ウェブ検索クエリに対する 周辺語を考慮した教師なしエンティティリンキング

豊田 樹生　夜久 真也　石川 葉子　土沢 誉太
Kulkarni Kaustubh　Bhattacharjee Anupam　宰川 潤二
ヤフー株式会社
{itoyota, syaku, yokishik, ytsuchiz, kkulkarn, abhattac, jsaikawa}@yahoo-corp.jp

1 はじめに

近年，商用検索エンジンにおいて，エンティティの概要を簡潔にユーザに提示するために知識パネル[1]が用いられるようになってきた．ウェブ検索クエリに対するエンティティリンキングはこの知識パネルを提示するための重要な構成要素の一つである．

ここで，ウェブ検索クエリのうちのどれだけが知識パネルの提示の対象になりうるかについては，いくつか報告がある [3, 9]．しかし，これらの報告は英語圏のユーザをターゲットとした商用ウェブ検索エンジンにおいての話であり，日本の商用検索エンジンの検索ログにおいてどのような分布になっているかは定かではない．

また，日本語ウェブ検索クエリに対するエンティティリンキングの研究はほとんど行われていない．齋藤ら [10] は日本語のウェブ検索クエリに対する教師なしのエンティティリンキング手法を提案している．しかし，この手法ではエンティティタームの周辺語を考慮したエンティティリンキングを行えていない．

そこで本研究では次のような貢献を行う．

(i) エンティティリンキングの観点から，日本の商用検索エンジンの実際のクエリログの分類を行う．
(ii) 日本語クエリに対する，周辺語を考慮した教師なしのエンティティリンキング手法を提案し，比較した手法間で最高性能を達成したことを示す．

2 関連研究

Roi ら [2] は FEL (Fast Entity Linker) 及びロジスティック回帰を用いたエンティティ-周辺語モデルを組み合わせた確率的モデルを提案している．しかし，本研究とは異なり，日本語に対する適用を考えた場合，周辺語が分散表現のエントリに存在しスコアリングできるか否かは，分かち書きの精度に依存するという問題がある．

Faegheh ら [4] は混合言語モデル (MLM) と commonness (CMNS) を組みわせた教師なしのエンティティリンキング手法を提案している．しかし，MLM は情報元となる各フィールドの重みを静的に設定している．本研究と異なり，クエリごとに動的に重要度を捉えることができていない．

3 問題定義

本研究ではあるウェブ検索クエリ q に対し， 知識ベースを検索し，最も適切なエンティティ，主要語，周辺語の組 $(e, s_s, s_c)^*$ を見つけるタスクを行う．

$$(e, s_s, s_c)^* = \underset{e \in E_q, (s_s, s_c) \in S_q}{\operatorname{argmax}} P(e, s_c|q) \quad (1)$$

$$= \underset{e \in E_q, (s_s, s_c) \in S_q}{\operatorname{argmax}} P(e|q)P(s_c|e) \quad (2)$$

ここで，式 2 では $q \perp s_c \mid e$ 及び $P(s_c|e, q) = P(s_c|e)$ を仮定している．

図 1 にクエリ “伏見 名古屋” での記号対応例を示す．

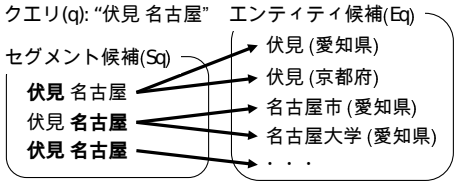

図 1: “伏見 名古屋” の例 ($S_q = \{$(伏見, 名古屋), (名古屋, 伏見), (伏見 名古屋, empty)$\}$， 太字は主要語)

[1] https://blogs.bing.com/search/2013/03/21/understand-your-world-with-bing/

ここで，本論文での主要な記号を表 1 に示す．

表 1: 主要な記号

記号	説明
E	システム中のすべてのエンティティ集合
S	システム中のすべてのセグメント集合
C	システム中のすべての情報元の集合
e	エンティティ (クリックログの URL は e に解決済み)
q	クエリ
s	セグメント (i.e., トークン列)
c	コンテンツの情報元 (c_w = Wikipedia, c_q = クリックログ)
d_e	エンティティe の仮想文書
s_c	周辺語を示すセグメント
s_s	主要語を示すセグメント
a_s	s はエイリアスか否か ($a_s \in \{0,1\}$) (i.e., リンクの付与されたセグメント)
S_e	エンティティe の持つセグメントの集合 (i.e., 名前や別名の集合)
S_q	クエリ q の持つ s_s,s_c ペアの集合
E_q	クエリ q の持つエンティティの集合
$a_{s,e}$	s は e を指すエイリアスか否か ($a_{s,e} \in \{0,1\}$)
$n(s,c)$	情報元 c における s の生起回数
$n(e,c)$	情報元 c における e の生起回数
$n(s,s_s,s_c,c_q)$	クリックログ c_q において 補間先セグメントが s のときの s_s,s_c の共起回数
$n(q,e,c_q)$	クリックログ c_q における q,e の共起回数
$n(s_c,d_e)$	仮想文書 d_e における s_c の生起回数

