社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2006－ＮＬ－174（３）
2006／7／27

ウェブ・ページ内での共起を使った同形異音語処理

隅田英一郎† 菅谷史昭‡

† NiCT & ATR 〒619-0288 京都府けいはんな学研都市光台 2-2-2
‡ KDDI 研究所 〒356-8502 埼玉県埼玉県上福岡市大原 2-1-15
E-mail: † eiichiro.sumita@nict.jp, ‡ fsugaya@kddilabs.jp

あらまし 本論文では、単語の曖昧性解消に必要な知識をウェブ・ページ上の単語周辺の情報に求め、これを学習データとして分類器を構成する手法を提案した。日本語固有名詞の同形異音語を対象とした読みの決定という問題に提案手法を適用して高い精度を確認した。さらに、頭字語とその定義の対応付けという別の問題でも有効性を確認した。学習データを取得する際に課さざるを得ない量的制約に起因して学習データと実データの分布に大きくズレが生じる場合があり性能が劣化することが観測されたが、これは今後の課題と考えている。

キーワード 同形異音語，発音，頭字語，多義解消，音声処理、ウェブに基づく自然言語処理

Heteronym disambiguation based on co-occurrence within Web page

Eiichiro Sumita† Fumiaki Sugaya‡

† NiCT & ATR 2-2-2 Hikaridai, Keihannna, Kyoto 619-0288, JAPAN
‡ KDDI R&D Labs 2-1-15 Ohara, Kamifukuoka-shi, Saitama 356-8502, JAPAN
E-mail: † eiichiro.sumita@nict.go.jp, ‡ fsugaya@kddilabs.jp

Abstract The authors claim that information around a word on the Web is useful for solving ambiguity related to the word. The proposed method learns a classifier based on the information. Appling the method to heteronym disambiguation demonstrated a high accuracy. Furthermore, it was effective for acronym disambiguation. However, when distribution of training data and that on the Web are largely different, we observed degradation, which is one of future works to be tackled.

Keyword Heteronym, Pronunciation, Acronym, Word Sense Disambiguation, Speech Processing, Web-based NLP

1. はじめに

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。同形異音語は稀ではない。例えば、本論文の実験で用いた日本語の地名辞書だけで、約 4,500 件ある[1]。

一方、読みはテキストを音声に変換する音声合成システムにとって不可欠の情報である。また、音声をテキストに変換する音声認識システムやテキストを外国語のテキストに変換する機械翻訳システムにおいても必要となる。

任意の入力を対象とするシステムでは、特に、固有名詞が課題となる。地名や人名などの固有名詞は同形異音語が多いにもかかわらず、処理に利用可能なメモリ量の制限などから、固有名詞はシステムの知識に登録されないことが多い。音声合成システムであれば、読めないことが頻繁に起こることになる。

[1] 固有名詞に同形異音語が多い理由は次のように推察される。日本語の漢字は１字ごと複数の読みがありうる。これが連なってできる単語の場合、各漢字の読みの全ての組み合わせの可能性がある。さらに、連濁や「の」の挿入など、さらに異音を増やす要因がある。また、地名や人名は、地域や家族など小さな集団の中で決定されることも多い。従って、独立に様々な読みが与えられ、同形異音の曖昧性が生じると考えられる。
例えば、「大平」の場合、「大」が「だい」「たい」「おお」と少なくても３通りに読め、「平」が「へい」「べい」「ひら」「びら」「たいら」「だいら」と６通りに読め、組み合わせとして１８通りあり、実際に「だいへい」、「たいへい」、「おおびら」、「おおひら」、「おおだいら」の５つが使われている。

また、頭字語（アクロニム）でも、同形異音語と同様の問題が生じる。頭字語と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。頭字語もその定義が複数になることが多く、この曖昧性を解消することが、機械翻訳システムやサーチ・エンジンなどのアプリケーションとって重要となる。

本稿では、上述の２つの現象のように、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。

本論文では、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手法を提案する。まず、同形異音語の問題と提案法と実験結果について、次に頭字語の問題と実験結果について述べ、さらに提案法の課題について議論し、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。

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[1]. 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、同形異音語の読み分けのために、Yarowsky [2]、 Li and Takeuchi [3]、Umemura and Shimizu [1]などが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人的コストはかからない。

