

最先端NLP勉強会

# Improving Entity Linking by Modeling Latent Relations between Mentions

Phong Le<sup>1</sup> and Ivan Titov<sup>1,2</sup> <sup>1</sup>University of Edinburgh <sup>2</sup>University of Amsterdam

ACL2018

豊田工業大学 知能数理研究室

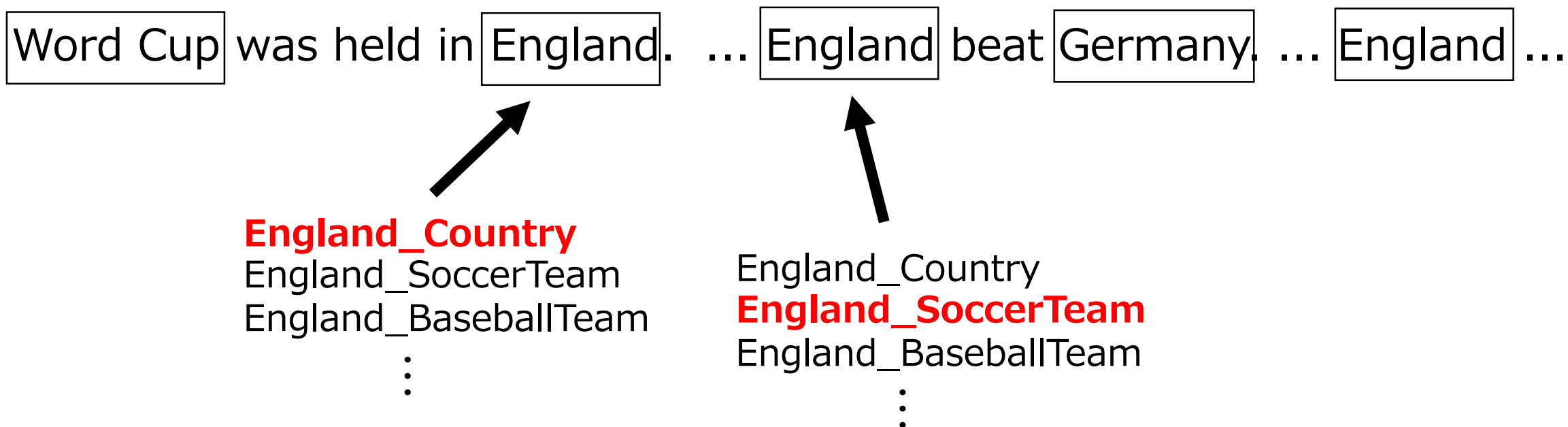
辻村有輝 (M2)

# Improving Entity Linking by Modeling Latent Relations between Mentions

- 著者： **Phong Le<sup>1</sup> and Ivan Titov<sup>1,2</sup>**    <sup>1</sup>University of Edinburgh    <sup>2</sup>University of Amsterdam
- ACL2018
- ニューラルネットでのEntity Linking
  
- Mention間の関係を考慮しながらEntity Linkingを解く
  - 関係は**教師無しで一緒に学習**
  - Coreferenceらしきものが学習された
- AIDA-CoNLLデータセットでSotA (in-domainな設定)
- Out-domainなテストセットでは既存手法に勝ったり負けたり

