

読み曖昧性解消のためのデータセット構築手法

Dataset Construction Method for Word Reading Disambiguation

西山 浩気^{*1}
Koki Nishiyama

山本 和英^{*1}
Kazuhide Yamamoto

中嶋 秀治^{*2}
Hideharu Nakajima

^{*1} 長岡技術科学大学
Nagaoka University of Technology

^{*2} NTTメディアインテリジェンス研究所
NTT Media Intelligence Laboratories

We propose a data construction method for word reading disambiguation. The method gives unique reading word to each reading of reading ambiguous word, collects sentences including the unique word, replaces the unique word in sentences to the original ambiguous word and tags readings of reading ambiguous words to the reading corresponding to the unique word. Through experiments, we confirmed the method collects data numerically balanced between readings.

1. はじめに

音声合成において、文章の区切りや読み・アクセントを決める言語解析は重要な技術である。単語の接続情報に基づいた形態素解析処理の結果によって、あるいは、単語の接続に関する統計情報に基づいて、単語に正しい読み・アクセントを付与することが概ね可能である[山森 99, 長野 06, 森 10 など]。しかし、それらの単語の接続情報は当該の単語の前後の数単語の情報を利用する程度に留まっており、誤る場合がある。その誤りの中でも、音声合成の目的において目立つ誤りとして、同形異音異義語間での読み付与の誤りがある。例えば、“方”という単語には、選択肢の一つや方角を意味する“ほう”という読みと人を意味する“カタ”という読みの曖昧性がある。このような読み曖昧性の解消が必要である。

このような曖昧性解消の研究には、できるだけ大規模なコーパスが必要である。一般に入手可能な現代日本語書き言葉均衡コーパス(BCCWJ)では、人手によって約 2,000 文への読み付与が行われているが、我々が必要とする同形異音異義語間での読み多義解消用のデータとして用いるには規模が十分ではない。意味の多義性解消における誤りの約 7 割が訓練データ不足による誤りであるとの報告[新納 15]から考えると、既存のコーパスのみでは十分な精度で多義性解消することは期待できず、読み・アクセント情報が付与されたデータセットを構築する必要がある。

多義語の語義毎の例文収集を目的として、言い換えを利用したデータ拡充手法が報告されている[戸田 16]。この手法に倣って、本研究では読み曖昧性解消用のデータセット構築を行なう。同形異音異義語は同じ字面でありながら、意味によって発音が異なる単語である(以後、多音語と呼ぶことにする)。多音語の読みごとのデータセットを構築するため、読みを 1 つしか持たない同義語(以後、一音語と呼ぶことにする)に置き換え、例文を収集することで、読みの定まった例文収集を行なう。同義語は文脈が似ていると期待できるから、一音語へ置き換え検索収集する。本手法では読みと語義ごとにデータの収集が可能であるため、多音語を検索キーとして例文収集を行い、後から読みを対応付ける手法に比べ、それぞれの読みに対応する例文を効率よく収集する手法であることを示す。

2. 提案手法

提案手法の概略を図 1 に示す。初めに、語義と読みの両方が異なる異音異義語、すなわち、多音語をそれぞれの同義語で読みが一通りである一音語へ置換する。置換した一音語を検索語として、クロールや大規模コーパスから、その検索語が含まれる文を収集する。この検索語を元の多音語に直し、多音語の複数の読みの中で前記の一音語に対応する読みを収集した文内での多音語の読みとして対応付ける。最後に、収集した文に付与された読みに対して人手による評価を行う。

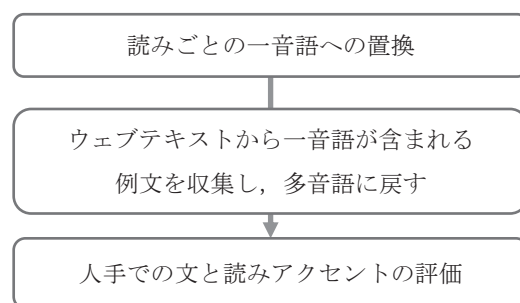


図 1 提案手法

2.1 読みごとの一音語への変換

多音語を読み毎に一音語へ置き換える。各単語の日本語 WordNet [Isahara 08]のエントリから各語義と等価な単語であり、かつ、読みの曖昧性の無い語を選定する。WordNet に検索語が存在しない語、あるいは、検索結果に適切な単語や複合語がない場合は人手で選定する。例として、“今日(コンニチ/キョー)”の読み毎に一音語へ選定する例を示す。初めに、日本語 WordNet で“今日”を検索すると以下のような結果が得られる。

