E_008

ウェブ・ページ内での共起を使った曖昧性解消
Disambiguation based on Co-occurrences within Web Pages

隅田　英一郎[1]　菅谷　史昭[2]
Eiichiro SUMITA　Fumiaki SUGAYA

1.はじめに

同形異音語とは、読みが複数ある単語のことである。読みの識別は音声合成では不可欠である。また、頭字語（アクロニム）は複数の単語からなる表現（本稿では、定義と呼ぶ）の省略形である。頭字語も通常その定義が複数あり、この識別も検索や翻訳などのアプリケーションとって重要である。

本稿では、上述の例ように単語とある表現の対応に存在する曖昧性に共通して利用できる解消方法を提案する。

提案手法では、ウェブから学習データを取得し、機械学習プログラムで分類器を生成して、これにより曖昧性を解消する。まず同形異音語と提案法と実験結果について、次に頭字語と実験結果について述べ、さらに提案法の課題や関連研究について議論した後、論文をまとめる。

2.同形異音語

音声合成では同形異音語の処理は重要である。同形異音語とは表記が同じで読みが異なる単語である。例えば、"bow" (ちょう形リボン) と "bow" (船首)は、英語の典型的な例である。

日本語の音声合成は 10%ほどの割合で読み誤るという報告がある(Umemura and Shimizu, 2000). この問題の主要な原因の一つに同形異音語の存在があり、Yarowsky (1996)、Li and Takeuchi (1997)、Umemura and Shimizu (2000)などが同形異音語の曖昧性解消手法を提案している。

本論文では、地名などの固有名詞に同形異音語が多いことを考慮して、固有名詞の処理に着目する。上述の先行研究では、読みは品詞または語義と連動すると仮定し、機械学習に基づく形態素解析や意味タグ付けによって読みを決定する手法を提案している。しかし、地名の場合は、品詞が「固有名詞」と分かっても意味タグが「場所」と分かっても読みは決定できない。提案手法は先行研究と同様に機械学習を用いているが、主要な違いは知識源にある。先行研究は、学習用コーパスに人手で読みを付与するため時間とコストが嵩むが、提案法では、ウェブの散在するページから学習データを取得するためコストはほとんどかからない。

2.1.提案手法

本稿では日本語に着目する。日本語は、同じ単語に複数の表記方法がある。すなわち、漢字、カタカナ、ひらがなである。後の二つは読みを表す。

2.2.手法 1

着想は地名の漢字による表記とカタカナ（または、ひらがな）による読みの表記が一つのウェブ・ページに頻繁に共起するという観察に基づく。

図 1に例をあげた。地名の漢字表記“大平”（実線の楕円でマークしてある）とカタカナによる読み表記“オオダイラ”（点線の楕円でマークしてある）が一つのページの近傍に共起している。Google によれば, 464 個のページで“大平”と“オオダイラ”は共起している。

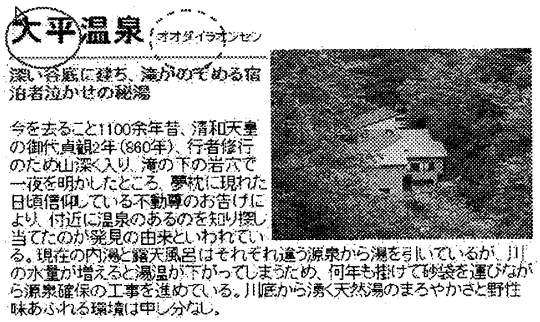

図 1　漢字とカタカナのウェブページ内での共起

手法 1 はページ・ヒットを直接利用する。すなわち、最もページ・ヒットが多い読みを選択する。

2.3.手法 2

手法 1 は選択した候補以外を無視しており、曖昧性解消と言い難い。そこで、次に述べるように、共起したページ内の当該単語の近傍の特徴を抽出し、これを訓練データとして、分類器を機械学習で作成し識別する。