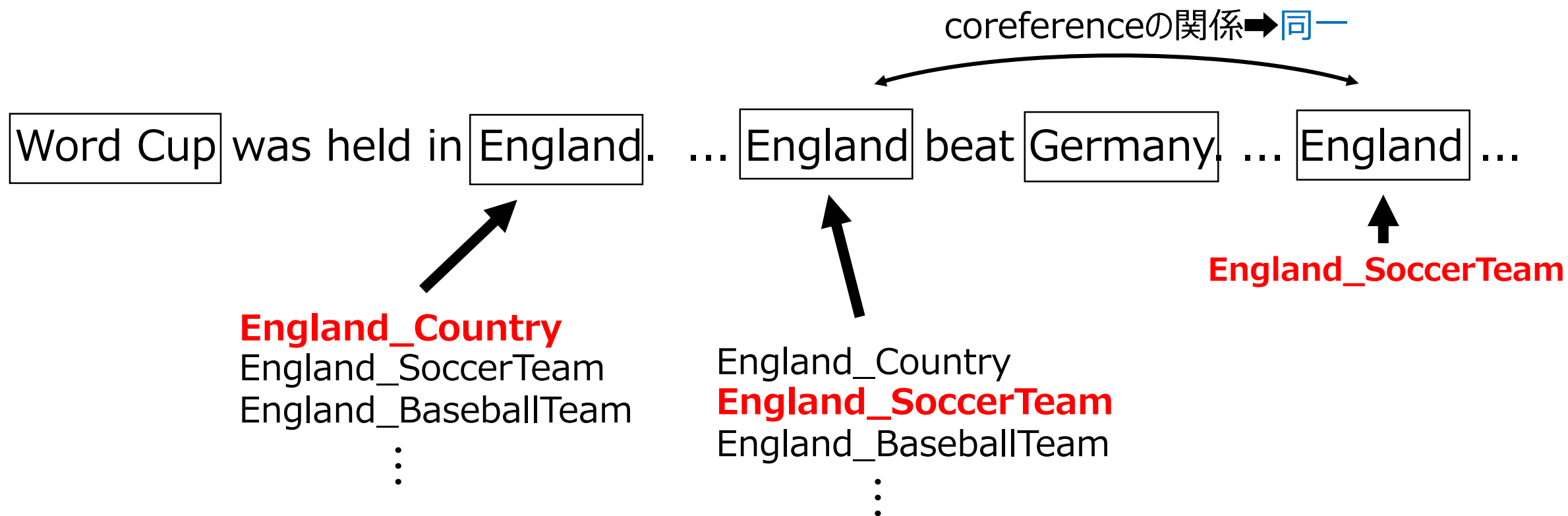
# Entity Linking

- ドキュメント内のMentionがデータベース上のどのエン트리と一致するか判定



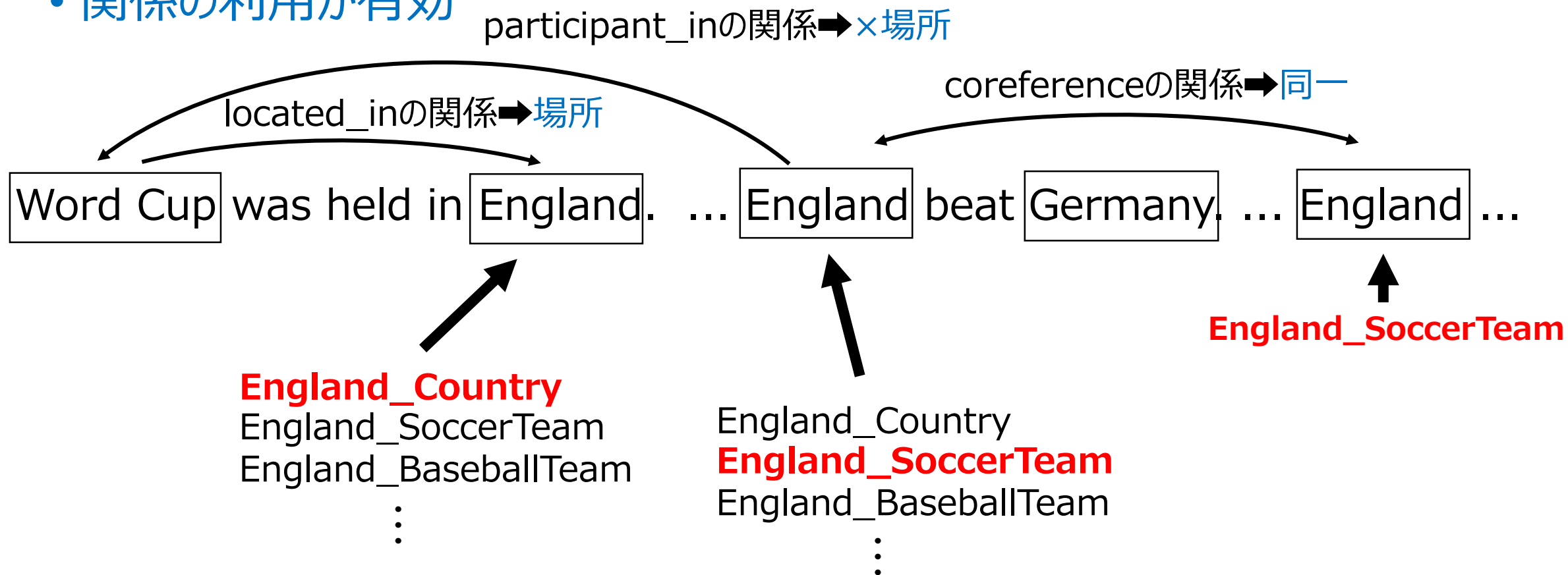
# Entity Linking

- ドキュメント内のMentionがデータベース上のどのエントリと一致するか判定
- 関係の利用が有効



# Entity Linking

- ドキュメント内のMentionがデータベース上のどのエントリと一致するか判定
- 関係の利用が有効



# Entity Linkingでの関係の利用

- 伝統的には主にCoreferenceのみが使われてきた
  - どの関係がEntity Linkingに有効か調べること自体が難しい問題だったから他の関係はあまり使われなかった（著者たちの考え）
  - しかしcoreference以外も有用なはず
- 関係は基本的には別システムで抽出した結果を持ってくるかもしくはヒューリスティックを用いて取り出す
  - ドメインごとにエキスパートの知識などがいる

# アイデア

Entity Linkingで有効な関係は  
Entity Linkingの学習を通じて自然と獲得されるはず

関係をネットワーク内の隠れ変数としてうまくモデリングし  
その関係に基づいてEntity Linkingを行うようにする

## メリット

- Coreference以外の関係も自然と利用できる
- 言語やドメインごとに必要なエキスパートの知識が減る

# アイデア

関係をネットワーク内の隠れ変数としてうまくモデリングし  
その関係に基づいてEntity Linkingを行うようにする

ある種 $k$ の関係の表現行列

$$R_k$$

$i, j$ 番目のMentionに対してリンクしようとして  
しているエントリの表現ベクトル

$$\mathbf{e}_i, \mathbf{e}_j$$



# アイデア

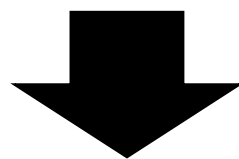
関係をネットワーク内の隠れ変数としてうまくモデリングし  
その関係に基づいてEntity Linkingを行うようにする

ある種 $k$ の関係の表現行列

$$R_k$$

$i, j$ 番目のMentionに対してリンクしようとして  