• WordNet における“今日”の検索結果の例

today : { 方今, 輒近, 頃来, 近頃, 近ごろ, 今, 今日, 今時, 今日このごろ, 最近, 当節, 現代, 現在, 今節, 今どき, 今日此の頃, 今日此頃 }

この検索結果から語義をひとつしか持たない単語をひとつ選択する。表 1 に“今日”に対して選定した一音語を示す。

表 1 一音語への変換の例「今日」

読み	一音語
コンニチ	最近
キョー	本日

2.2 例文収集

一音語を検索語として文の検索を行う。そのあと、収集した文に含まれる一音語を元の多音語の表記に置きなおすことで、読みや意味毎に対応づいた例文とする。

「今日」を例として読みごとの例文収集を説明する。2.1 節において「今日」の読み「コンニチ」「キョー」に対して一音語「最近」「本日」への置き換えをそれぞれ行った。置き換えた一音語を検索語として Web からクローリングを行なうなどして文を収集できる。次に、収集した例文の一音語を元の多音語に置き換えることで、読みと対応づいた例文を収集する。一音語で Web テキストを検索した結果と、その置換後の文について例を示す。

- “最近”が含まれる例文(コンニチに対応)
検索結果: 最近は寒の戻りの特異日ですが、
春の陽光たっぷりの一日になりそうです・・・
置換後 : 今日は寒の戻りの特異日ですが、
春の陽光たっぷりの一日になりそうです・・・
- “本日”が含まれる例文(キョウに対応)
検索結果: 私本日気になったので本屋を覗いて見ても
それらしい本はありませんでした。
置換後: 私今日気になったので本屋を覗いて見ても
それらしい本はありませんでした。

2.3 文と読みの妥当性に関する人手評価

前ステップで収集した文の一音語を多音語に置き換えた。このことによって文として成り立たなくなっているか、読みが正しいか否かを人手で評価する。評価は以下の 2 段階に分かれる。

- 評価 1. 文中の多音語の用法が正しいか否かの評価
- 評価 2. 用法が正しい場合は、文中の多音語の読みが正しいか否かを評価し、読みが誤っている場合は正しい読みを付与する。

多音語の複数の読みのうち取りうる読みをあらかじめ収集しておく、作業員へ提示する。例えば、“最近”を検索語として例文を収集した。“最近”を元の“今日(コンニチ)”に置き換えた文を挙げて以下に示す。

- 用例が“正しい”，読みが「コンニチ」の例
ちなみに今日の研究結果からは 3000 年前には既に弥生時代に入っていた、て結果だったよね。

上記の例においては文中での“今日”の用法は“正しい”，読みは“コンニチ”とラベル付けする。

3. 多音語の置き換えによるデータセットの構築実験

提案手法を用いて同形異音多音語の一音語への置換を行い、語義ごとの文収集を行ない、人手による評価を行った。本手法を用いることで読みごとにほぼ均一に例文を収集することができ、多音語の訓練データ不足を補う手法として有効であることを示す。

3.1 実験設定

今回の実験では、形態素辞書 UniDic 中の多音語を対象とした。同一の表記と品詞を持つ形態素のうち、読みの異なる形態素 522 組 2,128 語を抽出した。これらの単語を以下のように分類し、分類名 1) の異義語に含まれる 43 組 88 語の多音語を対象とした。

表 2 辞書から抽出した語の分類

分類名	事例
1) 異義語	今日:キョー/コンニチ
2) 清濁の違い	会社:カイシャ/ガイシャ
3) 音訓の違い	縁:エン/フチ
4) 同義語	下腹:シタハラ/カフク
5) 音の異なり	行く:イク/ユク

読みごとに置換する一音語選定には前記の日本語 WordNet を用いる。一音語を含む例文の収集にはクローリングも可能であるが、今回は、前もって収集されたウェブテキスト¹⁾内での検索によって例文の収集を行なった。このコーパスには IPAdic-2.7.0 の見出し語を検索語として Yahoo!WebAPI による検索結果に含まれるウェブページをクローリングした 396GB (数十億文)に相当するテキストが含まれる。収集した文の中から最大 100 文をその読みのデータセットとした。1 文に対して 2 名の作業員で評価を行い、前記の 2 つの観点からの評価を行なった。1 つの文に対して異なる 2 名の作業員を割り当て、作業員には互いに相談することなく作業を行わせた。

