

Entity Disambiguation using Embeddings

Student: Bc. Lejla Metohajrová
Advisor: Prof. Ing. Igor Farkaš, Dr.
Consultants: Prof. Dr. Philippe Cudré-Mauroux
Akansha Bhardwaj, MSc
Paolo Rosso, MSc

Entity Disambiguation: The Task

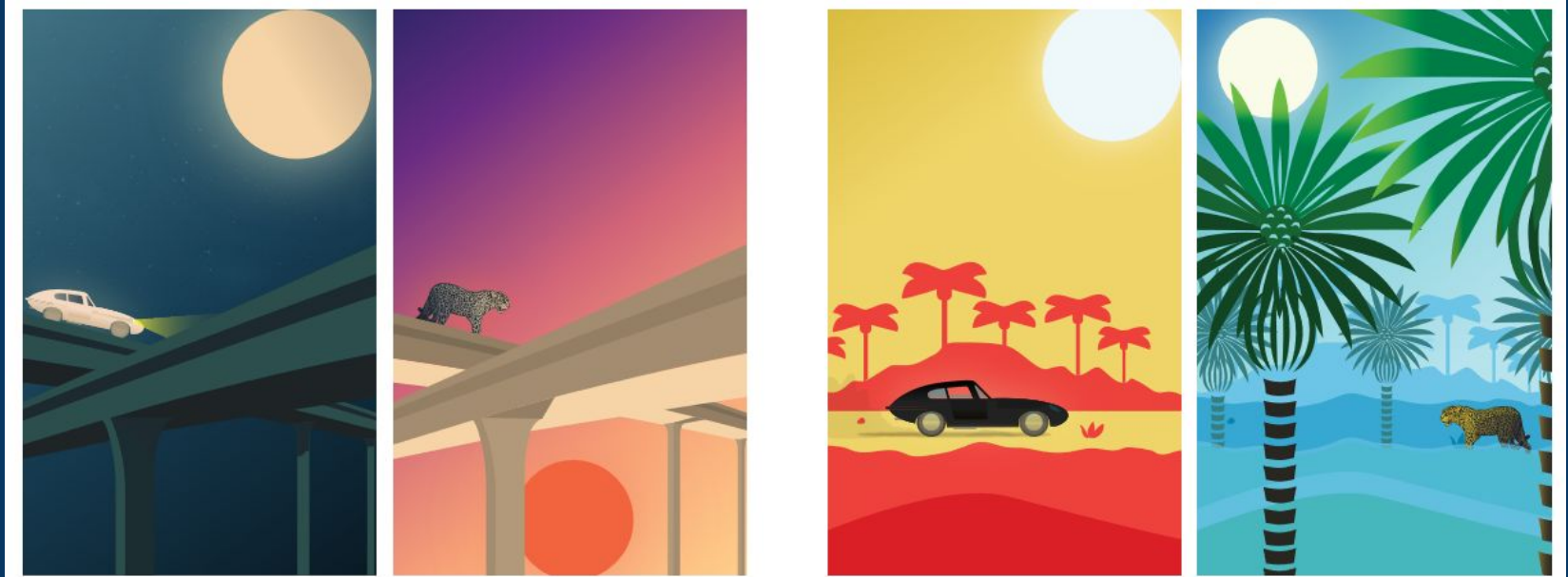
Woods is returning to the East Coast after winning the golf tournament in the UK.

Entity Disambiguation: The Task

Woods is returning to the East Coast after winning the golf tournament in the UK.

Mention	Candidates
Woods	Wood, Forest, Tiger_Woods
East Coast	East_Coast_of_the_United_States , National_Express_East_Coast, Eastern_states_of_Australia
golf	Golf , Volkswagen_Golf, Golf_(video_game)
UK	United_Kingdom , Ukrainian_language

Entity Disambiguation: The Task



The man saw a **Jaguar** speed on the highway.

The prey saw the **jaguar** cross the jungle.