しているエントリの表現ベクトル

$$\mathbf{e}_i, \mathbf{e}_j$$



$i, j$ 番目のMention間にある種 $k$ の関係が  
成り立つときのリンクの一貫性スコア

$$\mathbf{e}_i^T R_k \mathbf{e}_j$$

# アイデア

$i, j$ 番目のMention間にある種 $k$ の関係が  
成り立つときのリンクの一貫性スコアを

$$\mathbf{e}_i^T R_k \mathbf{e}_j \text{ で計算する}$$

このスコアを最大にするようなエン트리 $e_i, e_j$ を割り当てる

例えば $i$ 番目のMentionにEngland\_Countryを割り当てた時

- $R_k$ がcoreferenceに対応
  - $j$ 番目のMentionにEngland\_Countryを割り当て  $\rightarrow$  高スコアになるべき(一貫している)
  - $j$ 番目のMentionにEngland\_SoccerTeamを割り当て  $\rightarrow$  低スコアになるべき(矛盾している)
- $R_k$ がparticipant\_inに対応
  - $j$ 番目のMentionにEngland\_Countryを割り当て  $\rightarrow$  低スコアになるべき (矛盾している)
  - $j$ 番目のMentionにEngland\_SoccerTeamを割り当て  $\rightarrow$  高スコア になるべき(一貫している)

# アイデア

$i, j$ 番目のMention間にある種 $k$ の関係が  
成り立つときのリンクの一貫性スコアを

$$\mathbf{e}_i^\top R_k \mathbf{e}_j \text{ で計算する}$$

信念に基づいて $K$ 種の関係が存在するとする

$$\begin{aligned} & \mathbf{e}_i^\top R_1 \mathbf{e}_j + \mathbf{e}_i^\top R_2 \mathbf{e}_j + \cdots + \mathbf{e}_i^\top R_K \mathbf{e}_j \\ & = \mathbf{e}_i^\top (R_1 + R_2 + \cdots + R_K) \mathbf{e}_j \end{aligned}$$

