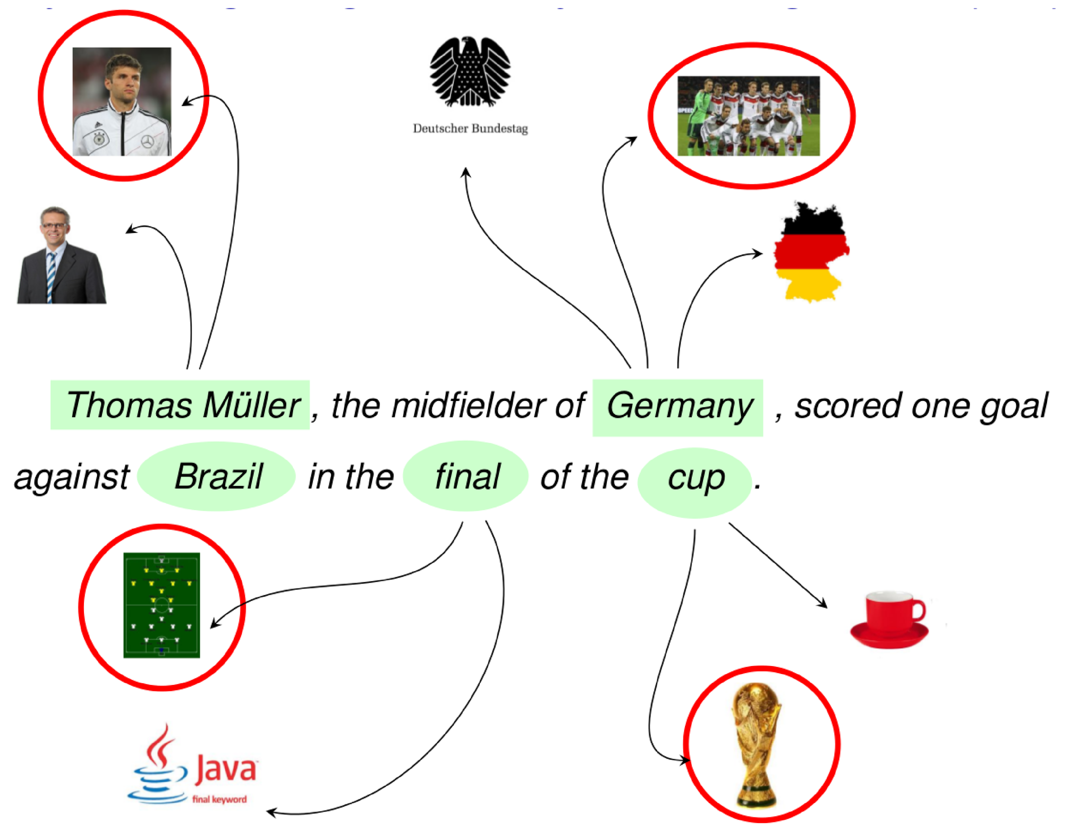
Deep Joint Entity Disambiguation with Local Neural Attention (EMNLP'17)

https://arxiv.org/abs/1704.04920

Octavian-Eugen Ganea, Thomas Hofmann

Department of Computer Science, ETH Zürich, Switzerland

Entity Disambiguation (ED): Problem Definition



Goal: leverage only neural features, no engineered (except $p(e|m)$)

1) Fast-Trainable Competitive Entity Embeddings

- Same space as word vectors (use pre-trained Word2Vec)
- $\hat{p}(w|e)$ estimated from : i) canonical pages ii) windows of hyperlinks
- Positive word distribution: $w^+ \sim \hat{p}(w|e)$
- Negative word distribution: $w^- \sim q(w) = \hat{p}(w)^\alpha$ for $\alpha \in (0,1)$
- Loss function (independent per each entity):

$$J(\mathbf{z}; e) = \mathbb{E}_{w^+|e}\, \mathbb{E}_{w^-}\left[\max\left(0, \gamma - \langle \mathbf{z}, \mathbf{x}_{w^+} - \mathbf{x}_{v^-}\rangle\right)\right]$$

