ウェブ検索ログとWikipedia内部リンクを用いたエンティティの曖昧性解消

石川 裕貴　　小林 健　　長田 誠也
ヤフー株式会社
{hishikaw, kenkoba, sosada}@yahoo-corp.jp

1 はじめに

テキスト中に出現するエンティティ(地名、人物、組織などの実存する概念) を表す表記を特定し、知識ベース内の対応するエンティティと結びつけるタスクはEntity Linkingと呼ばれ、NISTやマイクロソフトがコンテストを開催するなど、近年世界的に注目を集めている。知識ベースとしてはWikipediaが用いられることが多く、その場合は特にWikificationと呼ばれる[1]。Entity Linkingにおいて、表記が多義語の場合にはエンティティ候補が複数得られるため、曖昧性解消が必要となる。エンティティの曖昧性解消は、Entity Linkingにおける重要な課題の一つである。

本研究では、Wikificationにおけるエンティティの曖昧性解消に取り組む。知識ベースとしてWikipediaを用い、Wikipediaの各ページをエンティティとみなす。

Wikificationにおけるエンティティの曖昧性解消にあたっては、Wikipedia内の情報のみを利用するのが一般的である。そうした手法の問題点の一つとして、Wikipedia内のデータの偏りによる悪影響があげられる。例えば、「川崎」という表記は、一般のテキストでは「川崎市」のエンティティを指すことが多いと思われるが、Wikipedia内で張られるリンクの統計量のみを使うと、「川崎市」よりも「川崎フロンターレ」を指す確率のほうが大幅に大きいという誤った結果が得られてしまう。一方、ウェブ検索ログ (ウェブ検索クリックスルーログ) を見ると、クエリ「川崎」に対する検索結果の中で、「川崎市」のWikipediaページがクリックされる頻度が、「川崎フロンターレ」のWikipediaページがクリックされる頻度より大幅に高いことがわかる。ウェブ検索ログを利用することにより、Wikipedia内のデータの偏りを補正する効果が期待できる。

また、もう一つの問題点として、スパースネスの問題があげられる。一般のテキストには、Wikipediaに記載されていない語も多く出現するため、Wikipedia内の情報のみでは曖昧性解消のための手掛かりが十分得られないことも想定される。手掛かりとなるデータの量を補う意味でも、ウェブ検索ログの追加は精度向上に役立つと考えられる。

本研究では、Wikipedia内のデータのみを使った手法をベースラインとし、それに加えてウェブ検索ログを外部知識として利用した手法を提案手法として、曖昧性解消の精度が向上することを示す。

2 関連研究

Mihalceaら[1]は、テキストから特徴的な表記を抽出する処理と、抽出された表記に対してエンティティを決定 (曖昧性解消) する処理の2つを組合せてWikificationを実現している。曖昧性解消では、Wikipedia内のデータを元に、主に文脈に関する特徴量を考慮した手法を採用している。

Milneら[2]の手法は、文脈を考慮したエンティティの出現確率 (文脈確率) に加えて、表記とエンティティが対応する確率 (表記確率) を利用して曖昧性解消を行う点で本研究と類似している。Milneらの研究では、上記2種類に文脈の信頼度を加えた計3種類の特徴量を元に、機械学習を使った曖昧性解消を行っている。

日本語を対象にした研究としては、黒川ら[3]の手法が、文脈の類似度と表記確率を組合せたスコア関数を定義している点で本研究と類似している。確信度のスコアに応じて文脈幅を調整することにより曖昧性解消の精度を向上させることに成功している。

上記の研究はいずれも曖昧性解消にあたってWikipedia内の情報のみを用いているが、本研究ではそれに加えてウェブ検索ログを考慮した曖昧性解消を行う。また、いずれもWikipediaのページを対象に評価を行っているが、本研究では異なるドメインのテキストとしてニュース記事を対象に評価を行う。

3 曖昧性解消のための知識獲得

本研究では、曖昧性解消のための知識源として、Wikipedia 内部リンク及びウェブ検索ログの 2 種類のリソースを用いる。2 種類のリソースそれぞれから、後述の表記確率と文脈確率を求めることにより、合計 4 種類の知識を事前に獲得する。本節では、4 種類の知識獲得を行う手法について述べる。

本研究では、ベースラインとして Wikipedia 内部リンクから獲得した表記確率と文脈確率の 2 種類の知識を使った手法、提案手法として全 4 種類の知識を使った手法を用いて評価を行う。