Image from: <https://openai.com/blog/discovering-types-for-entity-disambiguation/>

Entity Disambiguation: The Task

$$F : (\mathcal{V}, \mathcal{C})^n \rightarrow \mathcal{E}^n$$

mentions $\mathbf{m} = (m_1, \dots, m_n) \in \mathcal{V}$

context windows $\mathbf{c} = (c_1, \dots, c_n) \in \mathcal{C}$

joint entity assignment $\mathbf{e} = (e_1, \dots, e_n) \in \mathcal{E}$

Entity Disambiguation: Probabilistic Model

Conditional probability model:


$$p(\mathbf{e}|\mathbf{m}, \mathbf{c})$$

Maximum a posteriori:

$$F(\mathbf{m}, \mathbf{c}) := \operatorname{argmax}_{\mathbf{e}' \in \mathcal{E}^n} p(\mathbf{e}'|\mathbf{m}, \mathbf{c})$$

Long vs. Short Documents

- A lot of data (long context, entity co-occurrence statistics, ...)
- Highly competitive state-of-the-art systems
- E.g. Wikipedia:
- Short, subjective context
- ~1.5 mentions per document
- Non-unified research
- Results leave out space for improvement
- E.g. Twitter:

Barack Hussein Obama II ([/bəˈrɑːk huːˈseɪn ouˈbɑːmə/](#)  listen)^[1] born August 4, 1961) is an American [attorney](#) and [politician](#) who served as the 44th [president of the United States](#) from 2009 to 2017. A member of the [Democratic Party](#), he was the first [African American](#) to be elected to the presidency. He previously served as a [U.S. senator](#) from [Illinois](#) from 2005 to 2008 and an [Illinois state senator](#) from 1997 to 2004.

No he isn't trying to ruin Obummer's non legacy, all [@realDonaldTrump](#) is doing is getting down to the truth about all the bad & illegal things [@BarackObama](#) did, when he was in office. [@POTUS](#) is too busy trying to clean up all of [#Obama](#) messes & [#KAG](#) w-out any [@dnc](#) help!

 1  11  32

Candidate Selection

- Mention-entity probability distribution: $p(e|m)$
 - based on Wikipedia and YAGO
- Top 30 potential entity candidates
- Top 4 based on $p(e|m)$
- Top 3 based on the local context-entity similarity: $\sum_{w \in c} \mathbf{x}_e^\top \mathbf{x}_w$
- Only mentions with entities in the KB

Local and Global Models

Local:

$$e_i^* = \operatorname{argmax}_{e'_i \in \mathcal{E}_{m_i}} \Psi(e'_i, c_i)$$

Global:

$$\mathbf{e}^* = \operatorname{argmax}_{\mathbf{e}' \in \mathcal{E}_{m_1} \times \dots \times \mathcal{E}_{m_n}} \sum_{i=1}^n \Psi(e'_i, c_i) + \sum_{i \neq j}^n \Phi(e'_i, e'_j)$$

Entity Embeddings

$$J(\mathbf{z}; e) := \mathbb{E}_{w^+|e} \mathbb{E}_{w^-} [h(\mathbf{z}; w^+, w^-)]$$

$$h(\mathbf{z}; w, v) := [\gamma - \langle \mathbf{z}, \mathbf{x}_w \rangle]_+$$

$$\mathbf{x}_e := \operatorname{argmin}_{\mathbf{z}: \|\mathbf{z}\|=1} J(\mathbf{z}; e)$$

(Ganea & Hofmann, 2017)

Local Model

Support score:

$$u(w) = \max_{e' \in \mathcal{E}_m} x_{e'}^\top \mathbf{A} x_w$$

Hard pruning:

$$\bar{c} = \{w \in c \mid u(w) \in \text{top}R(u)\}$$

Softmax:

$$\beta(w) = \begin{cases} \frac{\exp u(w)}{\sum_{v \in \bar{c}} \exp u(v)} & \text{if } w \in \bar{c} \\ 0 & \text{otherwise} \end{cases}$$

Context-based
local score:

$$\Psi(e', c) = \sum_{w \in \bar{c}} \beta(w) x_{e'}^\top \mathbf{B} x_w$$

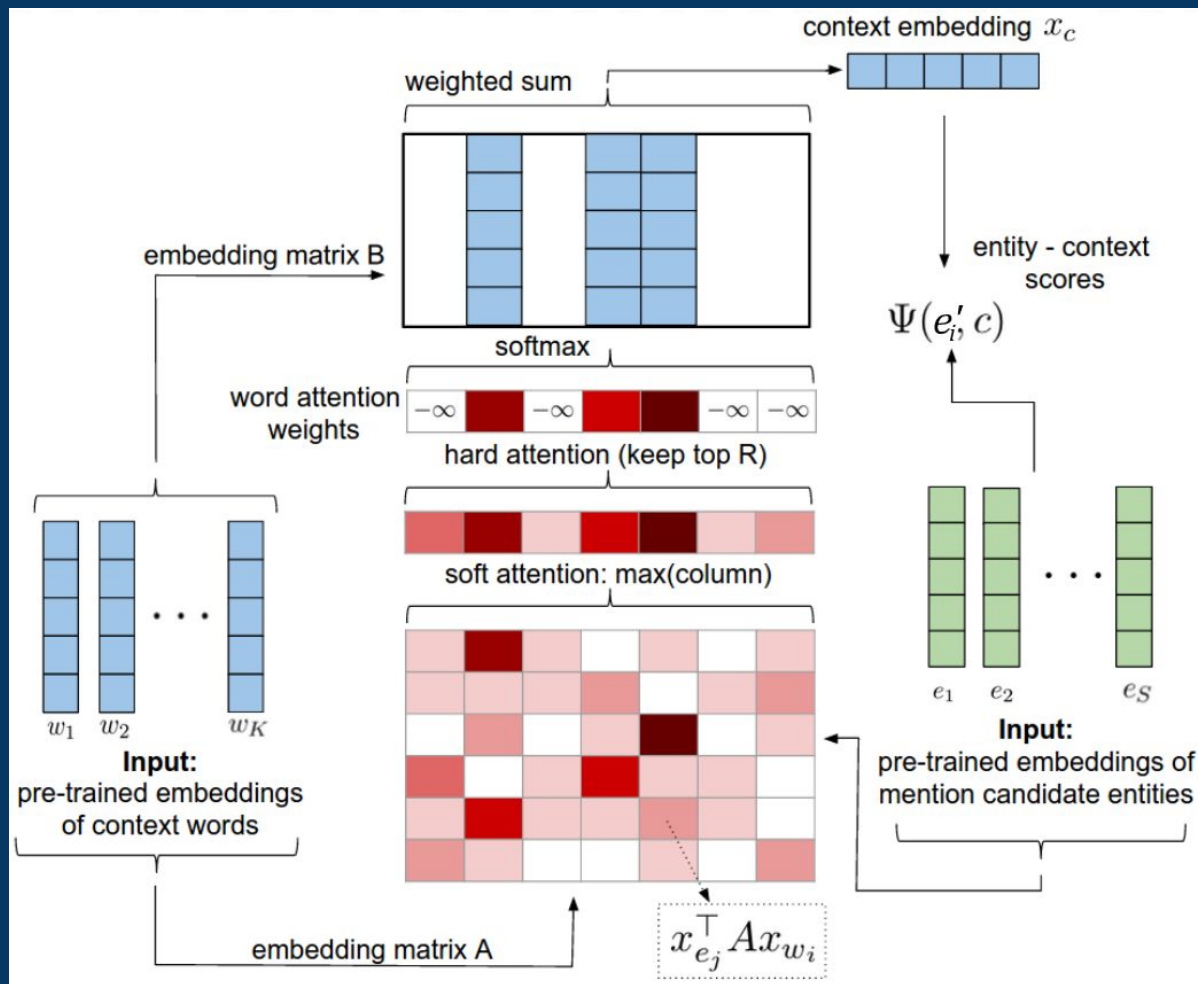
Local Model

Unary score:

$$\Psi(e', c) =$$

$$\sum_{w \in \bar{c}} \beta(w) x_{e'}^\top \mathbf{B} x_w$$

(Ganea & Hofmann, 2017)



Pairwise score

Pairwise score:

$$\Phi(e'_i, e'_j) = \sum_{q=1}^K \alpha_{ijq} \Phi_q(e'_i, e'_j)$$

Relation score:

$$\Phi_q(e'_i, e'_j) = e'_i{}^\top \mathbf{R}_q e'_j$$

Normalized score:

$$\alpha_{ijq} = \frac{1}{Z_{ijq}} \exp \left(\frac{f^\top(m_i, c_i) \mathbf{D}_q f(m_j, c_j)}{\sqrt{d}} \right)$$