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]。

2.1.1. 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。ウェブ・ページの作成時に、複数の読みがある語に関して注意が向いて、読み情報を入れることが多くなるのではと推測される。

図 1 に例[3]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 のサーチエンジンによれば, 464 個のページで “大平” と “オオダイラ” は共起している。

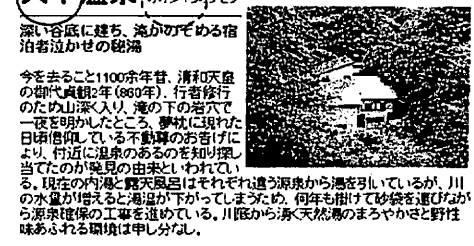
大平温泉（オオダイラオンセン）
深い谷底に建ち、滝かのぞめる宿
泊者泣かせの秘湯
今を去ること1100余年昔、清和天皇の御代貞観2年(860年)、行者修行のため山深く入り、滝の下の岩穴で一夜を明かしたところ、夢枕に現れた日頃信仰している不動尊のお告げにより、付近に温泉のあるのを知り探し当てたのが発見の由来といわれている。現在の内湯と露天風呂はそれぞれ違う源泉から湯を引いているが、川の水量が増えると湯温が下がってしまうため、何年も掛けて砂袋を運びながら源泉確保の工事を進めている。川底から湧く天然湯のまろやかさと野性味あふれる環境は申し分なし。

図 1 漢字とカタカナのページ内での共起

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。

2.1.2. 手法 2

手法 1 は選択した候補以外を無視しており、曖昧性解消とは言えない。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（当該の読みが使われるための条件あるいは文脈とみなせる）を抽出し、これを訓練データとして、当該単語に対して分類器を機械学習で作成する手法を提案する。

2.1.3. 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。読み候補は人手で作成した辞書でも良いし、自動的に作成してもよい。

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

訓練データ数を L で制限する。訓練データ数を制限せずに全てのヒットしたページを使う選択もありうる。

[2] 現代の日本語において、カタカナは主に外来語の表記に、ひらがなは地の文の表記に使われることが多いなど、若干の用法上の相違が見られる。

しかし、検索ヒット数は100万件を頻繁に超える性質を持ち、メモリ容量や学習時間など観点から現実的とは言えないことと、現在提供されている検索エンジンAPIはダウンロード数に上限があることを考慮して、訓練データ数を制限する。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。

ステップ iii)で、スニペット S_l から読み R_k を削除しているのは、学習で得られる分類器は、入力データに読みが含まれていない場合に動作する必要があるからである。

2.1.4. 分類器の訓練

訓練データ $T_l(W)$から特徴ベクトルを作成し、読み R_k とともに、決定木の機械学習アルゴリズム[4]に入力する。

$T_l(W)$ を $W_{-m}\ W_{-(m-1)} \cdots W_{-2}\ W_{-1}\ W\ W_1\ W_2 \cdots W_{m-1}\ W_m$, と書く。ここで m は 2 から M（ウインドウ・サイズ）まで動く。特徴ベクトルでは、ウインドウ中のキーワードの有無をビットで表わす。

```
香取 = Y: さわら (42.0/1.0)
香取 = N
|   白楊 = Y: さわら (31.0)
|   白楊 = N
|   |   銚子 = Y: さわら (28.0)
|   |   銚子 = N
|   |   |   山車 = Y: さわら (27.0)
|   |   |   山車 = N
|   |   |   |   北 = Y: さわら (18.0)
|   |   |   |   北 = N
|   |   |   |   |   千葉 = Y
|   |   |   |   |   |   1 = Y: さはら (2.0)
|   |   |   |   |   |   1 = N: さわら (67.0/1.0)
```

図 3 決定木の例

単語 W に関する学習データを全部集め、その中に含まれる単語の頻度順に上位 N 語[5]をここではキーワードとした。

また、図 3に単語「佐原」の読み「さはら」と「さわら」との曖昧性解消の決定木の一部を示す。

2.2. 実験データ

ここでは地名の実験について報告する。

2.2.1. 曖昧な地名のリスト

郵政公社が公開している住所とその読みを含む郵便番号データ[6]を利用した。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（表 1)。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。ウェブ上でのヒット数を考慮して推定するとウェブ上の地名の約 1/4 が曖昧である。

表 1 地名の曖昧性（郵便番号データの場合）

読みの種類	語数	%
1	70,232	94.3
2	3,443	5.7
3	599	
4	150	
5	45	
6	11	
7	4	
8	2	
11	1	
total	74,487	100.0

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、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一度も共起しないペアは 1 件のみであった。この意味で提案法は実現可能である.