3.1 Wikipedia 内部リンクからの知識獲得

3.1.1 表記確率の獲得

表記に対するエンティティの使われ易さには偏りがあると考えられ、その偏りを考慮して曖昧性解消の精度向上を図ることは広く行われている [2][3]。本研究でも、この偏りを考慮した曖昧性解消を行う。

Wikipedia には、あるページから関連する他のページへの内部リンクが張られている。内部リンクの中には、[[広島東洋カープ | 広島]] のような縦線付きのリンクがあり、「広島」がアンカーテキスト、「広島東洋カープ」がリンク先のページタイトルを表す。アンカーテキストとリンク先を計数することにより、ある表記があるエンティティとして使われる確率値を求める。

Wikipedia 内部リンクから獲得した表記確率を $P_{ws}(entity|surface)$ とする。

3.1.2 文脈確率の獲得

エンティティの曖昧性解消において、文脈の情報は重要な手掛かりとなる。例えば、表記「広島」に対し「広島東洋カープ」と「広島市」がエンティティ候補として得られたとき、文脈に「阪神タイガース」「野球」のエンティティがあれば、前者を選択すべきと考えられる。そこで、Wikipedia 内部リンクを利用してエンティティ間の関連度をスコア化し、その知識をエンティティの曖昧性解消に利用することが考えられる。

本研究では、Wikipedia 内のリンク/被リンク数及び、Wikipedia の全ページを対象に、ある固定サイズのウィンドウ内でのリンク同士の共起頻度を計数し、エンティティ間の関連度を確率的に表す。

Wikipedia 内部リンクから獲得した文脈確率を $P_{wc}(context\_entity|entity)$ とする。

3.2 ウェブ検索ログからの知識獲得

ウェブ検索ログ (ウェブ検索クリックスルーログ) の例を表 1 に示す。クエリと、検索結果の各 URL に対するクリック数を集計したデータである。本研究では、2013 年の Yahoo!検索の全ログのうち Wikipedia に遷移しているクリックを集計してデータを作成した。このデータのうち、例えばクエリが「広島」のデータを使うことで、表記「広島」は「広島市」の意味で使われやすいことや、クエリが「広島 阪神」となっているデータを使うことで、表記「広島」の文脈に表記「阪神」が出ている場合に、「広島」は「広島東洋カープ」の意味で使われている可能性が高いことなどがわかる。

本節では、ウェブ検索ログを使って表記確率と文脈確率を獲得する方法について述べる。

表 1: ウェブ検索ログの例

クエリ	遷移先 URL	clicks
広島	http://ja.wik.../広島市	12,748
広島	http://ja.wik.../広島東洋カープ	384
...	...	...
広島 阪神	http://ja.wik.../広島東洋カープ	135
広島 阪神	http://ja.wik.../広島市	2
...	...	...
広島 気候	http://ja.wik.../広島市	160
広島 気候	http://ja.wik.../広島東洋カープ	1
...	...	...

3.2.1 表記確率の獲得

表記確率の算出には、ウェブ検索ログの中で、クエリが 1 つのタームからなっているもの (スペースを含まないもの) を用いる。対象のクエリを表記、遷移先 URL をエンティティとして、条件付き確率を求める。

ウェブ検索ログから獲得した表記確率を $P_{cs}(entity|surface)$ とする。

3.2.2 文脈確率の獲得

文脈確率の算出には、ウェブ検索ログの中で、クエリが 2 つのタームからなっているもの (スペースを 1 つ含むもの) を用いる。

以下、表 1 のうち、「広島 阪神」から「広島東洋カープ」に遷移している行の例をとって説明する。まずはクエリに含まれるターム「広島」、「阪神」のうち、「広島東洋カープ」を「広島」と対応付ける。これは、上

記の表記確率が存在する表記と対応付けることにより行う。対応付かなかったターム「阪神」を、「広島東洋カープ」の文脈表記としてカウントする。上記をデータ全体に対して集計して確率値を求めることにより、エンティティ(遷移先 URL) に対する文脈表記の文脈確率を得る。

