with the log-prior probability using a FFNN



Step 5 (Extra!)– Long Range Context Attention

- Improves model in some cases by explicitly capturing long context dependencies.
- Using an attention model [3] collect one context embedding per mention
  - *based on informative context words*
  - *related to at least of the candidate entities*



Step 6 - Training

- Assuming corpus with documents and gold entity - mention pairs $G = \{(m_i, \mathrm{e}_i^*)\}_{i=\overline{1,K}}$
- Collect set M of all (potentially overlapping) token spans m for which $|C(m)| \geq 1$
- Train the parameters using the minimization procedure:

$$\theta^* = \arg\min_{\theta} \sum_{m \in M} \sum_{e \in C(m)} V(\Psi_\theta(e, m))$$

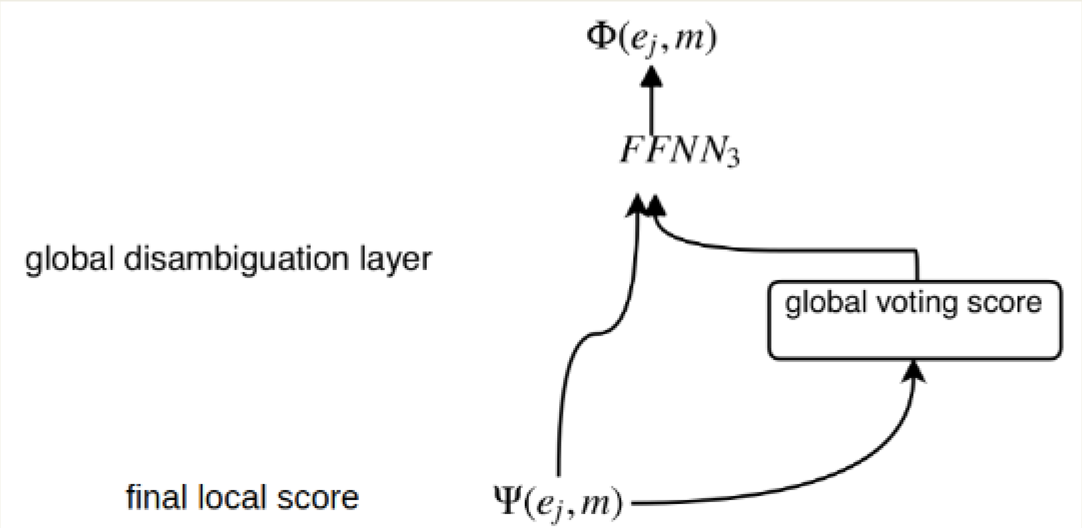
- V enforces the scores of gold pairs to be linearly separable from scores of negative pairs

Extra Step – Global Disambiguation

- Currently, the model performs disambiguation of each candidate span independently.

➢ Add extra layer that promotes coherence among linked and disambiguated entities inside the same document.

1. Define the set of mention-entity pairs that can participate in the global disambiguation voting (that have a **high score**)
2. Calculate **final "global" score** $G(e_j, m) = \cos\left(y_{e_j}, y_G^m\right)$
3. Combine this with the local score:



Coreference Resolution Heuristic

It is important to be able to solve simple coreference resolution cases

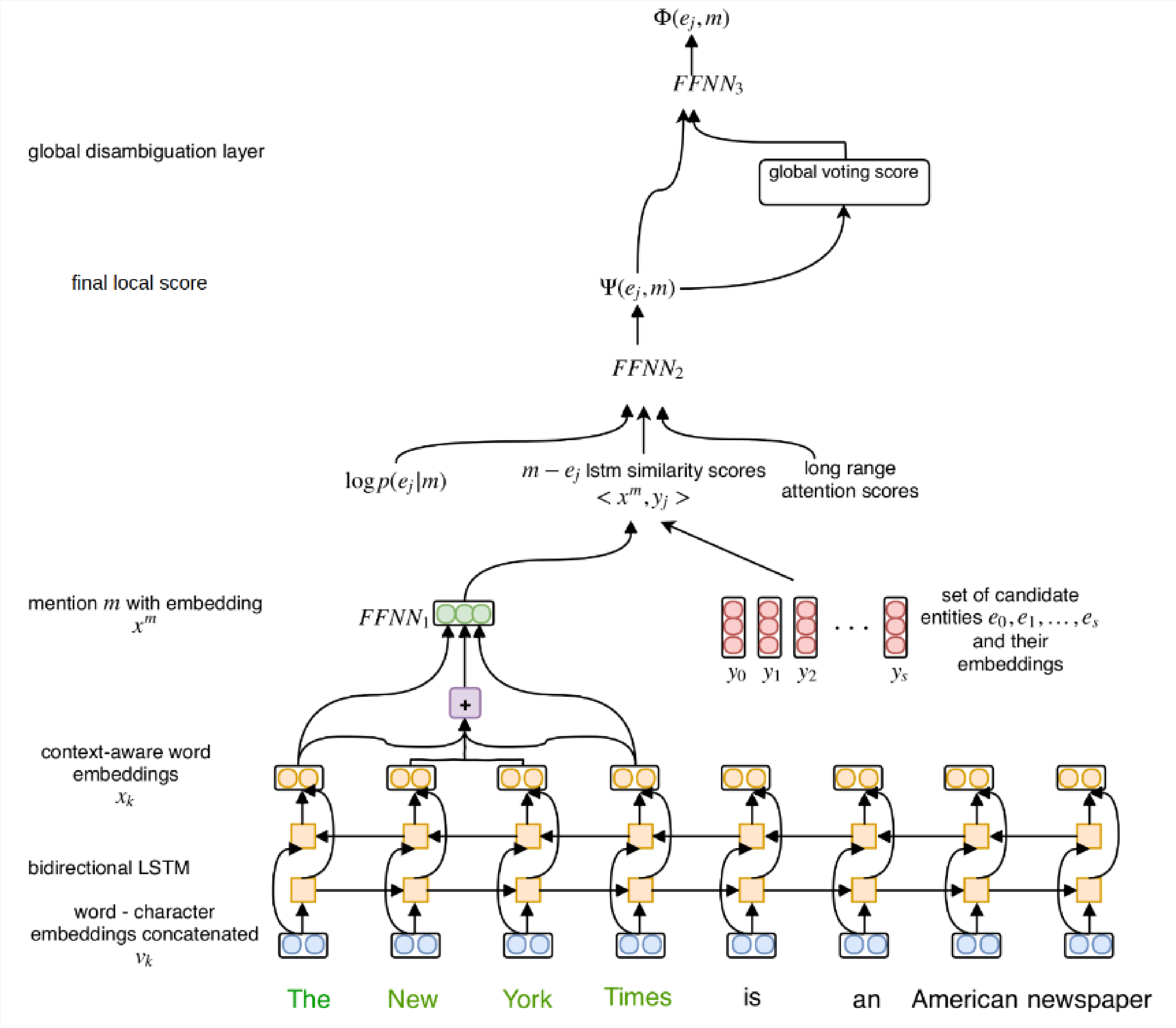
("Alan" referring to "Alan Shearer")

these cases are difficult to handle by the candidate selection strategy

Adopt simple heuristic [3]:

Observed between 0.5% and1% improvement on all datasets

The Model Architecture



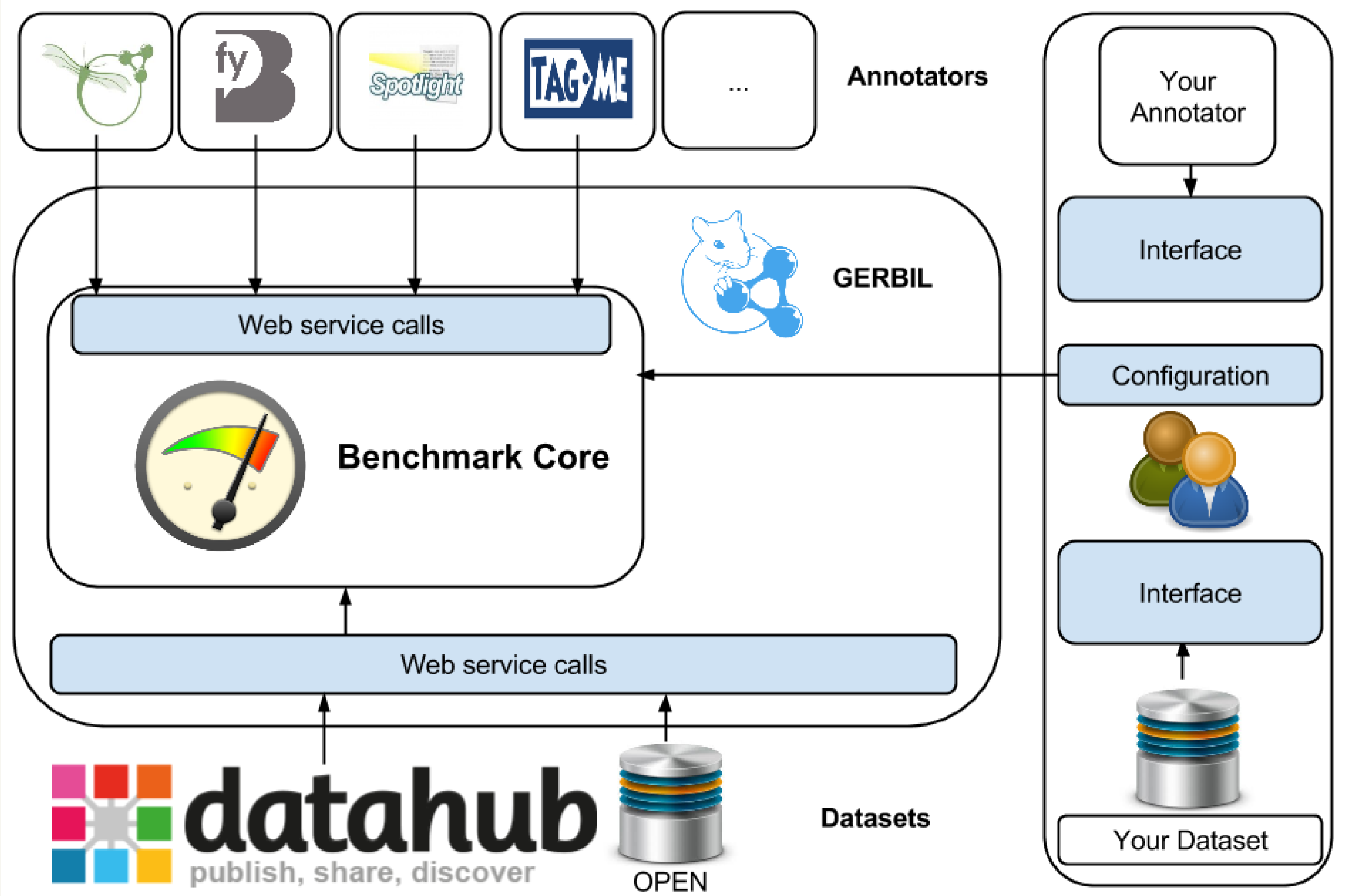
Experiments

- Used Wikipedia 2014 as KB.
- Conducted experiments on the most important public EL datasets using the **Gerbil** platform.
- For training, used the biggest publicly available EL dataset: AIDA/CoNLL

  Training Set
  - *18,448 linked mentions*
  - *946 documents*

  Validation Set
  - *4,791 mentions*
  - *216 documents*

  Test Set
  - *4,485 mentions*
  - *231 documents*



Results

- 4 different models used
  - **Base model**: only uses the mention local score and the log-prior.
  - **Base model + att**: the Base Model plus Long Range Context Attention.
  - **Base model + att + global***:* the Global Model
  - **ED base model + att + global StanfordNER:** ED Global model that runs on top of the detected mentions of the Stanford NER system [4].

