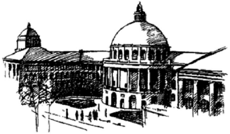


Deep Joint Entity Disambiguation with Local Neural Attention

Octavian Ganea   Thomas Hofmann

Department of Computer Science
ETH Zürich, Switzerland

Conference on Empirical Methods in Natural Language Processing, EMNLP 2017

Today's Menu

Text disambiguation system based on:

- entity embeddings (cheaply trained)
- a neural attention mechanism over local context windows
- a CRF equipped with a differentiable inference procedure

Entity Disambiguation (ED)

Thomas Müller, the midfielder of Germany, scored one goal against Brazil in the final of the cup.

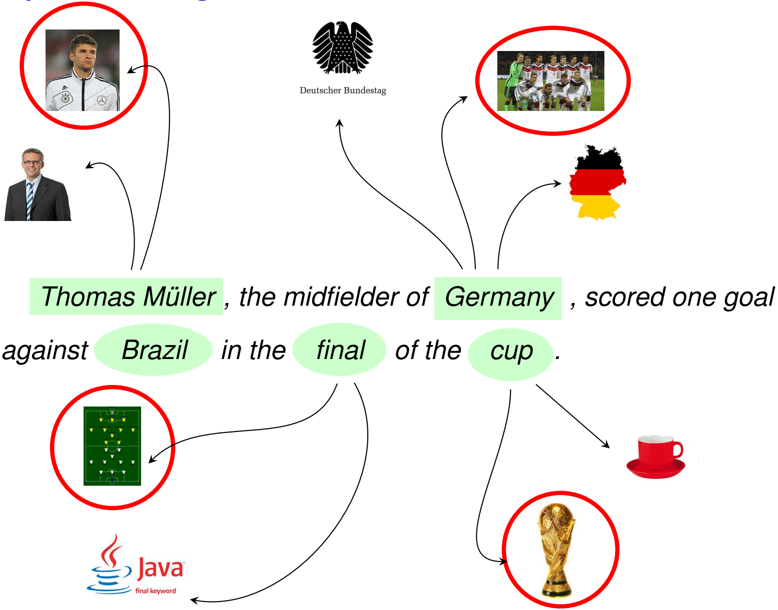
Entity Disambiguation (ED)

Deutscher Bundestag

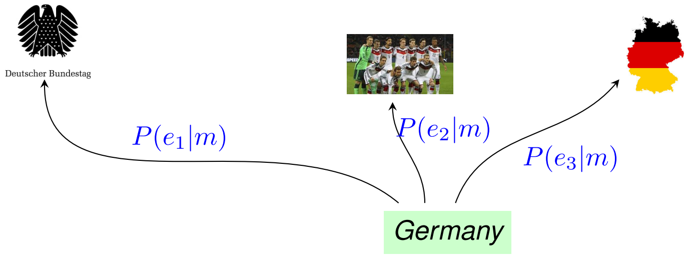
Thomas Müller , the midfielder of Germany , scored one goal against Brazil in the final of the cup .

Java
final keyword

Entity Disambiguation (ED)



Key Component: Mention - Entity Compatibility



- Mention - entity prior $p(e|m)$
- Estimated from statistics of Wikipedia hyperlinks

$$P(e|m) \approx \frac{\text{\# links with } m \text{ that point to } e}{\text{\# links with anchor } m}$$

- Trivial baseline: $e_i^* = \arg\max_{e \in \mathcal{E}} P(e_i|m_i)$
- Used for: i) scoring, ii) candidate selection

Key Component: Entity Embeddings (1)

- Same space as pre-trained word vectors (Word2Vec)
- Positive word distribution: $w^{+} \sim \hat{p}(w|e)$
  - estimated from context windows around occurrences of e
- Negative word distribution: $w^{-} \sim q(w) = \hat{p}(w)^{\alpha}$ for $\alpha \in (0,1)$
- Loss function:

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

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

Key Component: Entity Embeddings (2)

Advantages:

- Avoids entity co-occurrence sparse counts (as opposed to prior work)
- Only a subset of entities can be trained
- Rare entities
- Works well in practice