Normalization factor:

$$Z_{ijq} = \sum_{\substack{j'=1 \\ j' \neq i}}^n \exp \left(\frac{f^\top(m_i, c_i) \mathbf{D}_q f(m'_j, c'_j)}{\sqrt{d}} \right)$$

(Le & Titov, 2018)

Global Model

Global model:

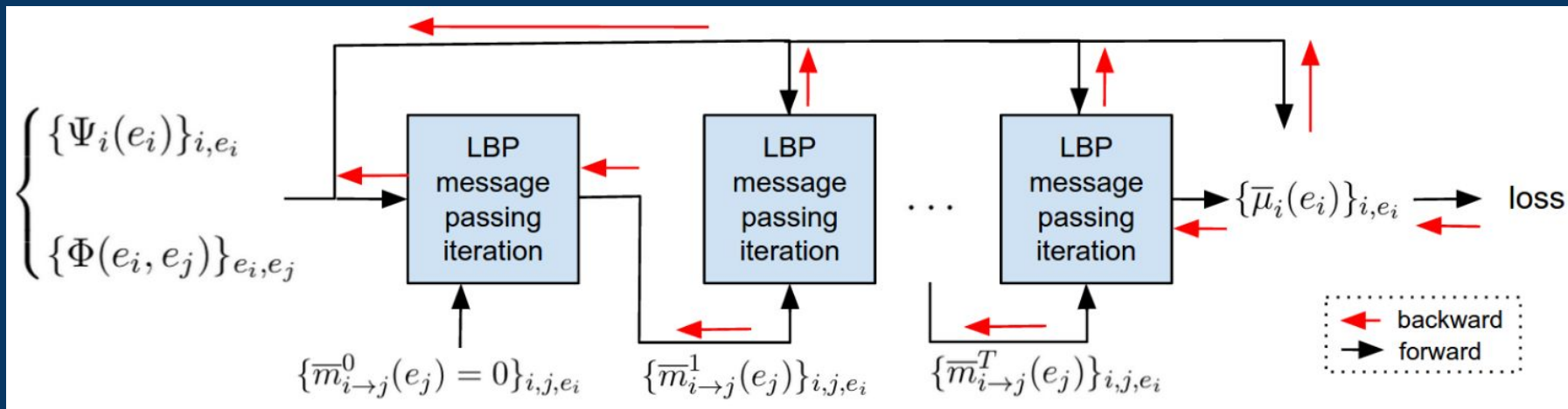
$$\text{CRF}(\mathbf{e}|\mathbf{m}, \mathbf{c}) = \exp \left(\sum_{i=1}^n \Psi(e_i) + \sum_{i<j} \Phi(e_i, e_j) \right)$$

Potentials:

$$m_{i \rightarrow j}^{t+1}(e) = \max_{e' \in \mathcal{E}_{m_i}} \{ \Psi(e') + \Phi(e, e') + \sum_{k \neq j} \bar{m}_{k \rightarrow i}^t(e') \}$$

$$\bar{m}_{i \rightarrow j}^t(e) = \log[\delta \cdot \text{softmax}(m_{i \rightarrow j}^t(e)) + (1 - \delta) \cdot \exp(\bar{m}_{i \rightarrow j}^{t-1}(e))]$$

Global Model



Marginals:

$$\mu_i(e) = \Psi(e) + \sum_{k \neq i} \bar{m}_{k \rightarrow i}^T(e)$$

(Ganea & Hofmann, 2017)

$$\bar{\mu}_i(e) = \frac{\exp[\mu_i(e)]}{\sum_{e' \in \mathcal{E}_{m_i}} \exp[\mu_i(e)]}$$

Model Training

Final score:

$$\rho_i(e) = g(\bar{\mu}_i(e), p(e|m_i))$$

Margin loss:

$$L(\theta) = \sum_{D \in \mathcal{D}} \sum_{m_i \in D} \sum_{e \in \mathcal{E}_{m_i}} h(m_i, e)$$

$$h(m_i, e) = [\gamma - \rho_i(e_i^*) + \rho_i(e)]_+$$

$$\theta = \{\mathbf{A}, \mathbf{B}, \mathbf{R}, \mathbf{D} \text{ and the weights of } f \text{ and } g\}$$