2.3.1 ウェブからの学習データの取得

手続を図 2に示す。入力は単語 W と読み候補の集合$\{R_k \mid k=1\sim K\}$である。

実験では L を 1,000 としたので、それぞれの読み R_k 毎に高々 1,000 個の訓練データ $\{T_l(W)\}$が得られる。

For all k =1~K **do**
- i)　クエリ「W **and** R_k」でウェブを検索する。
- ii)　W と R_k を含むスニペットの集合$\{S_l(W, R_k) \mid l=1\sim L\}$を取得する。
- iii)　S_l から R_k を削除し、学習データの集合$\{(T_l(W), R_k) \mid l=1\sim L\}$を取得する。

end

図 2　ウェブからの学習データの取得

[1] NiCT & ATR SLC
[2] KDDI Labs & ATR SLC

2.3.2 分類器の訓練

訓練データ $T_l(W)$から featureベクトルを作成し、読み R_k とともに、決定木の機械学習アルゴリズム[1]に入力する。$T_l(W)$ を $W_{-m}\ W_{-(m-1)}\ \ldots\ W_{-2}\ W_{-1}\ W\ W_1\ W_2\ \ldots\ W_{m-1}\ W_m$, と書く。ここで m は 2 から M（ウインドウ・サイズ）まで動く。ウインドウ中のキーワードの有無を feature とする。ここで、キーワード[2]とは、{ $T_l(W)$ }全体での単語の頻度を調べ、上位N語のこととした。

2.4.実験データ

ここでは地名の実験について報告する。

2.4.1 曖昧な地名のリスト

郵政公社は住所とその読みを含む郵便番号データを公開している。そこから 79,861 件の地名と読みのペアからなる地名リストを抽出した。5.7% が複数の読みをもっており曖昧であった。曖昧な単語に関して平均 2.26 個の読みがある。ウェブ上での出現頻度を加味すると、地名の出現の約 1/4 が曖昧である。

提案法は W と R のウェブページ上での共起に基づいているで、正しいペアの共起がないと動作しない。79,861 件の地名と読みのペアの中で同一ウェブページ上で一度も共起しないペアは 1 件のみであった。この意味で、カバー率はほぼ 100%である.

2.4.2 オープンデータ

実験は EDR コーパスを用いた。日本語の新聞記事からなり、形態素解析され、品詞と読みが振られている。本データはウェブ上に公開されていないので学習データと重ならない。上記の曖昧な地名リストに出現する単語を含む文[3]を抽出した。

2.5.実験結果

2.5.1 オープンテスト

まず、ページ・ヒットを直接利用する手法 1 を評価した(表 1)。読みの表記には 2 種類、ひらがなとカタカナとあるが、いずれの場合も高い精度を示した。

表 1　（ページ・ヒットを直接利用する）手法 1 の精度

読み表記	Accuracy
ひらがな	89.2
カタカナ	86.6

次に分類器を作成する手法 2 を評価した(表 2)。手法 1 に比べて、全てのウインドー・サイズにおいて、手法 2 がより高い精度を示している。ウインドー・サイズが大きい方が高精度で、M=10 で手法 1 を約 3.5% 上回っている（誤り削減率では 30%前後）.

表 2　（分類器を作成する）手法 2 の精度

読み表記	M=2	M=5	M=10
ひらがな	89.9	90.3	**92.9**
カタカナ	89.2	88.4	**89.9**

2.5.2 曖昧度と精度

ここでは、読みの曖昧度と精度の関係を調べた（「ひらがな」のクロス・バリデーションテスト）。

平均的な場合を調べるために、地名リストから 20 種類の単語をランダムに選択した[4]。平均の曖昧度は 2.1 である。手法 2 の平均精度は、90%前後であり、手法 1 より高い精度を示している。

表 3　平均的曖昧度の場合の手法 2 の精度

曖昧度	M=2	M=5	M=10	手法 1
2.1	**89.2 %**	**90.9 %**	**92.3 %**	67.5%

最も曖昧な場合を調べるために、地名リストから曖昧度の上位 20 位までの単語を選択した[5].平均曖昧度は 7.1 である。予想される通りに、手法 2 の性能は平均的な曖昧度の場合より低いが、それでも、約 70% から約 80 %と高い。また、手法 1 より高い精度を示している。