4 提案手法

本研究では，FEL[2] に対しクリックログから生成されるクエリ補間モデルを組み合わせ，クエリ-エンティティモデルを生成する．また，周辺語から構成される仮想文書に対し，Latent Dirichlet Allocation (LDA) を適用し，エンティティ-周辺語モデルを生成する．

4.1 クエリ-エンティティモデル

クエリ中 q が与えられたときエンティティe が生成される確率は式 3 で表現される:

$$P(e|q) \approx \max_{s \in \{s_s\} \cup S_{(s_s,s_c)},(s_s,s_c) \in S_q} P(e|s)P(s|s_s,s_c) \tag{3}$$

ここで，$P(e|s)$ は FEL [2] に基づいている．セグメント s が与えられたときエンティティe が生成される確率を表現している．$P(s|s_s,s_c)$ はクエリ補間モデルにおいて，s_s，s_c のペアが与えられたときに補間先のセグメント s が生成される確率を表現している．

4.1.1 セグメント-エンティティモデル

セグメント s が与えられたときエンティティe が生成される確率は式 4 で表現される:

$$P(e|s) = \sum_{\substack{c \in \\ \{c_q,c_w\}}} P(c|s)P(e|c,s) \tag{4}$$

ここで，c_q は情報元がクリックログであること，c_w は情報元が Wikipedia であることを示す．

セグメント s が与えられたときに情報元 c が生成される確率は式 5 で表現される:

$$P(c|s) = \frac{n(s,c)+1}{|C| + \sum_{c'} n(s,c')} \tag{5}$$

情報元 c 及びセグメント s が与えられたときエンティティe が生成される確率は式 6 で表現される:

$$P(e|c,s) = \sum_{a_s=\{0,1\}} P(a_s|c,s)P(e|a_s,c,s) \tag{6}$$

ここで，情報元 c 及びセグメント s が与えられたときエイリアス a_s が生成される確率はそれぞれ式7，式 8 で表現される:

$$P(a_s=0|c,s) = 1 - P(a_s=1|c,s) \tag{7}$$

$$P(a_s=1|c,s) = \frac{\sum_{s:a_s=1} n(s,c)}{n(s,c)} \tag{8}$$

また，エイリアス a_s，情報元 c，セグメント s が与えられたときエンティティe が生成される確率は式 9 で表現される:

$$P(e|a_s,c,s) = \frac{\sum_{s:a_{s,e}=1} n(s,c) + \mu_c \cdot P(e|c)}{\mu_c + \sum_{s:a_s=1} n(s,c)} \tag{9}$$

ここで，情報元 c が与えられたときエンティティe が生成される確率は式 10 で表現される:

$$P(e|c) = \frac{n(e,c)+1}{|E| + \sum_{e \in E} n(e,c)} \tag{10}$$

4.1.2 クエリ補間モデル

クエリ補間モデルはクエリ中の主要語 s_s 及び周辺語 s_c が与えられたときクエリ補間先のセグメント s の生成される確率を表現する:

$$P(s|s_s,s_c) = \frac{n(s,s_s,s_c,c_q) + \alpha^{I(s=s_s)}}{\sum_{s \in \{s_s\} \cup S_{(s_s,s_c)}} (n(s,s_s,s_c,c_q) + \alpha^{I(s=s_s)})} \tag{11}$$

ここで，α は正の整数である．また，$S_{(s_s,s_c)}$ は下記の手順により生成する．

1. クリックログの各レコード $(q, e, n(q, e, c_q))$ のうち q から展開可能なすべての s_s， s_c を列挙する．
2. $s(\in S_e)$ に対し，(s_s, s_c, e) の組が条件を満たすとき s を集合 $S_{(s_s,s_c)}$ に加える． ここで，条件とは $s(\in S_e)$ 及び s_s に小文字化，記号削除を行なったとき，s_s が s の部分文字列もしくは同一の文字列となることである．

4.2 仮想文書の生成

下記の手順により各エンティティに対して周辺語を集めた仮想文書 d_e を生成する．

1. クリックログの各レコード $(q, e, n(q, e, c_q))$ のうち q から展開可能なすべての s_s，s_c を列挙する．
2. S_e の要素のうち s_s と完全一致する要素がある場合， その e と対応する仮想文書 d_e に対して s_c を $n(q, e, c_q)$ 個加える．