# アイデア

K種の関係が存在するときの一貫性スコアを以下で表す

$$\mathbf{e}_i^\top (R_1 + R_2 + \dots + R_K) \mathbf{e}_j$$

実際は*i, j*番目のMentionに*k*番目の関係が成り立つとは限らない

➡ 成り立つかどうかの**係数**をつける

$$\mathbf{e}_i^\top (\alpha_{ij1} R_1 + \alpha_{ij2} R_2 + \dots + \alpha_{ijK} R_K) \mathbf{e}_j = \Phi(e_i, e_j)$$

このスコアを最大化するエントリがネットワークの予測値になる※

$$E^* = \operatorname{argmax}_{e_1, e_2, \dots, e_n} \sum_{i \neq j} \Phi(e_i, e_j)$$

学習にはマージンベースな  
損失関数を使う

係数 $\alpha_{ijk}$ の値は入力ドキュメントからニューラルネットで計算 (アテンションライク)  
スコアの最大化を通じて $\alpha_{ijk}$ の値も**関係の教師なし**でうまく計算するようになる

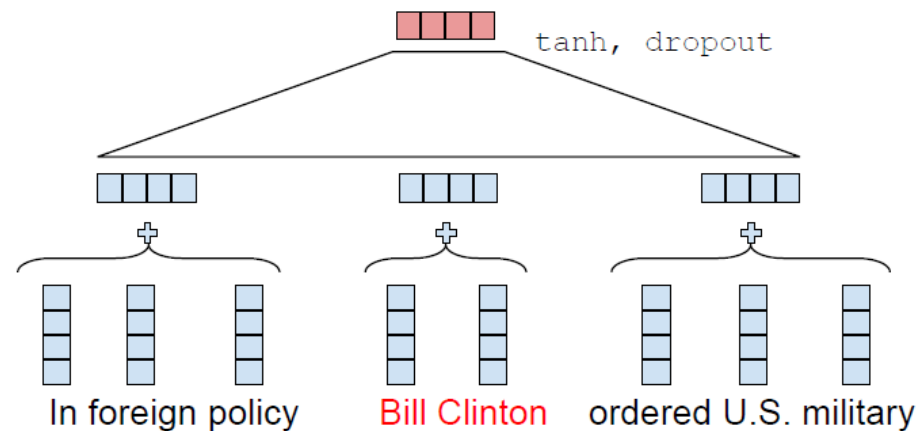
※実際は関係を気にしないスコアも加算され最終的な判断にはさらにもう一つ別のネットワークを通す

# ネットワーク構造

係数はアテンションのように計算

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

$c_i$  :  $i$ 番目のMentionのコンテキスト



$$\Phi(e_i, e_j) = \mathbf{e}_i^\top (\alpha_{ij1} R_1 + \alpha_{ij2} R_2 + \dots + \alpha_{ijK} R_K) \mathbf{e}_j$$

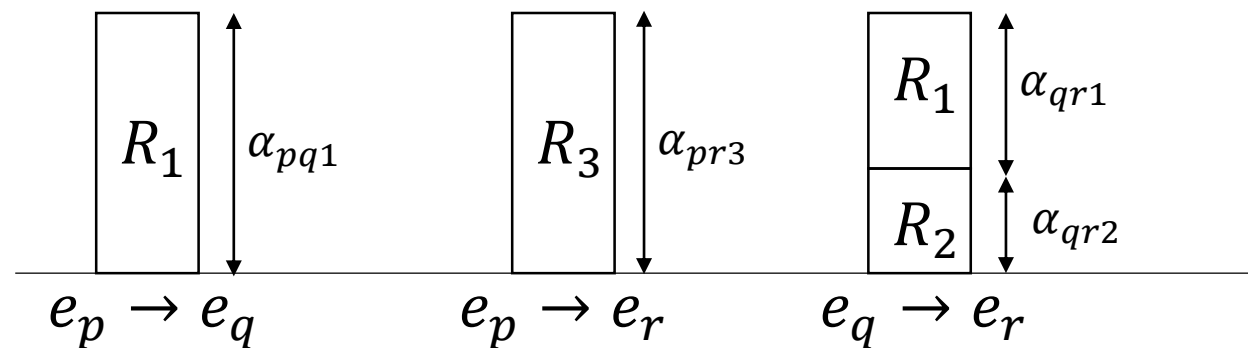
$$E^* = \operatorname{argmax}_{e_1, e_2, \dots, e_n} \sum_{i \neq j} \Phi(e_i, e_j)$$