2.2.2. オープンデータ

実験には EDR コーパス[7]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を抽出した。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.

2.3. 実験結果

2.3.1. 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 2)。

表 2 （ヒット数を利用する）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 3)。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

[3] http://oyudokoro.mimo.com/area/C/cd/tng/000370/
[4] http://www.cs.waikato.ac.nz/~ml/weka/
[5] 実験では N=100 とした。
[6] http://www.post.japanpost.jp/zipcode/
[7] http://www2.nict.go.jp/kk/e416/EDR/index.html

イズが大きい方が概して性能が良く、M=10 で約 3.5% 以上手法 1 より精度が高い（誤り削減率では 30%前後である）。

「ひらがな」の方が「カタカナ」より、概して性能が良いが、その理由については今後分析を進めて行きたい。

表 3 （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

2.3.2. 曖昧度と精度

ここでは、読みの曖昧度と手法 2 の精度の関係を調べた（「ひらがな」表記を使ったクロス・バリデーションテストを実施）。

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[8]。平均の曖昧度は 2.1 である。ほぼ 90%の精度を達成している。

表 4 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
2.1	89.2 %	90.9 %	92.3 %

次に、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順に、上位 20 位までの単語を選択した[9]. 平均曖昧度は 7.1 である。予想される通りに、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。

表 5 最も曖昧な場合の手法 2 の精度

曖昧度	M=2	M=5	M=10
7.1	73.9 %	77.3 %	79.9 %

2.3.3. 括弧を利用した訓練データ取得

訓練データの取得は、語と読みが共起するという条件で検索している。「読み」といっているが、「読み」としてページに記載されている保証はなく「読みに相当する文字列」に過ぎない。つまり、読みではないのに誤ってヒットする可能性がある。これを防ぐヒューリスティックスとして、括弧で囲まれた「読みに相当する文字列」を使う方法が考えられる（図 4の例[10]）。

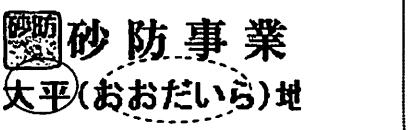

図 4 括弧の利用

ひらがな表記に関して、括弧の影響を調べたが、性能は大きく劣化した（表 6）。表中のヒット率は訓練データが少なくても 1 件あるものの割合である。括弧により精度は向上しても、逆に再現率が下がり、単語と読みを含むページが取得できかったと考えられる。

表 6 括弧の影響

読み表記	M=2	M=5	M=10	ヒット率
ひらがな	89.9	90.3	92.9	100.0
括弧付	86.9	85.1	82.8	92.3

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。頭字語は便利で広く使われ、自由に生み出される。従って、頭字語と定義の対応は曖昧になりやすい。

表 7 頭字語 ACL の出現例

定義	出現例
Anterior Cruciate Ligament (膝の怪我)	She ended up with a torn ACL, MCL and did some other damage to her knee.[11]
Access Control List (セキュリティの用語)	Calculating a user's effective permissions requires more than simply looking up that user's name in the ACL. [12]
Association for Computational Linguistics (学会名)	It will published in the upcoming leading ACL conference. [13]

例えば，表 7に例示したように、頭字語 ACL には少なくても 3 つの異なる定義がある。

8 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。

9 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。

10 http://www.cbr.mlit.go.jp/numazu/sand/sand20.html

11 http://aphotofreak.blogspot.com/2006/01/ill-give-you-everything-i-have-good html)

12 http //www.mcsa-exam.com/2006/02/02/effective-permissions.html

頭字語は本来長い定義を省略するためにあり、定義と文書内で共起しないことが多い。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。

一方、頭字語は、ある程度以上広く用いられる場合、その定義を公に明確にする必要がある。したがって、逆に、頭字語は定義とどこかで共起する(図 5の例[14])。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。