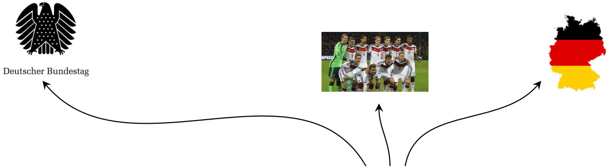
Method Metric	NDCG@1	NDCG@5	NDCG@10	MAP
WikiLinkMeasure [**MW08**]	0.54	0.52	0.55	0.48
(Yamada et al. 2016): d = 500	0.59	0.56	0.59	0.52
our (canonical pages): d = 300	0.624	0.589	0.615	0.549
our (canonical&hyperlinks): d = 300	**0.632**	**0.609**	**0.641**	**0.578**

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

Entity Embeddings - Examples

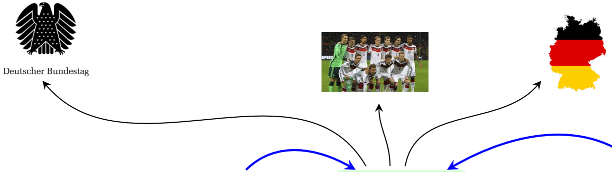
Entity	Closest words sorted by cosine similarity
Japan national football team	Japan player Shizuoka Yokohama played Asian USISL Saitama Okada Nakamura Tokyo Pele matches Japanese Korea players Tanaka soccer Chunnam game Suwon Takuya Kawaguchi Mizuno match Qatar team Eto Eiji football playing Confederations tournament Kagawa Chiba
Apple	apple fruit berry grape varieties apples crop pear potato blueberry strawberry growers peach orchards pears Prunus grower Rubus citrus spinosa tomato berries Blueberry peaches grapes almond juice melon bean apricot insect vegetable strawberries olive pomegranate Vaccinium cherries potatoes
Apple Inc.	Apple software computer Microsoft Adobe hardware company iPod PC product Dell laptop Mac computers Macintosh Flash video desktop iPhone Digital Windows app PCs Intel technology device iTunes Motorola Sony digital Multimedia iPad HP
Queen (band)	U2 band singer Avenged Rockers Coldplay concert Lynyrd Kiss Metallica Killers rerecorded song Beatles rock Stones recording Slash Singer touring musician music CD Dirty Moby rockers
Leicestershire	curacy town Yeomanry Buckinghamshire Leicestershire Bedfordshire Lichfield Wiltshire Shropshire almshouses Lancashire Stonyhurst
Leicestershire County Cricket Club	Warwickshire batsman England Hampshire Leicestershire Trott Glamorgan Nottinghamshire Northants Lancashire Middlesex Essex Giles fielding Porterfield Test Surrey cricketer centurion Gough Bevan Sussex Gloucestershire bowled Worcestershire Tests Martyn Croft Derbyshire Clarke overs bowler Lancastrian played Northamptonshire Kent Vaughan Fletcher captaining

Local Disambiguation - Idea



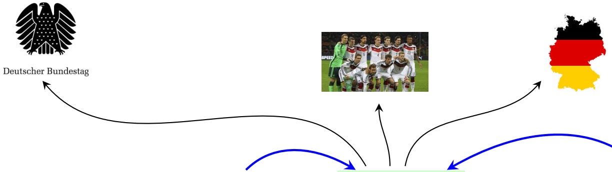
Thomas Müller, the midfielder of Germany , scored one goal against Brazil in the final of the cup.

Local Disambiguation - Idea



Thomas Müller, the midfielder of Germany , scored one goal against Brazil in the final of the cup.

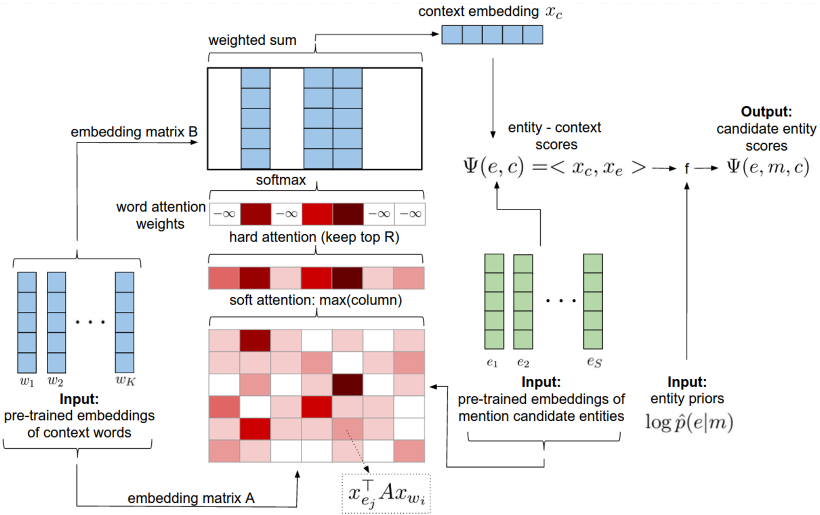
Local Disambiguation - Idea



Thomas Müller, the midfielder of Germany, scored one goal against Brazil in the final of the cup.