表 4　最も曖昧な場合の手法 2 の精度

曖昧度	M=2	M=5	M=10	手法 1
7.1	**73.9 %**	**77.3 %**	**79.9 %**	57.5%

3.　頭字語

頭字語（アクロニム）は複数の単語からなる表現（ここでは、定義と呼ぶ）の省略形である。頭字語は大変便利で広く使われ、どんな分野でも自由に他分野と無関係に生み出される。従って、頭字語と定義の対応は通常曖昧である。例えば, "ACL" にも沢山の定義がある。 "Anterior Cruciate Ligament (膝の怪我)," "Access Control List (コンピュータセ

---

[1] Weka による。

[2] 実験では N=100 とした。例えば、単語「山北町」の場合、キーワードは「文頭, か, 南足柄, 県, 開成, 4, 郡, 番号, やま, やす, ／, 松屋, 参照, して, す, 【, 〒, きた, ベスト, 神奈川, ください, くま, 電話, 山北町, ちょう, 箱根, 六, かいせい, が, 松田, 港町, ー, 安田, トップ, 新潟, 横浜, ：, 公式, ホームページ, ら, 市, み, 丸亀, ろく, １３０１, みなとまち, あし, 区, だ, ）, 。, ぐん, だちょう, 宣伝, リンク, 足柄上, 大和, 足柄下, 郵便, 山北, た, 町, 湯河原, 住所, よい, ・, がら, かわ, 」, 行, あわ, 、, まち, 岩船, は, ばん, や, 文末, （, 弥生, 粟田, かな, ち, , ま, 役場, ぐち, さ, さんぽ, 千, で, 人, 香川, の, みて, まつ, 「, 情報, 番地, と, 粟田口, を, ページ, 検索, 】」であった。

[3] 268 箇所に当該地名が出現し、単語の異なり数は 72 であった.

[4] 東浜町, 三角町, 宮丸町, 川戸 ,下坂田, 蓬田, 金沢町, 白木町, 神保町, 助谷, 新御堂, 糸原, 駿河町, 百目木, 垣内田町, 杉山町, 百戸, 宝山町, 出来島, 神楽町。

[5]小谷, 上原町, 上原, 小原, 西原, 上町, 大平, 葛原, 平田, 馬場町, 新田, 土橋町, 大畑町, 上野町, 八幡町, 柚木町, 長田町, 平原。

キュリティの専門用語)," and "Association for Computational Linguistics (学会名)."

頭字語は長い定義を使わずに済ますのが目的なので、定義なしで使われることが多い（表 5）。結果、そのような頭字語を含むテキストを解析したり、検索したり、翻訳するためには、頭字語の曖昧性の解消が必要になる。

表 5　頭字語 ACL の出現例

ACL の定義	出現例
Anterior Cruciate Ligament	She ended up with a torn **ACL**, MCL and did some other damage to her knee.
	(http://aphotofreak.blogspot.com/2006/01/ill-give-you-everything-i-have-good.html)
Access Control List	Calculating a user's effective permissions requires more than simply looking up that user's name in the **ACL**.
	(http://www.mcsa-exam.com/2006/02/02/effective-permissions.html)
Association for Computational Linguistics	It will published in the upcoming leading **ACL** conference.
	(http://pahendra.blogspot.com/2005/06/june-14th.html)

逆に、頭字語は定義とよく共起する(図 3). 例えば、頭字語 ACL はその定義のひとつである「Association for Computational Linguistics」と 211,000 回共起する（google.com）。

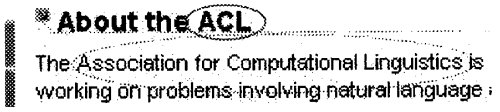

図 3　頭字語と定義のウェブページ内での共起

頭字語の曖昧性解消に、同形異音語で用いた方法（2.1 節の手法 1 と手法 2）が使える。頭字語の可能な定義のリストを作ることは、本提案の範囲外とする。実際、このリストの作成のためには、Nadeau and Turney (2005)などの先行研究があり高い性能が実現されている。また、すでに、この機能を提供するサイトがいくつもある。