[4] Finkel et al. 2005. Incorporating non-local information into information extraction systems by gibbs sampling.

F1@MA F1@MI	AIDA A	AIDA B	MSNBC	OKE-2015	OKE-2016	N3-Reuters-128	N3-RSS-500	Derczynski	KORE50
FREME	23.6 37.6	23.8 36.3	15.8 19.9	26.1 31.6	22.7 28.5	26.8 30.9	32.5 27.8	31.4 18.9	12.3 14.5
FOX	54.7 58.0	58.1 57.0	11.2 8.3	53.9 56.8	49.5 50.5	52.4 53.3	35.1 33.8	42.0 38.0	28.3 30.8
Babelfy	41.2 47.2	42.4 48.5	36.6 39.7	39.3 41.9	37.8 37.7	19.6 23.0	32.1 29.1	28.9 29.8	52.5 55.9
Entityclassifier.eu	43.0 44.7	42.9 45.0	41.4 42.2	29.2 29.5	33.8 32.5	24.7 27.9	23.1 22.7	16.3 16.9	25.2 28.0
Kea	36.8 40.4	39.0 42.3	30.6 30.9	44.6 46.2	46.3 46.4	17.5 18.1	22.7 20.5	31.3 26.5	41.0 46.8
DBpedia Spotlight	49.9 55.2	52.0 57.8	42.4 40.6	42.0 44.4	41.4 43.1	21.5 24.8	26.7 27.2	33.7 32.2	29.4 34.9
AIDA	68.8 72.4	71.9 72.8	62.7 65.1	58.7 63.1	0.0 0.0	42.6 46.4	42.6 42.4	40.6 32.6	49.6 55.4
WAT	69.2 72.8	70.8 73.0	62.6 64.5	53.2 56.4	51.8 53.9	45.0 49.2	45.3 42.3	44.4 38.0	37.3 49.6
Best baseline	**69.2** **72.8**	**71.9** **73.0**	**62.7** **65.1**	**58.7** **63.1**	**51.8** **53.9**	**52.4** **53.3**	**45.3** **42.4**	**44.4** **38.0**	**52.5** **55.9**
base model	86.6 89.1	81.1 80.5	64.5 65.7	54.3 58.2	43.6 46.0	47.7 49.0	44.2 38.8	43.5 38.1	34.9 42.0
base model + att	86.5 88.9	81.9 82.3	69.4 69.5	56.6 60.7	49.2 51.6	48.3 51.1	46.0 40.5	47.9 42.3	36.0 42.2
base model + att + global	86.6 89.4	82.6 82.4	73.0 72.4	56.6 61.9	47.8 52.7	45.4 50.3	43.8 38.2	43.2 34.1	26.2 35.2
ED base model + att + global using Stanford NER mentions	75.7 80.3	73.3 74.6	71.1 71.0	62.9 66.9	57.1 58.4	54.2 54.6	45.9 42.2	48.8 42.3	40.3 46.0

Results

Highlighted the best and second best models for the EL task.

Metrics computed in the **strong matching** setting:

- Requires the exact prediction of the gold mention boundaries and their entity annotations.

Main Goal Achieved!

- The joint EL offers the best model!
  - If enough training data is available with the same characteristics as the test data.
  - True not only when training on AIDA, but also for other types of datasets (queries, tweets)

- When testing data has different statistics than the training data:
  - Method works best with a state-of-the-art NER system.

Conclusion

Presented the first neural end-to-end entity linking model

Showed the benefit of jointly optimizing entity recognition and linking

Proved that engineered features can be almost completely replaced by modern neural networks

Q&A