図 5 頭字語と定義のページ内での共起

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 [4]などの先行研究があり高い性能が実現されている。また、この機能を提供するサイトがいくつもある。

3.1. 典型的な頭字語のリスト

まず、頭字語のリスを Wikipedia から取得し、英字大文字以外を含む、または、3 文字より短いものを除いた。続いて、頭字語の定義をより取得し[15]、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」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 8 曖昧度が 2 の場合の手法 2 の精度

曖昧度	M=2	M=5	M=10
2	88.7 %	90.1 %	92.4 %

曖昧度が 2 の場合（表 8）90%前後の精度が得られた。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 9 曖昧度が 5 の場合の手法 2 の精度

曖昧度	M=2	M=5	M=10
5	78.6 %	82.6 %	86.0 %

曖昧度が 5 の場合（表 9）は曖昧度が 2 の場合より

[13] http://pahendra.blogspot.com/2005/06/june-14th.html
[14] http://www.aclweb.org/
[15] http://www.acronymsearch.com/

は性能が下がるがおよそ 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。

頭字語の曖昧性解消でも、読みの曖昧性解消と同程度の高精度が確認できたといえる。

4. 議論

4.1. データのバイアス

2 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(図 6)。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。

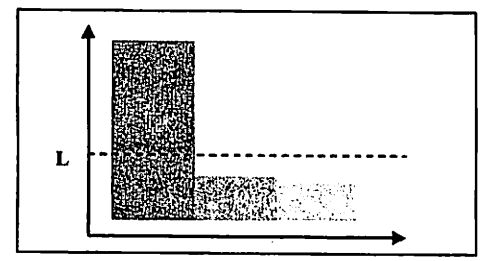

図 6 上限値 L によるデータ・バイアス

手法 2 は、訓練データ数の制限が決定木をある場合、誤動作しうる。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 10) 。一方、訓練データの分布は先に述べたように L で抑えられるため平坦である。従って、手法 2 が手法 1 に劣る結果になる。

表 10 頭字語 ISP の定義の分布

定義	ヒット数
Internet Service Provider	3,590,000
International Standardized Profile	776
Integrated Support Plan	474
Interactive String Processor	287
Integrated System Peripheral control	266

これと違って、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 11)。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。この場合は決定木がよく働き、手法 2 が手法 1 に勝る。

表 11 頭字語 CEC の定義の分布

定義	ヒット数
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

図 7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考えられる。

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。

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問題と頭字語と定義の対応付けの問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考えている。

文 献

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

[2] 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.

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

[4] David Nadeau and Peter D. Turney, 2005. "A super-vised learning approach to acronym identification," 18th Canadian Conference on Artificial Intelligence, LNAI3501.

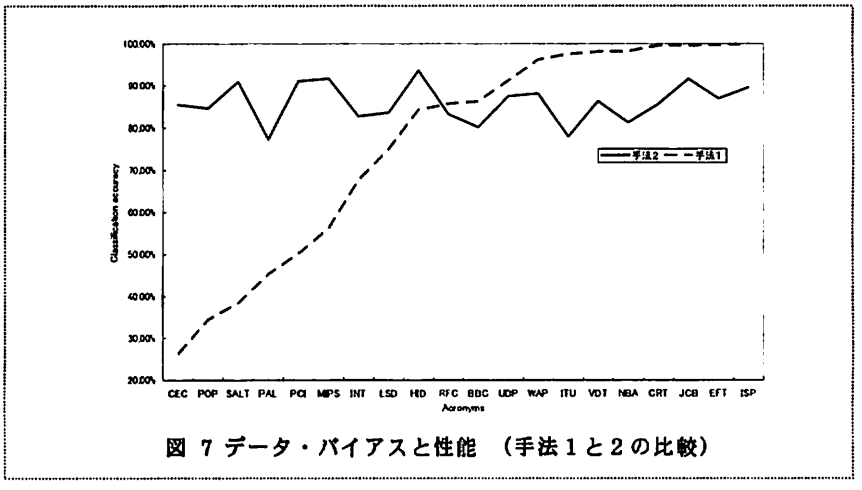

図 7 データ・バイアスと性能 （手法1と2の比較）