3.1. 実験データ

まず、頭字語を Wikipedia から取得し、英字大文字以外を含むもの、または、3 文字より短いものを除いた。続いて、頭字語の定義を http://www.acronymsearch.com/ より取得し、5 種類未満の定義しか持たない頭字語も除外した。最後に、その中からランダムに 20 個の頭字語を選択した。

これを典型的な頭字語のリストとして、以下の実験を行った。

3.2. 実験結果

ここでは、定義の曖昧度と精度の関係を調べた。クロス・バリデーションテストを行った。

表 6　曖昧度が 2 の場合の手法 2 の精度

曖昧度	M=2	M=5	M=10	手法 1
2	**88.7 %**	**90.1 %**	**92.4 %**	82.3%

曖昧度が 2 の場合（表 6）90%前後の精度が得られた。手法 2 は手法 1 を上回っている。M はウインドウ・サイズであり、M が長いほど、精度は高い傾向が見られる。

表 7　曖昧度が 5 の場合の手法 2 の精度

曖昧度	M=2	M=5	M=10	手法 2
5	**78.6 %**	**82.6 %**	**86.0 %**	76.5%

曖昧度が 5 の場合（表 7）は曖昧度が 2 の場合よりは性能が下がるが、それでも約 80 %と高い。他は、曖昧度が 2 の場合と同様の現象が観察された。

4. 議論

4.1. データのバイアス

2 節、3 節の実験で見たように、平均性能では、手法 2 は手法 1 より性能が良い。しかし、個別に見ると逆転することがある。

なぜなら、手法 1 はサーチエンジンのお陰でウェブの全出現が考慮されているが、手法 2 では 訓練データ数は L で制限されている。L の制限をなくすと、全共起データを用いて、分類器を作成することになり、データ量や処理時間などの観点から実用的でなくなる。頭字語の実験では、平均訓練データ数は 830 であり、訓練データの分布は平坦である。

これが分類器を誤動作させることがある。例えば、頭字語"ISP" の最大頻度の定義は 99.9% のシェアがあり非常に分布が急峻である(表 8) 。一方、訓練データの分布は先に述べたように平坦 である。従って、手法 2 が手法 1 に劣る結果になる。

表 8　ISP の分布

Definition	Page hits
Internet Service Provider	3,590,000
International Standardized Profile	776
Integrated Support Plan	474
Interactive String Processor	287
Integrated System Peripheral control	266

頭字語 CEC"の最大頻度の定義は 26.3% のシェアであり分布が平坦である (表 9)。訓練データの分布が実データの分布に類似していると言える。この場合は分類器がよく働き、手法 2 が手法 1 に勝る。

図 4 にあるように、手法 2 が勝る場合に大きなゲインが観測され、逆の場合には小さな劣化が観測される。

データのバイアスによる手法２の不具合は、本実験のように単純に決定木を学習するのではなく、データのバイアスを考慮した機械学習手法を採用することで克服できると想定される。

表９ CEC の分布

Definition	Page hits
California Energy Commission	161,000
Council for Exceptional Children	159,000
Commission of the European Communities	138,000
Commission for Environmental Cooperation	77,400
Cation Exchange Capacity	76,400

4.2. 訓練データと実データの関係

訓練データは W と R が共起するデータであり、実データは共起する場合もしない場合もある。つまり、訓練データと実データが似ている保障はない。従って、学習して得られた分類器が実データをうまく処理できるとは限らない。ただ、2.5.1 節のオープンデータを用いた同形異音語の識別実験では良い結果が得られており、上記の懸念が必ずしも当たらない可能性が示唆される。

4.3. 関連研究

近年、ウェブをコーパスとして活用する手法が盛んに研究されている(Kilgarriff and Grefen-stette, 2003)。本提案の手法１と同様に、ページ・ヒットを利用する手法が主流である。