4.3 エンティティ-周辺語モデル

エンティティ-周辺語モデルは式 12 で表現される:

$$P(s_c|e) \approx \begin{cases} \beta \cdot P_{\mathrm{lda}}(s_c|e) & \text{if } s_c \neq \text{empty} \\ 1.0 & \text{otherwise} \end{cases} \quad (12)$$

ここで $P(s_c|e)$ はエンティティ e が与えられたとき周辺語 s_c が生成される確率である．

また， β は周辺語のつきやすさを示す係数である:

$$\beta = \frac{\sum_{s_{c'}} n(s_{c'}, d_e) - n(s_c = \text{empty}, d_e)}{\sum_{s_{c'}} n(s_{c'}, d_e)} \quad (13)$$

4.3.1 Latent Dirichlet Allocation

Xing ら [8] の式 8 に基づき，本研究では Online LDA[5] により $P_{\mathrm{lda}}(s_c|e)$ の分布を求める．ここで Xing ら [8] における w は s_c，d は d_e に本研究ではそれぞれ対応する．ただし，各 d_e から empty は除外する:

$$P_{\mathrm{lda}}(s_c|e) = P(s_c|d_e, \hat{\theta}, \hat{\phi}) \quad (14)$$

$$= \sum_{z=1}^{K} P(s_c|z, \hat{\phi}) P(z|d_e, \hat{\theta}) \quad (15)$$

ここで K はトピック数であり，$\hat{\theta}$ 及び $\hat{\phi}$ はそれぞれ θ 及び ϕ の予測分布である．

5 評価実験

5.1 訓練用事例及びパラメータ

- **知識ベース** 2018 年 11 月 19 日付ダンプ[2]
- **クリックログ** 2018 年 01 月 01 日から 11 月 30 日の期間に Yahoo!検索に発行された飛び先がモバイル版 Wikipedia(ja) のクリックログ
- **Wikipedia** 2018 年 11 月 01 日付 Wikipedia(ja) の pages-articles.xml
- **分散表現** 2018 年 04 月 20 日付 Wikipedia(ja) より学習された word2vec モデル [3]
- **パラメータ** $\alpha = 50$ または $\alpha = 500$[4]， $\mu_c = 0.1$, $K = 500$

5.2 評価用事例

クエリログの分類 2018 年 12 月 01 日から 12 月 14 日の期間に Yahoo!検索に発行されたクエリのうち，9,542 クエリ (計 10,000 imps) を重み付きでランダムサンプリングした．そして，エンティティクエリ [7] のみを抽出した[5]．結果，全体の約 23%にあたる，2,020 クエリ (計 2,257 imps) がエンティティクエリであった．表 2 に正負例を示す．

表 2: エンティティクエリか否かの正負例

正	中居正広	脂漏性皮膚炎
	関ジャニ∞ 安田	ulu 歌手
	ナヴィ パズドラ	西門クリニック 相模原
負	宇野昌磨 ツイッター	冬 コーデ
	天気 東京	出発点 英語
	占い 無料	ツムツム イベント

アノテーション 前述のエンティティクエリのうちの 1,915 クエリ[6]に対して，Y-ERD[4] のガイドラインに従い，候補となりうるエンティティを割りあて，クエリ-エンティティペアを生成した. そして，スコアの付与を行い，評価用事例とした．ここで，評価スケールは次の 3 つである: 関連が強い (スコア 1)， 関連が弱い (スコア 0.5)， 関連がない (スコア 0)．

[2]非公開の統合的知識ベース．主たる情報源は Wikidata, Wikipedia，Freebase

[3]https://github.com/Kyubyong/wordvectors

[4]$S_{(s_s,s_c)}$ の生成が記号削除及び小文字化処理のみで完結し，部分一致による照合を含まない場合は前者のパラメータを用いた

[5]このとき，全ての参加システムにおいて検索結果数がゼロのクエリは対象外とした

[6]クエリとひも付きうるエンティティの同一性が確認できない場合は除外した．具体的な事例としては，住所不明でエンティティの位置を確認できない，名称以外の値が不明などが挙げられる．

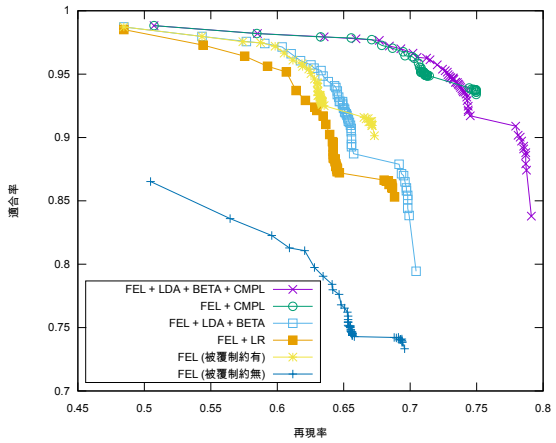

図 2: 再現率-重み付き適合率@1 グラフ

5.3 ベースラインと提案手法との精度比較

ベースラインは 3 モデル用意した: FEL (被覆制約[7]の有無の違いで 2 つ)， FEL + LR (FEL とロジスティック回帰モデルの組み合わせ[8]) [2].

