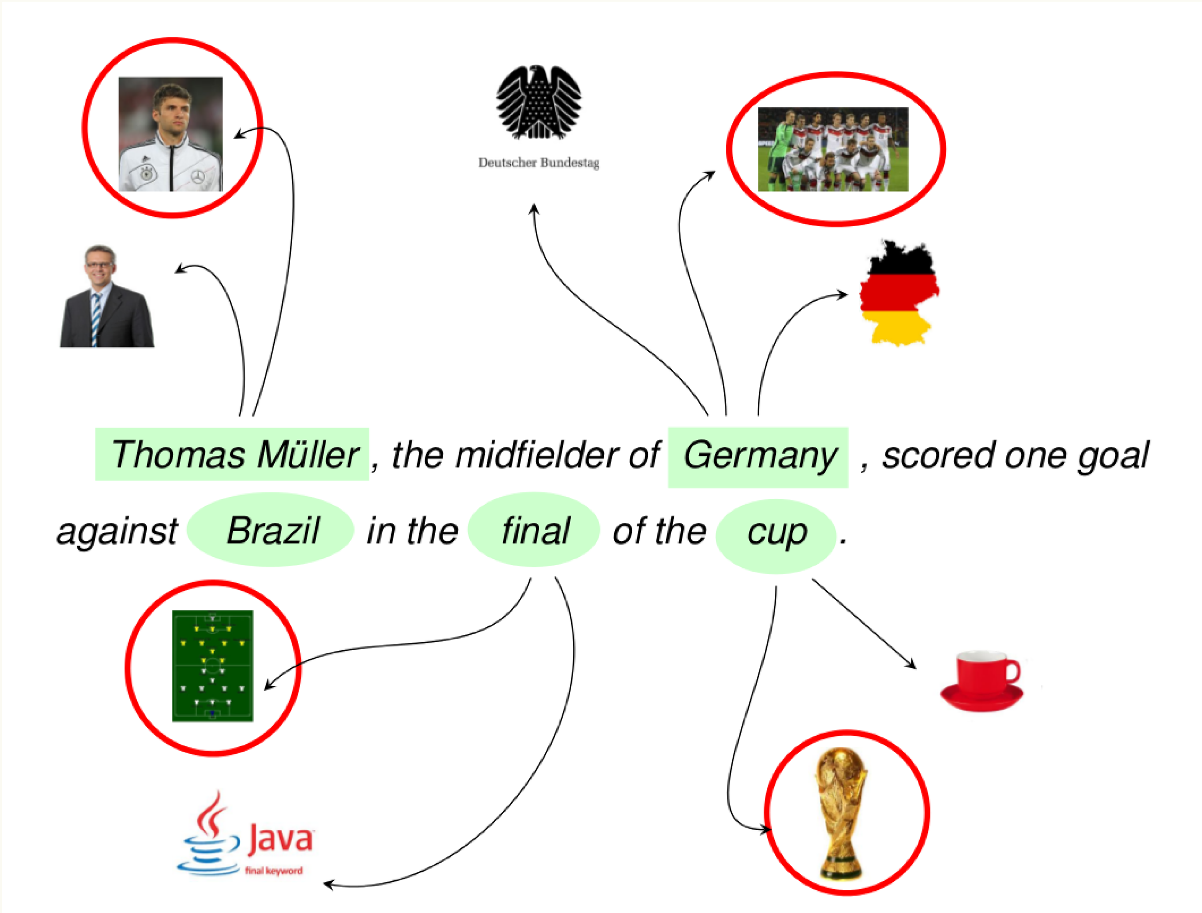
Probabilistic Bag-Of-Hyperlinks Model for Entity Linking

Octavian-Eugen Ganea · Marina Ganea · Aurelien Lucchi · Carsten Eickhoff · Thomas Hofmann