# $\alpha_{ijk}$ の正規化

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

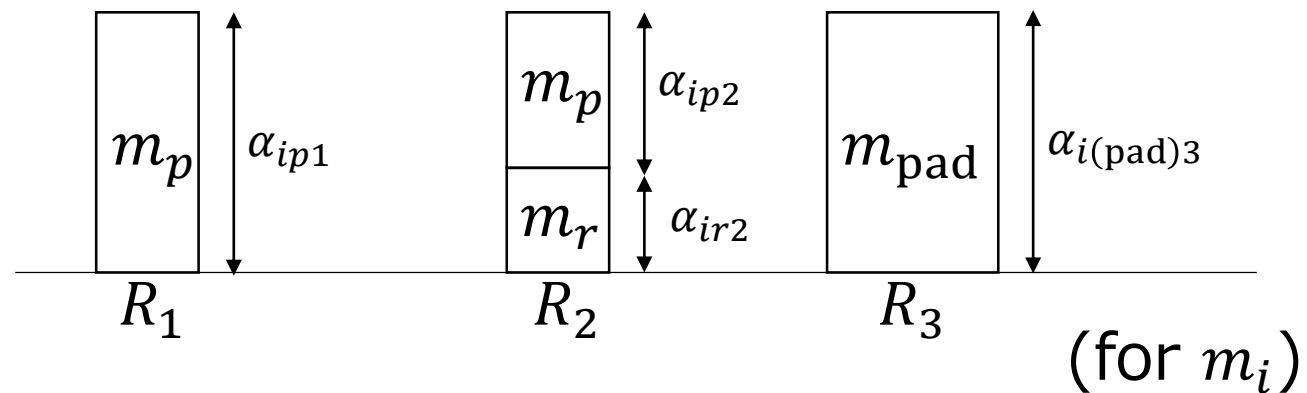
Rel-norm : 代表関係を選ぶ  $\sum_{i \neq j} \mathbf{e}_i^\top (\alpha_{ij1} R_1 + \alpha_{ij2} R_2 + \dots + \alpha_{ijK} R_K) \mathbf{e}_j$

$$Z_{ijk} = \sum_k \exp\left(\frac{f^\top(m_i, c_i) D_k f(m_j, c_j)}{\sqrt{d}}\right)$$



Ment-norm : 代表エントリを選ぶ  $\sum_k \sum_i \mathbf{e}_i^\top R_k (\alpha_{i1k} \mathbf{e}_1 + \alpha_{i2k} \mathbf{e}_2 + \dots + \alpha_{ink} \mathbf{e}_n)$

$$Z_{ijk} = \sum_{j (\neq i)} \exp\left(\frac{f^\top(m_i, c_i) D_k f(m_j, c_j)}{\sqrt{d}}\right)$$



# 実際のところ

スコアは関係を気にしない (他のMentionを気にしない) 項もある

$$\Psi(e_i, c_i) = \mathbf{e}_i^T \mathbf{B} f(c_i)$$

$c_i$  :  $i$ 番目のMentionのコンテキスト  $f(c_i)$  : アテンションベースのネットワーク

$$E^* = \operatorname{argmax}_{e_1, e_2, \dots, e_n} \sum_i \Psi(e_i, c_i) + \frac{1}{n-1} \sum_{i \neq j} \Phi(e_i, e_j)$$

ベースにした先行研究のモデルも同じ形でスコアを計算

- 結局のところ違いは $\Phi$ の式

$$\Phi(e_i, e_j) = \mathbf{e}_i^T R \mathbf{e}_j$$

( $K=1$ のrel-normはこれと等価)

提案手法 :  $\sum_k \mathbf{e}_i^T a_{ijk} R_k \mathbf{e}_j$

# 正則化

各関係が別々の関係を表現するように正則化項を追加

$$\lambda_1 \sum_{i,j} \text{dist}(\mathbf{R}_i, \mathbf{R}_j) + \lambda_2 \sum_{i,j} \text{dist}(\mathbf{D}_i, \mathbf{D}_j)$$

$$\text{dist}(\mathbf{x}, \mathbf{y}) = \left\| \frac{\mathbf{x}}{\|\mathbf{x}\|_2} - \frac{\mathbf{y}}{\|\mathbf{y}\|_2} \right\|_2$$