これらと異なり、本提案の手法２と Sarikaya (2005)による Web-based Language Modeling と Mihalcea (2002)による Boot-strapping Large Sense-Tagged corpora は、ページの内容を利用している。

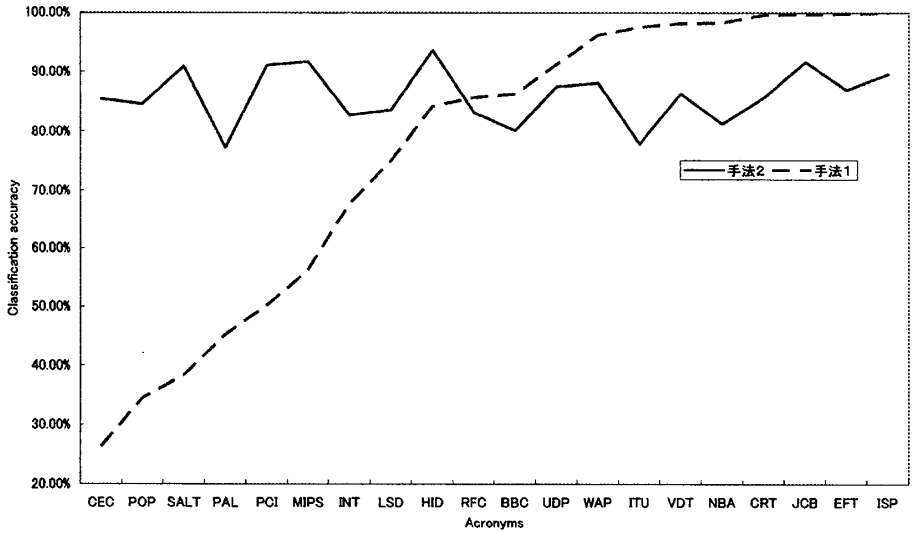

図４　データのバイアスと二つの手法の比較

5. まとめ

本論文では、単語の曖昧性解消に、ウェブ・ページ内の共起データを利用して、単語の周辺の情報を考慮する手法を提案した。

日本語の地名と読みの対応付けという課題と頭字語と定義の対応付けの課題という異なる問題に提案手法を適用して、高い精度を確認した。

学習データを取得する際に課さざるを得ない制約により、学習データと実データの分布にズレが生じる場合に性能が出ないことがあったが、機械学習方法を変更すれば解消できると想定している。

参考文献

Hang. Li and Jun-ichi Takeuchi. 1997. Using Evidence that is both string and Reliable in Japanese Homo-graph Disambiguation, SIGNL119-9, IPSJ.

Yoshiyuki Umemura and Tsukasa Shimizu. 2000. Japa-nese homograph disambiguation for speech synthe-sizers, Toyota Chuo Kenkyujo R&D Review, 35(1):67-74.

David Yarowsky. 1996. Homograph Disambiguation in Speech Synthesis. In J. van Santen, R. Sproat, J. Olive and J. Hirschberg (eds.), Progress in Speech Synthesis. Springer-Verlag, pp. 159-175.

A. Kilgarriff and G. Grefenstette. 2003. Introduction to the special issue on the Web as a corpus. Computational Linguistics 29(3): 333-348.

Rada. F. Mihalcea, 2002. Bootstrapping Large Sense-Tagged Corpora, Proc. of LREC, pp. 1407-1411.

David Nadeau and Peter D. Turney, 2005. "A supervised learning approach to acronym identification," 18th Canadian Conference on Artificial Intelligence, LNAI3501.

Ted Pedersen and Rada. F. Mihalcea, Advances in Word Sense Disambiguation, tutorial at ACL 2005. http://www.d.umn.edu/~tpederse/WSDTutorial.html.

Ruhi Sarikaya, Hong-kwang Jeff Kuo, and Yuqing Gao, 2005. Impact of Web-Based Language Modeling on Speech Understanding, Proc. of ASRU, pp. 268-271.