提案手法は 3 モデル用意した: FEL + LDA + BETA (FEL と LDA によるエンティティ-周辺語モデルの組み合わせ)， FEL + CMPL (FEL とクエリ補間モデルの組み合わせ)， FEL + LDA + BETA + CMPL (すべてのオプションの組み合わせ).

システムが出力したクエリ-エンティティペアに対して重み付き適合率 [6]@1 による評価を行なった．各再現率での，重み付き適合率@1 を図 2 に示す.

また，表 3 に示すように，各モデル間での F1 値の比較を行った．提案手法の FEL + LDA + BETA + CMPL が比較した手法間で最大の F1 値 0.839 を達成した.

表 3: F1 値の比較． いずれの提案手法もベースラインを上回った (Bootstrap[1]， p<0.05)

モデル	被覆制約	Best F1
FEL + LDA + BETA + CMPL	有	**0.839 (+0.123)**
FEL + CMPL	有	0.832 (+0.116)
FEL + LDA + BETA	有	0.774 (+0.058)
FEL + LR [2]	有	0.764 (+0.048)
FEL [2]	有	0.773 (+0.057)
FEL [2]	無	0.716

[7]クエリ中のすべてのトークンが過不足なくモデルのエントリと照合可能である場合のみを許容する制約

[8]前述の仮想文書を使用．現実的な実行時間での実行が容易ではなかったため，一文書あたり最大 50 セグメントが正例となるように復元抽出で再標本化した．最適化には L-BFGS-B を使用 ($\lambda = 10$, $\rho = 5$).

具体的な回答事例の比較

次に，一部の事例について回答事例の比較を行った．クエリ “はやぶさ 新幹線” に対して，ベースラインの手法は誤った回答をしていた．原因は，形態素解析器の分かち書きのずれにより “新幹線” に対応する分散表現を保持しておらず， “新幹線” が周辺語に来る回答事例が候補から外れてしまったことである．提案手法は形態素解析器に対する依存がないため，このような事例でも正しく回答することができた.

また，クエリ “move 車” に対してもベースラインの手法は誤った回答をしていた．原因は，ベースラインの手法ではクエリ補間ができなかったため，“ダイハツ・ムーヴ” が回答候補から外れてしまったことである.

表 4: 回答事例の比較

クエリ	FEL+LR [2]	FEL + LDA + BETA + CMPL
はやぶさ 新幹線	新幹線	はやぶさ (新幹線)
move 車	move (音楽ユニット)	ダイハツ・ムーヴ

6 おわりに

本研究では実際の日本語ウェブ検索クエリをエンティティリンキングの観点から分類し，エンティティクエリの割合は全体の約 23%であると報告した．また，日本語ウェブ検索クエリに対する周辺語を考慮した教師無しエンティティリンキング手法を提案し，比較した手法間で最高性能を達成したことを示した.

参考文献

[1] Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. An empirical investigation of statistical significance in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 995–1005. Association for Computational Linguistics, 2012.

[2] Roi Blanco, Giuseppe Ottaviano, and Edgar Meij. Fast and space-efficient entity linking for queries. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pp. 179–188. ACM, 2015.

[3] Jiafeng Guo, Gu Xu, Xueqi Cheng, and Hang Li. Named entity recognition in query. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pp. 267–274. ACM, 2009.

[4] Faegheh Hasibi, Krisztian Balog, and Svein Erik Bratsberg. Entity linking in queries: Tasks and evaluation. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*, pp. 171–180. ACM, 2015.

[5] Matthew Hoffman, Francis R Bach, and David M Blei. Online learning for latent dirichlet allocation. In *advances in neural information processing systems*, pp. 856–864, 2010.

[6] Patrick Pantel and Marco Pennacchiotti. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pp. 113–120. Association for Computational Linguistics, 2006.

[7] Jeffrey Pound, Peter Mika, and Hugo Zaragoza. Ad-hoc object retrieval in the web of data. In *Proceedings of the 19th international conference on World wide web*, pp. 771–780. ACM, 2010.

[8] Xing Wei and W Bruce Croft. LDA-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 178–185. ACM, 2006.

[9] Xiaoxin Yin and Sarthak Shah. Building taxonomy of web search intents for name entity queries. In *Proceedings of the 19th international conference on World wide web*, pp. 1001–1010. ACM, 2010.

[10] 齋藤智輝, 豊田樹生, 夜久真也, 岩澤宏希. Web 検索クエリに対する教師なしエンティティリンキング. 言語処理学会第 24 回年次大会発表論文集, pp. 412–415, 2018.