WWW '16, Proceedings of the 25th International Conference on World Wide Web

Entity Disambiguation



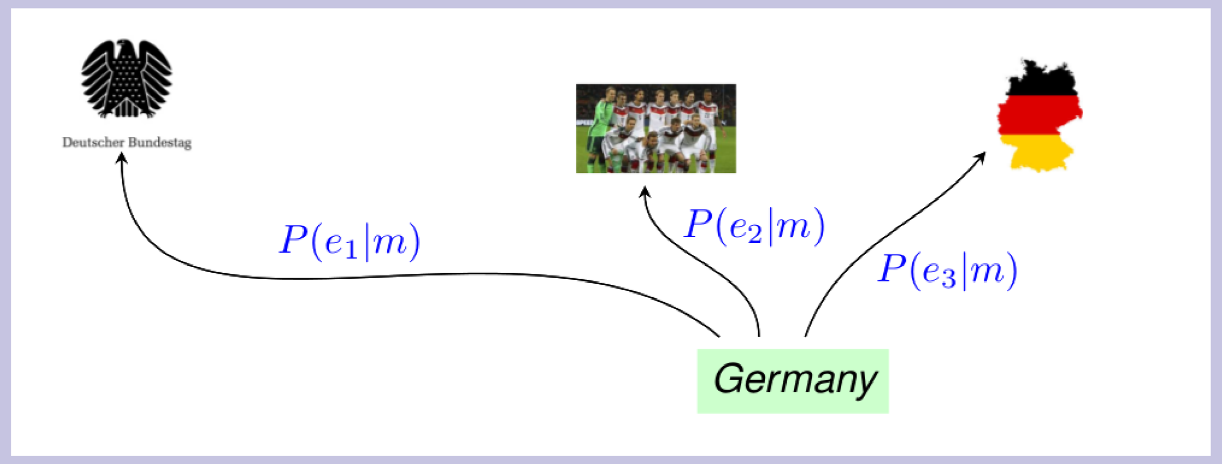
Goal

Clean and effective solution for Entity Disambiguation

Contributions:

- ✔ Rigorous probabilistic semantics for ED.
- ✔ No engineered features. Model explains observed sufficient statistics.
- ✔ Very fast training - scalable to massive data.
- ✔ Sufficiently fast prediction method for real-time usage.
- ✔ Easy to reproduce.
- ✔ Competitive/state-of-the-art performance.
- ✔ Generalizes well on different unseen datasets.

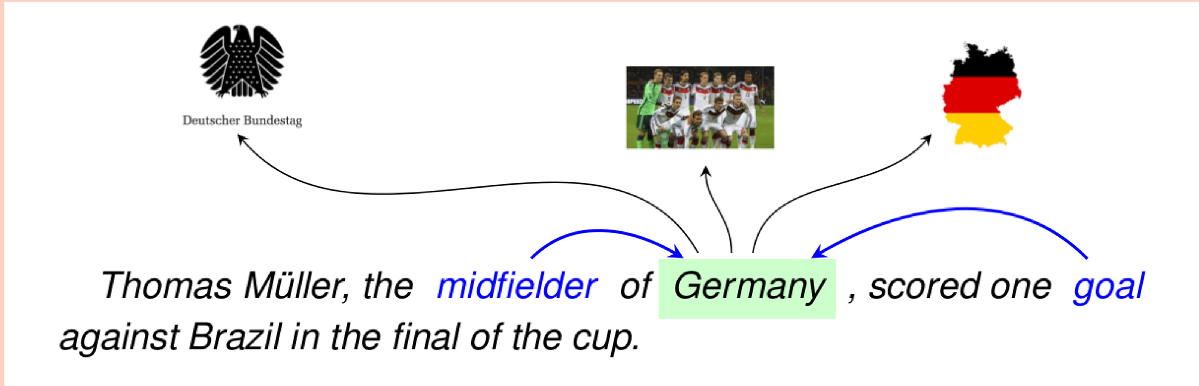
Mention - Entity Prior



- Strong baseline: $e_i^* = \arg\max_{e \in \mathcal{E}} P(e_i|m_i)$
- Estimated from Wikipedia + Crosswikis:

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

Local Context



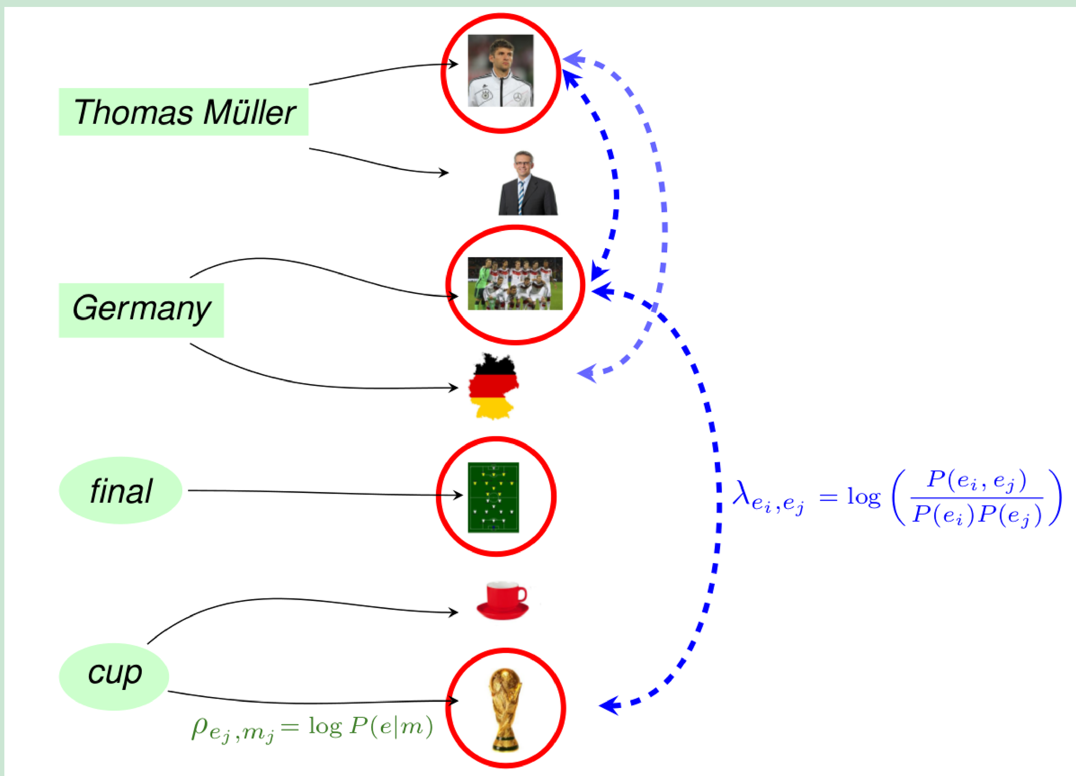
- Context = text window around mention
- Bag-of-words context: $P(c|e) = \prod_{w \in c} P(w|e)$