3.2 収集例文数と作業員の信頼性評価

一音語を多音語に置き換えたことで日本語テキストとして意味の通らない文が存在するかもしれない。これは、多義性解消機能自体で評価もできるが、集めた文をより詳しく評価するため、人手評価における「用法は正しいか否か」のラベルを通して妥当な置き換えが行われたか否かを判別した。2 名の作業員がいずれも正しいと判断した文であり、かつ同一の読みを付与した文であれば訓練データとして利用可能であると考えた。Web コーパスから収集した例文数のうち、2 名の作業員がいずれも正しいと判断した例文数(評価 1)と、評価 1 かつ同一の読みに編集した例文数(評価 1&2)を集計し、表 3 に示す。両作業員が同一の読み仮名を選択した場合に限り、正解とした。

表 3 総例文と収集したデータの総数

例文数	評価 1	評価 1&2
11, 653	5, 790	5, 389

表 3 より例文数と評価 1&2 を比較すると、収集された例文は Web コーパスから収集した例文数の 4 割である。また、2 名の作業員が正しいと判断した文に関してはほぼ一致した読み情報が付与されていることがわかる。しかしながら、作業員の付与ラベルが正確であるとは限らないため、次にラベルの信頼性について述べる。作業員が“正しい”とラベル付けた文に対して、人手で文中の多音語の用法が正しいか否かを再度評価し、作業員との一致率を確認した。“作業員 2 名のうち、いずれかの作業員 1 名が正しいと判断した文”と“作業員 2 名が正しいと判断した文”のそれぞれをランダムに 100 文抽出し、人手で再度評価したラベルと比較した一致率を表 4 に示す。

表 4 作業員が”正しい”と判断した文と
人手で再評価した結果との一致率

作業員 1 名	作業員 2 名
42%	98%

¹ テキストアーカイブ・日本語ウェブコーパス2010,
<http://syata.jp/corpus/nwc2010/texts/>

表 6 平均採用率 10%以上の多音語における語義ごとに獲得した文の例

異義語	語義/一音語	例文
今日	キョー/本日	8月に入った途端、暑い夏が続いていますが、暦では秋・今日は立秋です
	コンニチ/最近	ちなみに今日の研究結果からは 3000 年前には既に弥生時代に入っていた、て結果だったよね。
寒気	サムケ/悪寒	思っていたとしても、一歩間違えると私と姑の関係のようになってしまう寒気がします
	カンキ/冷たい風	午前には寒気が強くなっていたんですが午後には止み、JSB 決勝の頃はほぼカンペキなドライコンディション

作業員 1 名のみでは再評価したラベルとの一致率は低く、十分な信頼性は得られないが、2 名の作業員がいずれも正しいとする例文での一致率は非常に高く、十分な信頼性が得られていることがわかる。従って、収集した 11,653 文のうち評価 1,2 を満たす 5,389 文、割合で示すと 46.2% の文がデータセットとして利用可能であると言える。

3.3 単語ごとの例文収集効率について

前項では、データセットとして利用可能な例文数を示した一方で、収集した例文数に比べ利用可能な例文の割合が 4 割程度と効率の悪い収集方法であるように見える。そのため本項では多音語ごとに評価 1&2 を満たす例文数を評価することで、例文の収集効率について言及する。読みごとに Web テキストから収集した例文数と作業 1&2 の比率で採用率を計算し、その平均した結果を表 5 に表す。この結果からは、データセット全体での採用率は 6 割程度であるように見えるが、これは採用率が 10% 未満の著しく採用率の低い一音語が、対象とする多音語 82 語のうち 15 語存在しているためである。この採用率 10% 未満の語を除いた例文の採用数を同様に表 5 に示す。この結果から本稿で対象とした 82 語のうち、15 語を除いた 67 語に対しては平均 8 割程度の採用率で訓練データを収集することができる。

表 5 データの語義ごとの収集率

対象	数量	平均採用率[%]
全語義	82 語	60.4
採用率 10%以上の語	67 語	80.4

次に、読みごとの例文収集効率について、単純に多音語を検索キーと比較した場合と比較する。“今日”を検索キーとして Web コーパス内を検索し、検索結果からランダムに 200 文を抽出、人手で“今日”の読み仮名を評価した。その結果、“キョー”が 197 件、“コンニチ”が 3 件となり、使用頻度に非常に大きな偏りがあることがわかった。その一方で、本手法では両方の読みを 200 文ずつ集めることができており、読み毎に例文を収集する手法として効率よく収集ができていたといえる。