Statistics and Representation

- Entity-mention probability map $p(e|m)$ based on Wikipedia (Feb, 2014)
- Pre-trained Word2Vec (Mikolov, 2013) in local model
- Pre-trained GloVe (Pennington, 2014) in global model
- Pre-trained entity embeddings (Ganea & Hofmann, 2017) based on Wikipedia (Feb, 2014)
- All embeddings: dim = 300

Data: Long Documents

Dataset	#Mentions	#Docs	Mentions per Doc
AIDA-train	18448	946	19.5
AIDA-A (val)	4791	216	22.1
AIDA-B (test)	4485	231	19.4
MSNBC	656	20	32.8
AQUAINT	727	50	14.5
ACE2004	257	57	4.5
WNED-CWEB	11154	320	34.8
WNED-WIKI	6792	345	19.7

Data: Short Documents

Dataset	#Mentions	#Docs	Mentions per Doc
Microposts2016*-train	4905	3333	1.5
Microposts2016*-dev	148	97	1.5
Microposts2016*-test	368	286	1.3
Brian	1585	1603	1
Mena	510	162	3.1
Microposts2014	3819	2339	1.6
Tw-train	3662	2058	1.8
Tw-val	457	418	1.1
Tw-test	457	421	1.1

Experiments: AIDA Corpus

Dataset	mul-rel	sgl-rel	sgl-norm	BASELINE
AIDA-A	90.66 ± 0.2	83.87 ± 1.8	90.35 ± 0.1	73.73
AIDA-B	91.66 ± 0.2	83.79 ± 3.1	91.64 ± 0.2	71.79
MSNBC	93.38 ± 0.6	91.91 ± 1.4	93.70 ± 0.8	89.67
AQUAINT	86.92 ± 0.9	86.41 ± 0.8	88.30 ± 1.4	84.48
ACE2004	88.73 ± 0.4	87.32 ± 0.4	88.40 ± 2.1	87.32
WNED-CWEB	77.47 ± 0.4	74.60 ± 0.8	77.77 ± 0.2	69.74
WNED-WIKI	76.84 ± 0.5	73.30 ± 1.8	76.54 ± 0.9	63.96
Tw-train	69.83 ± 2.2	75.10 ± 2.0	71.91 ± 10.9	80.12
Tw-val	63.57 ± 2.4	68.36 ± 5.0	65.65 ± 13.8	80.96
Tw-test	63.84 ± 4.0	67.09 ± 4.6	64.55 ± 12.7	77.24
Micro2016*-train	72.03 ± 3.3	75.77 ± 2.0	72.72 ± 13.2	81.57
Micro2016*-dev	56.25 ± 13.1	72.30 ± 7.9	50.68 ± 25.2	90.54
Micro2016*-test	54.42 ± 5.1	70.43 ± 3.1	56.34 ± 30.1	83.15
Micro2014	57.55 ± 2.3	64.96 ± 1.7	60.10 ± 11.5	69.05
Mena	79.89 ± 2.2	81.80 ± 1.3	80.86 ± 4.0	83.32
Brian	60.69 ± 1.0	59.28 ± 0.3	60.97 ± 3.6	59.64

Experiments: Tw Corpus

Dataset	mul-rel	sgl-rel	sgl-norm	BASELINE
Tw-train	87.23 ± 0.7	84.42 ± 1.4	87.16 ± 0.9	80.12
Tw-val	85.34 ± 0.5	82.58 ± 1.2	85.82 ± 1.0	80.96
Tw-test	84.86 ± 0.9	81.14 ± 1.0	84.99 ± 0.8	77.24
Micro2016*-train	85.05 ± 0.5	83.86 ± 0.7	85.56 ± 0.9	81.57
Micro2016*-dev	90.41 ± 1.4	88.51 ± 5.0	91.62 ± 1.4	90.54
Micro2016*-test	84.62 ± 0.6	81.20 ± 4.8	85.49 ± 1.6	83.15
Micro2014	74.46 ± 0.5	72.15 ± 0.8	74.56 ± 0.8	69.05
Mena	84.54 ± 1.0	84.14 ± 0.3	84.42 ± 0.8	83.32
Brian	66.71 ± 0.2	63.28 ± 1.9	66.74 ± 0.5	59.64
AIDA-A	81.30 ± 4.7	77.66 ± 1.6	83.62 ± 6.1	73.73
AIDA-B	80.86 ± 5.7	77.49 ± 1.9	84.36 ± 7.7	71.79
MSNBC	91.81 ± 2.3	91.54 ± 0.4	92.30 ± 1.9	89.67
AQUAINT	88.48 ± 4.8	88.28 ± 0.7	89.29 ± 4.3	84.48
ACE2004	88.21 ± 0.9	88.13 ± 0.5	88.37 ± 1.1	87.32
WNED-CWEB	74.49 ± 4.7	72.93 ± 0.6	75.51 ± 4.7	69.74
WNED-WIKI	71.16 ± 6.9	68.16 ± 0.8	72.22 ± 6.6	63.96