$$\mathbf{x}_e = \arg\min_{\mathbf{z}:\|\mathbf{z}\|=1} J(\mathbf{z}; e)$$

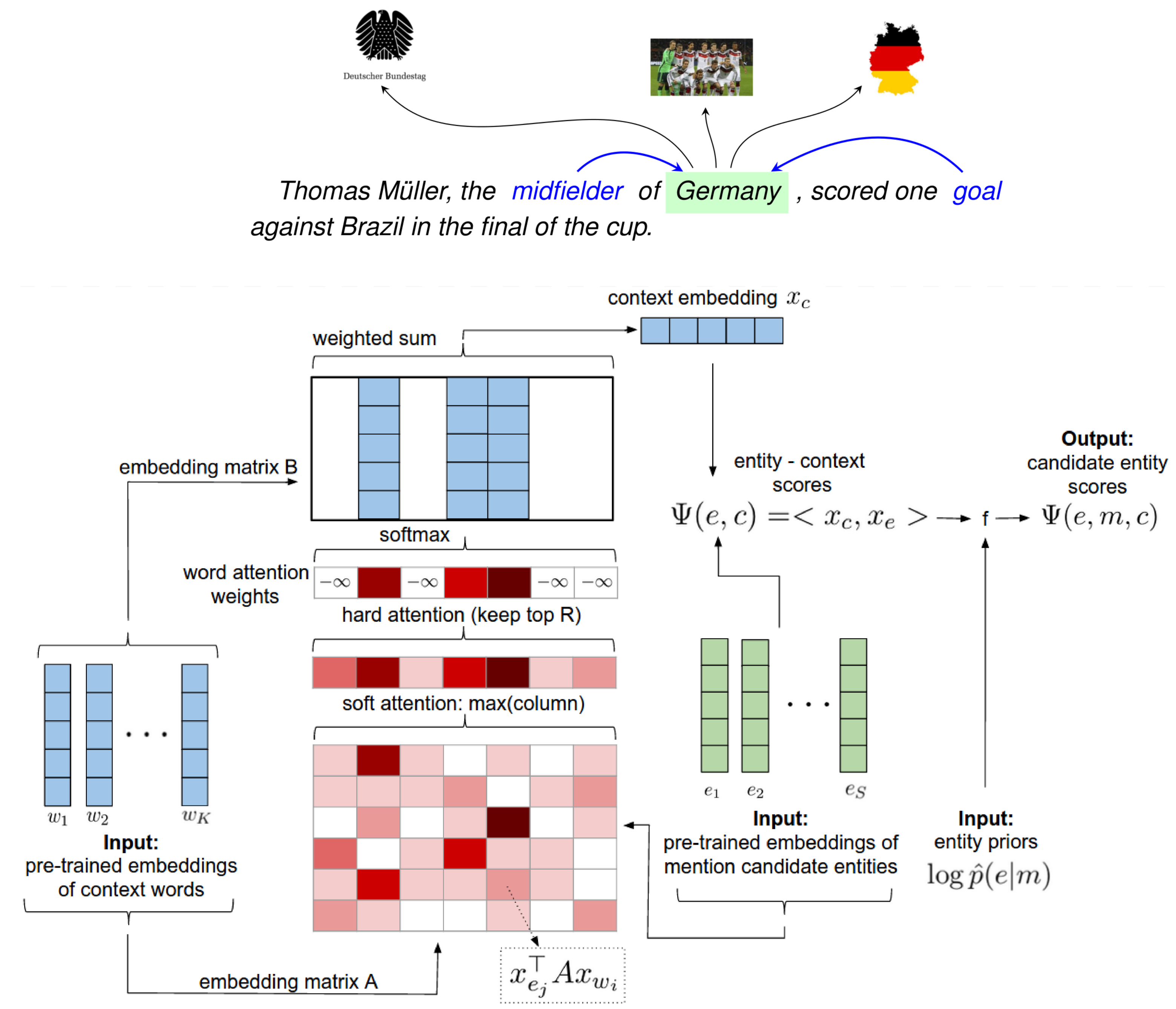
Advantages:

- Avoids entity co-occurrence counts (suffer from sparsity, memory footprint, rare entities)
- Any subset of entities can be trained (no need for full set of entities)
- Entities can join later without full retraining

Method Metric	NDCG@1	NDCG@5	NDCG@10	MAP
WikiLinkMeasure	0.54	0.52	0.55	0.48
(Yamada,2016)	0.59	0.56	0.59	0.52
our (canonical pages)	0.624	0.589	0.615	0.549
our (canonical+hyperlinks)	**0.632**	**0.609**	**0.641**	**0.578**

Table: Entity relatedness results on the test set of (Ceccarelli, 2013).