$$P(w|e) \approx \frac{\#\text{times } w \text{ in context of link to } e}{\#\text{words surrounding links to } e}$$

- Bayes' rule (conditional indepence: $c \perp\!\!\!\perp m \mid e$):

$$P(e|m,c) \propto P(e|m)P(c|e) = P(e|m)\prod_{w \in c} P(w|e)$$

Joint Disambiguation



Log-linear model:

$$P(\mathbf{e}|\mathbf{m}) = \frac{1}{Z(\mathbf{m})} \exp\left[\sum_{1\le i\le n} \rho_{e_i,m_i} + \sum_{1\le i<j\le n} \lambda_{e_i,e_j}\right]$$

- Leverage entity - entity co-linking statistics:

$$P(e,e') \approx \frac{\#\text{articles have links to } e \text{ and } e'}{\#\text{articles}}$$

- Markov Random Field factorization ⇒ plug-in estimators for $\boldsymbol{\rho}$, $\boldsymbol{\lambda}$ (see above figure)

Probabilistic Bag Of Hyperlinks model (PBOH)

$$\log P(\mathbf{e}|\mathbf{m},\mathbf{c}) \propto \underbrace{\sum_{i=1}^{n} \log P(e_i|m_i)}_{\text{Mention - Entity compatibility}} + \zeta \underbrace{\sum_{i=1}^{n}\sum_{w_j \in c_i} \log P(w_j|e_i)}_{\text{Context - Entity interactions}} + \underbrace{\frac{2\tau}{n-1}\sum_{i<j} \log\left(\frac{P(e_i,e_j)}{P(e_i)\,P(e_j)}\right)}_{\text{Entity - Entity coherence}}$$

Candidate selection:

- First, top 64 based on $P(e|m)$
- Then, top 10 based on $P(e|m,c)$.

Inference:

- MAP inference w/ Loopy Belief Propagation:

$$\mathbf{e}^* = \arg\max_{\mathbf{e} \in \mathcal{E}^n} P(\mathbf{e}|\mathbf{m},\mathbf{c})$$

- Fast empirical convergence (typically < 3 iterations, 400ms/doc)

Incremental Accuracy

	Datasets			
	CoNLL test A		CoNLL test B	
Baselines	R@MI	R@MA	R@MI	R@MA
Mention-Entity	69.73	69.30	67.98	72.75
Joint Uncalibrated	69.77	69.95	75.87	75.12
Joint Calibrated	75.09	74.25	74.76	78.28
Local Context	82.50	81.56	85.46	84.08
PBoH	**85.53**	**85.09**	**87.51**	**86.39**

Future Research

- Alleviate data sparseness using low-dimensional entity vector representations
- Joint mention detection and entity disambiguation w/ deep representations

Experiments (see paper)

- On Gerbil platform of [1]:
  - State-of-the-art performance on 11/14 datasets
  - 2nd best on 2/14 datasets
  - 10 state-of-the-art competitors
- Very good generalization performance across many datasets
- Works also for short texts (e.g. micro-blog posts)
- More experiments in the paper

References

[1] Ricardo Usbeck et al. Gerbil: general entity annotator benchmarking framework. WWW '15, 2015.

[2] Octavian-Eugen Ganea et al. Probabilistic bag-of-hyperlinks model for entity linking. WWW '16, 2016.

[3] Neil Houlsby and Massimiliano Ciaramita. A scalable gibbs sampler for probabilistic entity linking. In *Advances in Information Retrieval*. 2014.

[4] Zhaochen Guo and Denilson Barbosa. Robust entity linking via random walks. CIKM, 2014.

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