- Context = text window around mention (of size K)
- Bag-of-words context model.
- Idea: only few words are informative

Local Disambiguation - Model



Local Disambiguation - Model

- Final model:
  - neural network w/ 100 hidden units and ReLU
  - norm bound regularization

$$\Psi(e, m, c) = f(\Psi(e, c), \log \hat{p}(e|m))$$

- Max-margin loss:

$$\theta^* = \arg\min_{\theta} \sum_{m} \sum_{e \in \Gamma(m)} [\gamma - \Psi(e^*, m, c) + \Psi(e, m, c)]_+$$

Local Disambiguation - Attended Words

Mention	Gold entity	$\hat{p}(e\|m)$ **of gold entity**	**Attended contextual words**
Scotland	Scotland national rugby union team	0.034	England Rugby team squad Murrayfield Twickenham national play Cup Saturday World game George following Italy week Friday selection dropped row month
Wolverhampton	Wolverhampton Wanderers F.C.	0.103	matches League Oxford Hull league Charlton Oldham Cambridge Sunderland Blackburn Sheffield Southampton Huddersfield Leeds Middlesbrough Reading Coventry Darlington Bradford Birmingham Enfield Barnsley
Montreal	Montreal Canadiens	0.021	League team Hockey Toronto Ottawa games Anaheim Edmonton Rangers Philadelphia Caps Buffalo Pittsburgh Chicago Louis National home Friday York Dallas Washington Ice
Santander	Santander Group	0.192	Carlos Telmex Mexico Mexican group firm market week Ponce debt shares buying Televisa earlier pesos share stepped Friday analysts ended
World Cup	FIS Alpine Ski World Cup	0.063	Alpine ski national slalom World Skiing Whistler downhill Cup events race consecutive weekend Mountain Canadian racing

Effects of Hyperparemeters

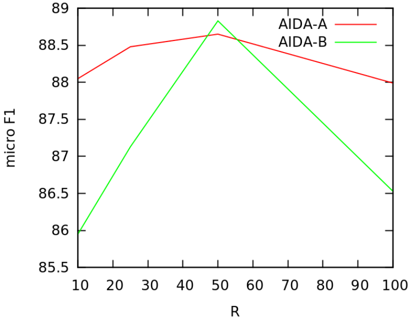
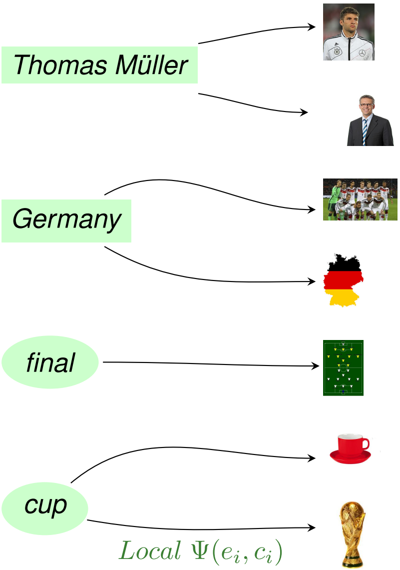


Table: Hard attention improves accuracy of a local model with K=100.

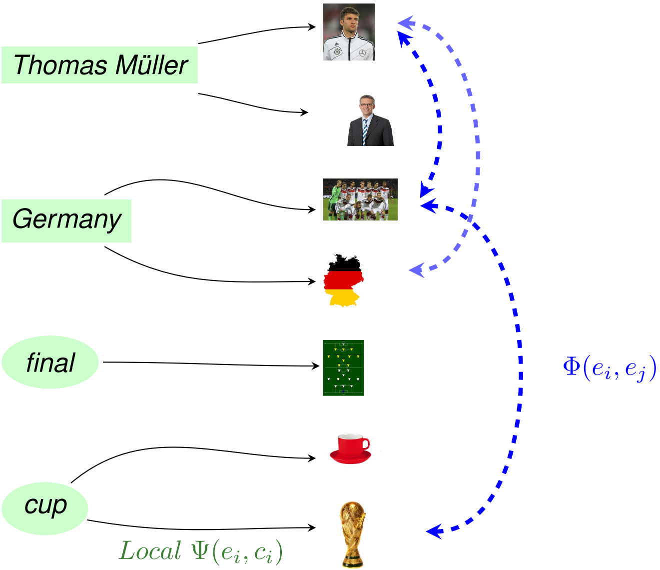
Global Disambiguation

Thomas Müller, *the midfielder of* *Germany* , *scored one goal against* *Brazil* *in the* *final* *of the* *cup* .