Experiments: Microposts2016 Corpus

Dataset	mul-rel	sgl-rel	sgl-norm	BASELINE
Micro2016*-train	85.09 ± 2.8	83.69 ± 2.8	84.36 ± 3.9	81.57
Micro2016*-dev	91.22 ± 1.7	91.89 ± 1.7	92.12 ± 1.0	90.54
Micro2016*-test	86.14 ± 1.2	85.42 ± 5.5	87.50 ± 5.5	83.15
Tw-train	81.71 ± 2.5	81.80 ± 0.3	81.66 ± 2.0	80.12
Tw-val	81.11 ± 2.5	80.16 ± 3.0	81.69 ± 1.1	80.96
Tw-test	79.65 ± 5.2	79.29 ± 1.1	79.14 ± 5.1	77.24
Micro2014	70.33 ± 2.5	70.46 ± 1.0	70.59 ± 2.2	69.05
Mena	83.45 ± 2.9	83.59 ± 0.6	84.07 ± 1.1	83.32
Brian	61.74 ± 0.2	60.76 ± 3.3	61.04 ± 0.3	59.64
AIDA-A	79.53 ± 1.9	77.42 ± 6.6	79.78 ± 7.7	73.73
AIDA-B	78.63 ± 1.8	76.31 ± 7.2	79.75 ± 11.0	71.79
MSNBC	92.22 ± 0.2	91.46 ± 0.8	92.17 ± 1.2	89.67
AQUAINT	89.23 ± 1.3	87.37 ± 2.9	89.09 ± 0.9	84.48
ACE2004	88.13 ± 1.0	87.86 ± 1.5	87.99 ± 1.2	87.32
WNED-CWEB	74.82 ± 0.7	72.45 ± 3.2	74.73 ± 2.5	69.74
WNED-WIKI	71.20 ± 1.5	67.80 ± 4.7	70.58 ± 6.2	63.96

*Microposts2016 data with only ~54% of mentions.

State of the Art

Method	Microposts 2016
NEEL (Rizzo et al., 2016)	53.6 micro F1
PBOH (Ganea et al., 2016)	72.1 micro F1
#KEA (Waitelonis and Sack, 2016)	75.2 micro F1
S-MART (Yang and Chang, 2016)	81.1 micro F1

State of the Art

Method (EL)	Microposts 2014
Microsoft (Cano et al., 2014)	<u>70.1</u> micro F1
Method (ED)	Microposts 2014
DEC (Feng et al., 2018)	53.9 micro F1
$p(e m)$ baseline	69.1 micro F1
sgl-norm trained on Tw	74.6 micro F1

State of the Art

Method (EL)	Brian	Mena
TwitterNEED (Habib and Van Keulen, 2016)	55.5 micro F1	70.1 micro F1
Method (ED)	Brian	Mena
$p(e m)$ baseline	59.6 micro F1	83.3 micro F1
sgl-norm trained on Tw	66.7 micro F1	84.4 micro F1

Discussion

Embeddings	Tw	AIDA
Pretrained	84.99	91.64
All Randomized	81.62	74.49
Randomized Word	84.68	85.99
Randomized Entity	81.47	74.45
BASELINE	77.24	71.79

Thank you

Conditional Random Field

$$p(Y_v | \mathbf{X}, Y_w; v \neq w) = p(Y_v | \mathbf{X}, Y_w; v \sim w)$$

(Lafferty, J., McCallum, A., & Pereira, F. C., 2001)

CRF(X,Y):

$$\exp \left(\sum_k \mu_k \Psi_k(Y_i, \mathbf{X}, i) + \sum_j \lambda_j \Phi_j(Y_{i-1}, Y_i, \mathbf{X}, i) \right)$$