ウェブ検索ログから獲得した文脈確率を $P_{cc}(context\_surface|entity)$ とする。

4 曖昧性解消の手法

エンティティの曖昧性解消の全体像を図 1 に示す。以下、曖昧性解消の対象となる表記 s_t に対する処理を例にとって説明する。

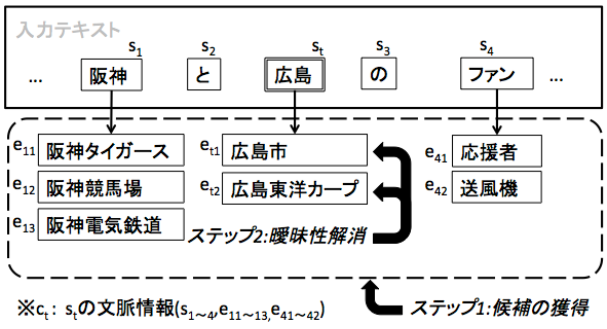

図 1: エンティティの曖昧性解消の全体像

1 つめのステップとして、入力テキスト中の各表記に対しエンティティの候補を獲得する。

2 つめのステップとして、s_t に対するエンティティ候補の中で、確信度スコアが最高のものをエンティティの曖昧性解消結果として出力する。確信度スコアは、前節で求めた 4 種類の確率値を使って算出する。

本節では、それぞれのステップの詳細を述べる。

4.1 ステップ 1:エンティティ候補の獲得

入力文書中の各表記に対するエンティティ候補は、表記確率が存在するエンティティを列挙することにより獲得する。

例えば、図 1 中の「広島」に対して、Wikipedia 内部リンクとウェブ検索ログから得られた表記確率のデータを参照して、確率値を持つ「広島市」、「広島東洋カープ」を候補とする。

4.2 ステップ 2:エンティティの曖昧性解消

図 1 において、曖昧性解消の対象となる表記 s_t と文脈 c_t が与えられた時に、確率 $P(e_{ti}|s_t, c_t)$ が最も大きくなるエンティティ e_{tx} を曖昧性解消の結果として出力することを考える。

$$x = \arg\max_i P(e_{ti}|s_t, c_t) \tag{1}$$

$P(e_{ti}|s_t, c_t)$ は各種独立性の仮定、i に依存しない確率値の削除を行うことにより以下のように展開できる。

$$\begin{aligned} P(e_{ti}|s_t, c_t) &= \frac{P(s_t, c_t|e_{ti})P(e_{ti})}{P(s_t, c_t)} \\ &= \frac{P(s_t|e_{ti})P(c_t|e_{ti})P(e_{ti})}{P(s_t, c_t)} \\ &\propto P(s_t|e_{ti})P(c_t|e_{ti})P(e_{ti}) \\ &= \frac{P(e_{ti}|s_t)P(s_t)}{P(e_{ti})}P(c_t|e_{ti})P(e_{ti}) \\ &= P(e_{ti}|s_t)P(s_t)P(c_t|e_{ti}) \\ &\propto P(e_{ti}|s_t)P(c_t|e_{ti}) \end{aligned} \tag{2}$$

$P(e_{ti}|s_t)$ は前述の表記確率を使って近似する。

$$P(e_{ti}|s_t) \approx P_{ws}(e_{ti}|s_t)P_{cs}(e_{ti}|s_t) \tag{3}$$

$P(c_t|e_{ti})$ は前述の文脈確率を使って近似する。

$$P(c_t|e_{ti}) \approx \prod_l \max_m P_{wc}(e_{lm}|e_{ti}) \prod_n P_{cc}(s_n|e_{ti}) \tag{4}$$

上記の結果を踏まえ、提案手法では確信度スコア $Score(e_{ti}, s_t, c_t)$ が最大となるエンティティを選択する事によりエンティティの曖昧性解消を行う。

$$x = \arg\max_i Score(e_{ti}, s_t, c_t) \tag{5}$$

$$\begin{aligned} Score(e_{ti}, s_t, c_t) &= (P_{ws})^\alpha \cdot (P_{cs})^\beta \cdot \\ &\quad (\prod_l \max_m P_{wc})^\gamma \cdot (\prod_n P_{cc})^\delta \end{aligned} \tag{6}$$

ここで α、β、γ、δ はそれぞれの確率値をどの程度考慮するかのパラメータである。また、各確率値はラプラス法によりスムージングを行う。

5 実験

5.1 データセットの作成

Yahoo!ニュースのスポーツカテゴリの、2013 年の全記事から 70 記事をランダムサンプリングし、パラメータ推定及び評価に使うデータセットを作成した。Wikipedia 内で「曖昧さ回避ページ」のタイトルになっている表記を、曖昧性を解消すべき多義語とみなし、記事中に出現する多義語に対して、人手で正解のエンティティを付与することにより作成した。