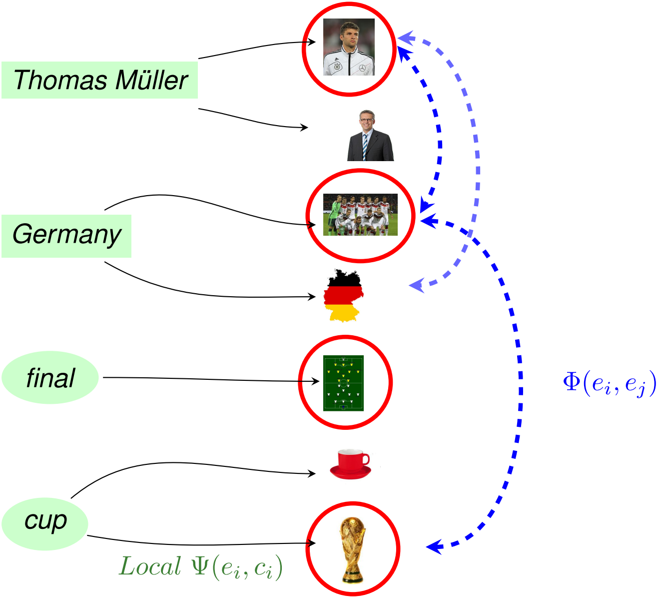
Global Disambiguation



Global Disambiguation



Global Disambiguation



Global Disambiguation - Model

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

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

Global Disambiguation - Model

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

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

- $\Psi(e_i, c_i)$ - local scores (w/o mention prior)

Global Disambiguation - Model

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

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

- $\Psi(e_i, c_i)$ - local scores (w/o mention prior)
- $\Phi(e_i, e_j) = \frac{2}{n-1} \, \mathbf{x}_{e_i}^{\top} \mathbf{C} \, \mathbf{x}_{e_j} \, ,$

Global Disambiguation - Model

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

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

- $\Psi(e_i, c_i)$ - local scores (w/o mention prior)
- $\Phi(e_i, e_j) = \dfrac{2}{n-1} \; \mathbf{x}_{e_i}^{\top} \mathbf{C} \, \mathbf{x}_{e_j}$,
- Mention prior gets combined with the approximate marginals.
- Jointly solve: **e** $\in \Gamma(m_1) \times \cdots \times \Gamma(m_n)$

Training and Prediction in Loopy Graphical Models

- Traditionally:
  - learning via maximum likelihood
  - prediction via approximate inference (e.g. message passing)

Training and Prediction in Loopy Graphical Models

- Traditionally:
  - learning via maximum likelihood
  - prediction via approximate inference (e.g. message passing)
- Problems:
  - intractable partition function

Training and Prediction in Loopy Graphical Models

- Traditionally:
  - learning via maximum likelihood
  - prediction via approximate inference (e.g. message passing)
- Problems:
  - intractable partition function
  - approximation errors of the inference algorithm are not captured during training
    - unless training with inner loop inference for each example - slow

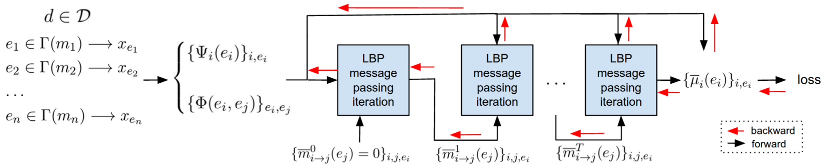
Training and Prediction in Loopy Graphical Models

- Traditionally:
  - learning via maximum likelihood
  - prediction via approximate inference (e.g. message passing)

- Problems:
  - intractable partition function
  - approximation errors of the inference algorithm are not captured during training
    - unless training with inner loop inference for each example - slow
  - model mis-specification (too strong assumptions, thus no *true* parameters)