3.4 採用率 10%以上の語の誤り分析

採用率の高い語とその例文について述べる。Web コーパスからの例文収集数が 100 文、かつ作業員 2 名にその全ての例文が正しいと判断された「今日(キョー/コンニチ)」「寒気(サムケ/カンキ)」について、各語義とその例文について表 6 に示す。「今日」について収集した例文では、その日一日を表す語義である(キョー)とその日だけでなくそれまでの日も表す語義(コンニチ)の例文が収集できていることがわかる。「寒気」でも同様に不快な寒気を表す語義と単に冷たい風を表す語義の例文を収集することができた。これらの語においては単語を置き換えたことにより、例文の内容が変わってしまうことは無く、適切な一音語への置き換えができていたといえる。一方で誤りとなる場合について

ランダムに 100 文を抽出し、以下の表 7 のように 3 種類の誤りに分類した。

表 7 収集率 10%以上の語の誤りの分類と割合

誤り要因	該当語数
1) 一音語ではない語への置換	55 語
2) 複合名詞の置換	43 語
3) その他	2 語

(1) 一音語ではない語への変換

[多音語“避け(ヨケ)”を“よけ”に置き換えた誤り]

原文：定点観測が 5/28 以降更新されていませんが、現在の状況をご存知の方、よければ情報を教えてください

置換後：定点観測が 5/28 以降更新されていませんが、現在の状況をご存知の方、避ければ情報を教えてください

“よけ”には“良け”、“除け”など曖昧性があるため、“除け”など曖昧性を回避する表現に置換するなどして、より適切な一音語へ置き換えることで改善できる。

(2) 複合名詞の変換

[多音語“頭(カシラ)”を“上部”に置き換えた誤り]

原文：赤石トンネルを上村側に抜けると、坑口上部に「赤石隧道」の表札が掲げられている

置換後：赤石トンネルを上村側に抜けると、坑口頭に「赤石隧道」の表札が掲げられている

複合名詞の多くはその組み合わせが異なると違和感のある文になりやすいのではないかと考えられる。複合名詞になる例文はなるべく収集しないように改良する必要がある。

(3) その他

[多音語“何人(ナニジン)”を“どこの国の人”に置き換えた例]

原文：「私もこの人(資料請求した)どこの国の人と聞いたことがある」と話していたことである

置換後：「私もこの人(資料請求した)何人ナニジンと聞いたことがある」と話していたことである

プログラムの誤り(多音語の直後に読みが挿入される)、作業員への指導の誤り、数は少ないので無視できる。

3.5 採用率 10%未満の語の誤り分析

採用率が 10%未満の 15 語についても誤り分析を行った。誤りを 3 つに分類し、読みごとに該当する誤りを分類した。誤りの要因とその該当語数について表 8 に示す。

表 8 収集率 10%未満の語の誤りの分類と割合

誤り要因	該当語数
1) 不適切な一音語への置換	10 語
2) 一音語ではない語への置換	3 語
3) その他	2 語

次に、誤りの要因の例とその誤りについての今後の対策について挙げる。

(1) 不適切な一音語への変換

[多音語“面(ツラ)”を“顔面”に置き換えた誤り]

原文：ただ普通に勝負したら人間が勝つのは目に見えるので、顔面に洗濯ばさみというスタイルで挑みます。

置換後：ただ普通に勝負したら人間が勝つのは目に見えるので、面に洗濯ばさみというスタイルで挑みます。

このカテゴリは一音語を多音語の置き換え語として選択したが、誤りとなった語義が含まれる。例においては“面”は不適切な置換であるが、“相手の面に蹴りを入れる”など俗語的な使われ方をする場合には適切な場合がある。従って、多音語の出現する文脈を考慮した置換が必要であるが、多音語と同一の文脈で出現する適切な一音語が見つからない場合には、収集効率は低下することを踏まえた上で、必要数より多く見積もって例文を収集しておく必要がある。

(2) 一音語ではない語への置換

[多音語“人気(ヒトケ)”を“人の気配”に置き換えた誤り]

原文：蠍座田中支配人の気配りで、お客様がより時間に制約されないようにと、午前と午後の 1 日 2 回のプログラムを組んでくれました。

置換後：蠍座田中支配人人気で、お客様がより時間に制約されないようにと、午前と午後の 1 日 2 回のプログラムを組んでくれました。

“気配”には“ケハイ”と“キクバリ”の読み曖昧性があるため、表記の一致をとる場合に“キクバリ”と読む文を誤って収集してしまっている。“気配り”と書く場合は“キクバリ”という読みを予め原文に割り当てることでこの曖昧性を回避し、文の収集ができる。