作成したデータセットに対して、半分の 35 記事を式 (6) の α、β、γ、δ のパラメータ推定用、残りの半分を評価用のデータセットとした。

5.2 実験結果

ベースラインとしては、従来の Wikipedia 内部リンクの情報のみを用いる手法で実験を行った。式 (6) のパラメータのうち、β と δ を 0 に固定して、α と γ をパラメータ推定用データセットを使って推定し、評価用データセットを使って精度評価を行った。

提案手法としては、Wikipedia 内部リンクの情報に加えてウェブ検索ログの情報を考慮する手法で実験を行った。ベースラインと α、γ を揃えて、β、δ の推定を行い、精度評価を行った。

精度評価結果を表 2 に示す。ベースラインと比較して提案手法のほうが高い精度を得られていることがわかる。β、δ の双方に 0 よりも大きい定数を与えたほうが精度が高くなっており、ウェブ検索ログを追加した効果が得られている。

表 2: 精度評価結果

モデル	α,β,γ,δ	精度 [Accuracy]
ベースライン	1, 0, 4, 0	63.70% (286/449)
提案手法	1, 1, 4, 1	66.36% (298/449)

提案手法での改善例を図 2 に示す。Wikipedia 内部リンクに基づく表記確率及び文脈確率はともに、「西武園競輪場」のほうが「西武園ゆうえんち」より大きな値を持つ。一方、ウェブ検索ログに基づく表記確率及び文脈確率はともに、「西武園ゆうえんち」のほうが「西武園競輪場」よりも大きな値を持つ。双方の知識を統合することにより、「西武園ゆうえんち」のほうが「西武園競輪場」よりも確信度スコアが高くなり、正しい結果が得られる。

> ... W久保英恵（３０）が３日、埼玉・西武園でトークショーを行った。スケートリンク...
>
> 正解:「西武園ゆうえんち」
> ベースライン:「西武園競輪場」
> 提案手法:「西武園ゆうえんち」

図 2: 改善例

提案手法での改悪例を図 3 に示す。Wikipedia 内部リンクに基づく情報を使った場合、表記「千葉」に対する「ジェフユナイテッド市原・千葉」及び「千葉県」の表記確率は同程度となるため、文脈確率が大きい「ジェフユナイテッド市原・千葉」を正しく選択できている。一方、ウェブ検索ログに基づく情報を使った場合、表記「千葉」に対する「千葉県」の表記確率のほうが、「千葉」に対する「ジェフユナイテッド市原・千葉」の表記確率よりも大幅に大きくなるため、「千葉県」が選ばれるという誤った結果となる。ウェブ検索ログに基づく情報を、より適切なバランスで確信度スコアに反映することは今後の課題である。

> ... 藤田氏は、磐田や名古屋、熊本、千葉で活躍し、オランダのユトレヒトにも在籍した。そ...
>
> 正解:「ジェフユナイテッド市原・千葉」
> ベースライン:「ジェフユナイテッド市原・千葉」
> 提案手法:「千葉県」

図 3: 改悪例

6 おわりに

本研究では、Wikification におけるエンティティの曖昧性解消において、Wikipedia 内の情報のみを利用した手法と、それに加えてウェブ検索ログを外部知識として利用した手法を使って比較実験を行い、後者の曖昧性解消の精度の方が高くなることを示した。

本研究で用いたデータセットはパラメータ推定用、評価用それぞれ 35 記事と大きくはないため、統計的有意性を検証するには十分でないと思われる。今後の課題として、より大規模なデータセットを使った検証を行いたい。また、ウェブ検索ログの情報をより適切に確信度スコアに反映する方法についても検討を進めたい。

参考文献

[1] Rada Mihalcea and Andras Csomai. Wikify!: Linking Documents to Encyclopedic Knowledge. In *Proc. of CIKM '07*, pp. 233–242, New York, NY, USA, 2007. ACM.

[2] David Milne and Ian H. Witten. Learning to Link with Wikipedia. In *Proc. of CIKM '08*, pp. 509–518, New York, NY, USA, 2008. ACM.

[3] 黒川, 新里, 黒橋. 段階的文脈拡張による多義性解消. 言語処理学会第 17 回年次大会, pp. 544–54, 2011.