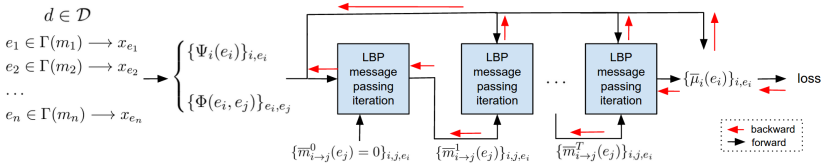
Global ED - Differentiable Inference Procedure

- *"truncated fitting"* (Domke, 2013) of (loopy) belief propagation
- directly optimize approximate marginals used for prediction; maximize their accuracy



Global ED - Differentiable Inference Procedure

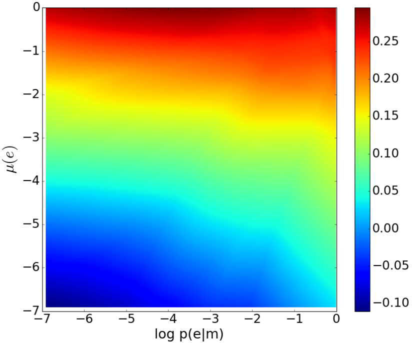
- *"truncated fitting"* (Domke, 2013) of (loopy) belief propagation
- directly optimize approximate marginals used for prediction; maximize their accuracy



- same model for learning and prediction
- much faster compared to double-loop likelihood training
- one of the first to investigate differentiable message passing in NLP
- (Domke, 2013): marginal-based loss functions are more resistant to model mis-specification

Global Disambiguation - Adding Mention Prior

- Approximate marginals: $\rho_i(e) := f(\overline{\mu}_i(e), \log \hat{p}(e|m_i))$
- Same max-margin loss as for the local model.



Experiments

Dataset	Number mentions	Number docs	Mentions per doc	Gold recall
AIDA-train	18448	946	19.5	-
AIDA-A (valid)	4791	216	22.1	96.9%
AIDA-B (test)	4485	231	19.4	98.2%
MSNBC	656	20	32.8	98.5%
AQUAINT	727	50	14.5	94.2%
ACE2004	257	36	7.1	90.6%
WNED-CWEB	11154	320	34.8	91.1%
WNED-WIKI	6821	320	21.3	92%

Table: ED datasets. *Gold recall* is the percentage of mentions for which the entity candidate set contains the ground truth entity.

Experiments

Methods	AIDA-B
Local models	
prior $\hat{p}(e\|m)$	71.9
(Lazic et al., 2015)	86.4
(Yamada et al. 2016)	87.2
(Globerson et al. 2016)	87.9
our (local, K=100, R=50)	**88.8**
Global models	
Huang et al. 2015)	86.6
(Ganea et al. 2016)	87.6
(Chisholm et al. 2015)	88.7
(Guo and Barbosa, 2016)	89.0
(Globerson et al. 2016)	91.0
(Yamada et al. 2016)	91.5
our (global)	**92.22** $\pm$ **0.14**

Table: In-KB accuracy for AIDA-B test set. All baselines use KB+YAGO mention-entity index.

Experiments

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

Table: Micro F1 results for other datasets.

Effects of Hyperparemeters

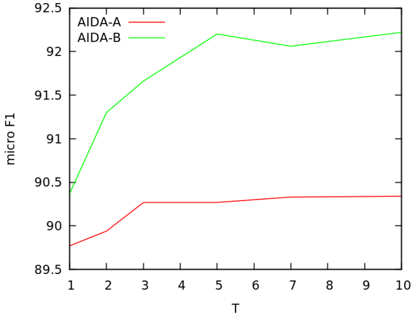


Table: A low T (e.g.5) is already sufficient for accurate approximate marginals.

Infrequent / Low Prior Entities

Freq 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%

References

Domke, Justin (2013)
Learning graphical model parameters with approximate marginal inference

Yamada, Ikuya et al. (2016)
Joint Learning of the Embedding of Words and Entities for Named Entity Disambiguation

Ceccarelli, Diego et al. (2013)
Learning relatedness measures for entity linking

Ferragina, Paolo and Scaiella, Ugo (2010)
Tagme: on-the-fly annotation of short text fragments (by wikipedia entities)

Hoffart, Johannes et al. (2011)
Robust disambiguation of named entities in text

Guo, Zhaochen and Barbosa, Denilson (2014)
Robust Entity Linking via Random Walks

Milne, David and Witten, Ian H (2008)
Learning to link with wikipedia

Ratinov, Lev et al. (2011)
Local and global algorithms for disambiguation to wikipedia

Thank you!