---

$$\Phi(e_i, e_j, D) = \sum_{k=1}^K \alpha_{ijk} \Phi_k(e_i, e_j, D)$$

$$\Phi_k(e_i, e_j, D) = \mathbf{e}_i^T \mathbf{R}_k \mathbf{e}_j \quad \alpha_{ijk} = \frac{1}{Z_{ijk}} \exp \left\{ \frac{f^T(m_i, c_i) \mathbf{D}_k f(m_j, c_j)}{\sqrt{d}} \right\}$$



# 結果 (in-domain)

Methods	Aida-B
Chisholm and Hachey (2015)	88.7
Guo and Barbosa (2016)	89.0
Globerson et al. (2016)	91.0
Yamada et al. (2016)	91.5
K=1のrel-normと等価 → Ganea and Hofmann (2017)	92.22 ± 0.14
rel-norm K=6	92.41 ± 0.19
ment-norm K=3	<b>93.07</b> ± 0.27
ment-norm ( $K = 1$ )	92.89 ± 0.21
ment-norm (no pad) K=3	92.37 ± 0.26

Table 1: F1 scores on AIDA-B (test set).

- rel-normもment-normも $K > 1$ で性能向上
- 学習にかかったエポック数は $K = 1$ で1250epoch,  $K = 6$ で120epoch (rel-norm)

# 結果 (out-domain)

AIDA-CoNLLで学習して、外部ドメインのテストセットで評価

Methods	MSNBC	AQUAINT	ACE2004	CWEB	WIKI	Avg
Milne and Witten (2008)	78	85	81	64.1	81.7	77.96
Hoffart et al. (2011)	79	56	80	58.6	63	67.32
Ratinov et al. (2011)	75	83	82	56.2	67.2	72.68
Cheng and Roth (2013)	90	<b>90</b>	86	67.5	73.4	81.38
Guo and Barbosa (2016)	92	87	88	77	<b>84.5</b>	<b>85.7</b>
Ganea and Hofmann (2017)	93.7 ± 0.1	88.5 ± 0.4	88.5 ± 0.3	<b>77.9 ± 0.1</b>	77.5 ± 0.1	85.22
rel-norm	92.2 ± 0.3	86.7 ± 0.7	87.9 ± 0.3	75.2 ± 0.5	76.4 ± 0.3	83.67
ment-norm	<b>93.9 ± 0.2</b>	88.3 ± 0.6	<b>89.9 ± 0.8</b>	77.5 ± 0.1	<u>78.0 ± 0.1</u>	<u>85.51</u>
ment-norm ( $K = 1$ )	93.2 ± 0.3	88.4 ± 0.4	88.9 ± 1.0	77.0 ± 0.2	77.2 ± 0.1	84.94
ment-norm (no pad)	93.6 ± 0.3	87.8 ± 0.5	<b>90.0 ± 0.3</b>	77.0 ± 0.2	77.3 ± 0.3	85.13

Table 2: F1 scores on five out-domain test sets. Underlined scores show cases where the corresponding model outperforms the baseline.

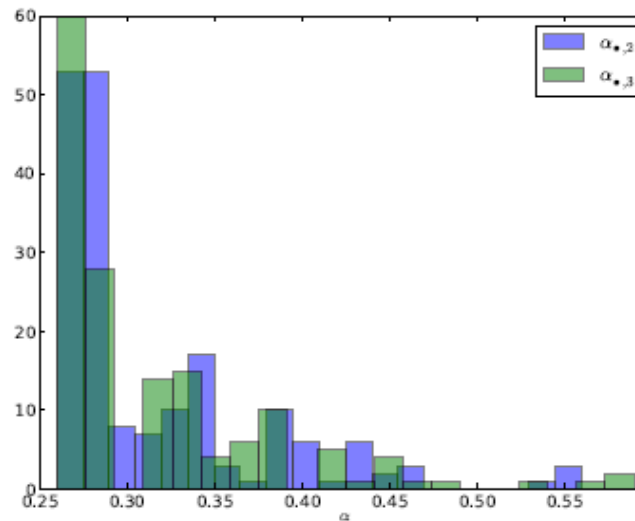
- 外部ドメインのテストセットではrel-normは $K > 1$ で性能低下
- Ment-normも既存手法に勝てるドメインと負けるドメインがある