(3) その他

3.4 項同様の誤り、無視できる範囲で事例が少ないため記述を割愛した。

以上を踏まえて、適切な一音語への置換を行うためには、異義語の語義に近い単語かつ多義を持たない場合であれば有効であると考ええる。

面(ツラ)のような同義語を人手で考えることが難しい単語に対して、類似語検索等の補助を行うことで、より適切な一音語への変換が可能になると考えられる。また、置き換え先の語は一定以上の出現頻度が必要である。“中日(ナカビ)”を“中間の日”に置き換えを行ったが、Web コーパスにおいての頻度が非常に低く 3 文程度のデータ収集に留まった。実験に用いた Web コーパスの量は十分であることから、クロール量を増やす以前にある程度の頻度を持つ単語への置き換えが重要である。

最後に例文の収集にかかる時間について考察する。1 人当たり約 4,000 文の評価に費やした時間は、納品までに約 2 日間である。例文が正しいか否かを判断するために、例文を全て読む必要があるが、例文が正しいか否かを判断した後の読みの読み判別は日本語母語話者であれば比較的迅速に作業を行うことができていた。新たな例文を収集するために本手法では既に Web テキストとして存在している文の一部を改変する形で拡充しているが、もうひとつの手法としてクラウドソーシングを用いた人手での例文生成が考えられる。しかし、本稿の手法と比較すると 1 文を生成するために必要なコストが高く、アノテート済みのデータを大量に作ることは向かないと考える。

作業が簡単であるため、1 人あたりの金額を抑えて募集することができた。更に多くのデータを収集する際に低コストで大量に収集するための手法として有効である。

4. おわりに

本稿では、既存のコーパスでは不足している同形異音異義語(本稿では多音語と定義)の曖昧性解消のためのデータセットの構築を行った。多音語の語義ごとに語義を一つしか持たない語へ人手で置換を行い、置換した語を検索キーとして Web コーパスから置換した語を含む例文を収集した。40 組 82 語の異義語が含まれる計 11,653 文の例文に対して、1 文あたり 2 名の作業者が人手でのアノテート作業を行い、例文が利用可能か否かと読みアクセント情報を付与した。その結果、訓練データとして利用可能であると判断された 5,790 のデータセットを構築した。本手法では異義語を一音語へ置換することで、“今日(コンニチ)”のような語義によって出現頻度に大きな偏りがある語であっても、例文を十分数収集できることを示した。また、15 語の著しく例文の収集率の低い異義語を除いた 67 語の異義語に対して、Web コーパスから収集した例文のうち約 80.4%を訓練データとして利用可能であることを示し、語義ごとの例文収集手法として有効な手法であることを示した。

参考文献

- [山森 99] 山森和彦: 未来ねっと技術シリーズ 4 メディア処理技術。オーム社, 1999.
- [長野 06] 長野徹, 森信介, 西村雅史: N-gram モデルを用いた音声合成のための読みおよびアクセントの同時推定. 情報処理学会論文誌, Vol. 47, No. 6, pp. 1793-1801, 2006.
- [森 10] 森信介, Graham Neubig: 仮名漢字変換ログの活用による言語処理精度の自動向上. 言語処理学会第 16 回年次大会, 2010.
- [新納 15] 新納浩幸, 白井清昭, 村田真樹, 福本文代, 藤田早苗, 佐々木稔, 古宮嘉那子, 乾孝司: クラスタリングを利用した語義曖昧性解消の誤り原因のタイプ分け, 自然言語処理, Vol. 22, No. 5, pp. 319-362, 2015.
- [Okumura 11] Okumura, M., Shirai, K., Komiya, K., and Yokono, H.: “On SemEval-2010 Japanese WSD Task.”, 自然言語処理, Vol. 18, No. 3, pp. 293-307, 2011.
- [西尾 94] 西尾実, 岩淵悦太郎, 水谷静夫, 岩波国語辞典第 5 版, 岩波書店, 1994.
- [戸田 16] 戸田勇馬, 村田真樹, 馬青: 言い換えと機械学習を用いた日本語単語の多義性解消, 言語処理学会第 22 回年次大会, Vol. 18, No. 3, pp. 172-175, 2016.
- [Isahara 08] Hitoshi Isahara, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama and Kyoko Kanzaki: Development of Japanese WordNet, LREC-2008, Marrakech, 2008.