2) Local Neural Attention Mechanism



Loss: $L(\theta) = \sum_{D\in\mathcal{D}} \sum_{m_i\in D} \sum_{e\in\Gamma(m_i)} [\gamma - \Psi(e_i^*, m_i, c_i) + \Psi(e, m_i, c_i)]_+$

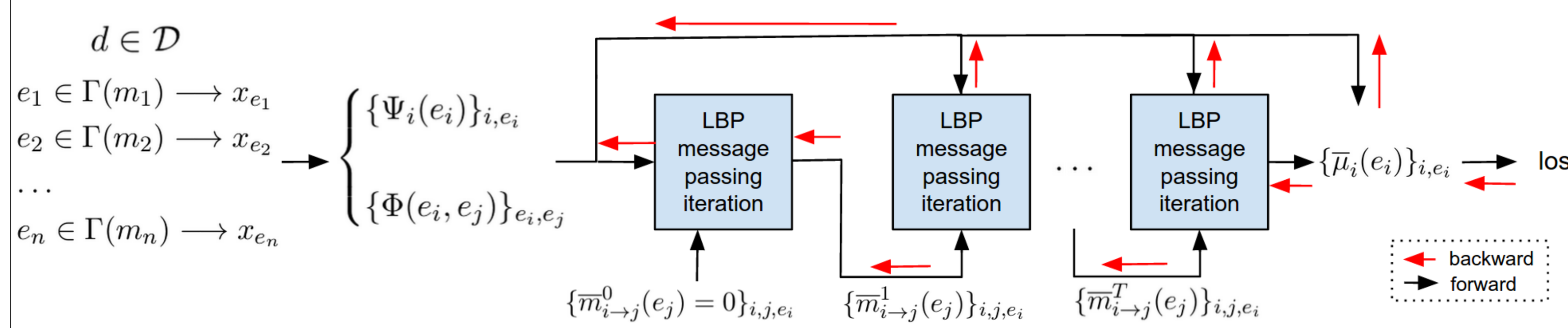
3) Differentiable Joint Inference Stage

i) Fully-connected pairwise CRF (log-scale):

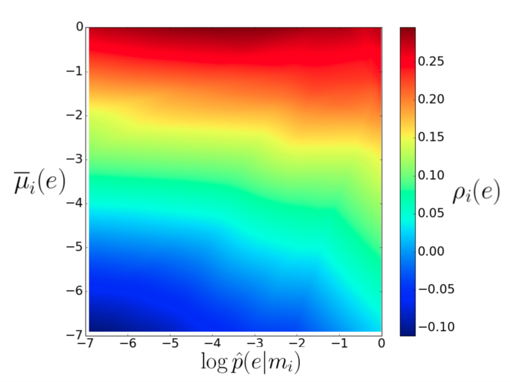
$$g(\mathbf{e}, \mathbf{m}, \mathbf{c}) = \sum_{i=1}^{n} \Psi_i(e_i) + \sum_{i<j} \Phi(e_i, e_j)$$

- unary factors: $\Psi_i(e_i) = \Psi(e_i, c_i)$ (context - entity similarity from above)
- pairwise factors: $\Phi(e, e') = \frac{2}{n-1} \mathbf{x}_e^\top \mathbf{C}\, \mathbf{x}_{e'}$,

ii) Unrolled loopy belief propagation (LBP) in a deep neural network:



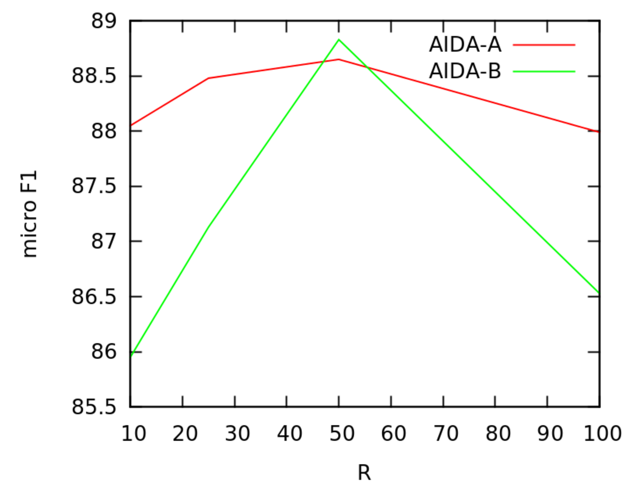
iii) Approximate LBP marginals combined with the mention prior: $\rho_i(e) := f(\overline{\mu}_i(e), \log \hat{p}(e|m_i))$

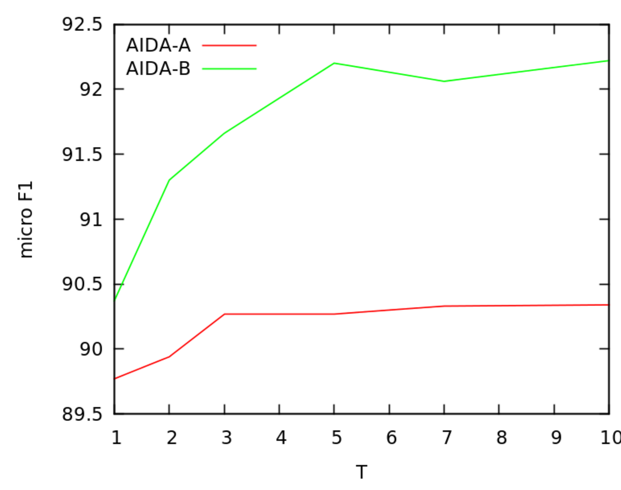


iv) Loss: $L(\theta) = \sum_{D\in\mathcal{D}} \sum_{m_i\in D} \sum_{e\in\Gamma(m_i)} [\gamma - \rho_i(e_i^*) + \rho_i(e)]_+$

Experiments (More in the Paper)

Methods	AIDA-B
Local models	
prior $\hat{p}(e|m)$	71.9
(Lazic, 2015) (Plato)	86.4
(Globerson, 2016)	87.9
(Yamada, 2016)	87.2
our (local)	**88.8**
Global models	
(Huang, 2015)	86.6
(Ganea, 2016)	87.6
(Chisholm, 2015)	88.7
(Guo, 2016)	89.0
(Globerson, 2016)	91.0
(Yamada, 2016)	91.5
our (global)	**92.22 ± 0.14**





Global methods	MSB	AQ	ACE	CWEB	WW
prior $\hat{p}(e|m)$	89.3	83.2	84.4	69.8	64.2
(Fang, 2016)	81.2	88.8	85.3	-	-
(Ganea, 2016)	91	89.2	**88.7**	-	-
(Milne, 2008)	78	85	81	64.1	81.7
(Hoffart, 2011)	79	56	80	58.6	63
(Ratinov, 2011)	75	83	82	56.2	67.2
(Cheng, 2013)	90	**90**	86	67.5	73.4
(Guo, 2016)	92	87	88	77	**84.5**
our (global)	**93.7 ± 0.1**	88.5 ± 0.4	**88.5 ± 0.3**	**77.9 ± 0.1**	77.5 ± 0.1

Table: Micro F1 results for other datasets.

Frequency gold entity	Number mentions	Solved correctly
0	5	80.0 %
1-10	0	-
11-20	4	100.0%
21-50	50	90.0%
> 50	4345	94.2%

$\hat{p}(e|m)$ **gold entity**	Number mentions	Solved correctly
≤ 0.01	36	89.19%
0.01 - 0.03	249	88.76%
0.03 - 0.1	306	82.03%
0.1 - 0.3	381	86.61%
> 0.3	3431	96.53%

Table: ED accuracy on AIDA-B for our best system splitted by Wikipedia hyperlink frequency and mention prior of the gold entity.