# エラー解析

rel-norm	on Friday , Liege police said in	ment-norm
██████████   ████	(1) missing teenagers in <b>Belgium</b> .	████ ████
██████   ██████████	(2) UNK <b>BRUSSELS</b> UNK	████ ████
██████████   ████	(3) UNK <b>Belgian</b> police said on	████ ████
██████████   ████	(4) . " a <b>Liege</b> police official told	██████   ████ ████
██████████   ████	(5) police official told <b>Reuters</b> .	████ ████
██████████   ████	(6) eastern town of <b>Liege</b> on Thursday ,	██████   ████ ████
██████████   ████	(7) home village of <b>UNK</b> .	████ ████
██████   ██████████	(8) link with the <b>Marc Dutroux</b> case , the	██████   ████ ████
██████████   ████	(9) which has rocked <b>Belgium</b> in the past	████ ████

$a_{ij1}$   $a_{ij2}$   $a_{ij3}$

$e_i$  (pointing to the text)
   
 $e_j$  (pointing to the ment-norm column)



同じMentionに対して大きな重み  $\rightarrow$  coreference (?)

2番目と3番目の関係に対する係数は表で見る限りはほぼ同じに見えるがヒストグラムで見ると同一ではないので違う関係が取れていると主張されている

# エラー解析

$$e_i^T \alpha_{ijk} R_k e_j$$

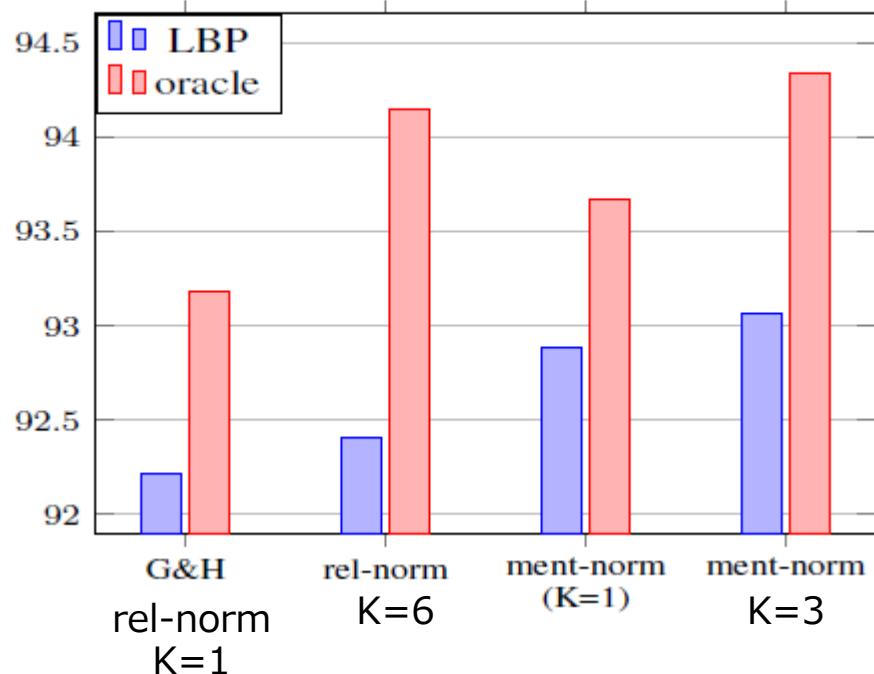
間違った時  
考えていい？

あるMentionに割り当てられたエンタリが間違いであるせいで

関係考慮時に反対側のMentionのスコア計算でエラーが伝播してしまう

➡ 一つのMentionの予測時に他すべてに正解のエンタリを与えてスコア計算

## Oracle



- Rel-normのほうがセンシティブ
- K=1のment-normより Multi-Relationalなrel-normのほうが Oracleでは高いスコア

# まとめ

- ネットワーク内で関係成立の重み付けとそれを考慮したスコアの計算を行い in-domainなデータセットでSotAなF値
- 関係は教師なし

$$\mathbf{e}_i^T (\alpha_{ij_1} R_1 + \alpha_{ij_2} R_2 + \dots + \alpha_{ij_K} R_K) \mathbf{e}_j$$

# 感想

- 汎用性は高いように見える
- Kが小さい (rel-normで6, ment-normで3)
- coreferenceだと主張されたが、本当にそうかは正直これだけだとわからない
  - 同じMentionだがcoreferenceではないものに低い重みが付いた結果が見たかった

おわり