決定リストによる同形異音語の読み分け

梅村 祥之　　清水 司

豊田中央研究所　機械認識研究室

1 はじめに

カーナビゲーションを初めとする種々の情報機器が自動車に搭載され，様々な情報通信サービスが広がりを見せている．その中には，交通情報，電子メール，新聞記事等の情報提供が含まれる．情報提示方法として，ドライバへの提示には，音声が望ましいが，テキスト音声変換に未解決の課題がある．

通常，漢字仮名混じりの文章を音読する技術は，1）文章の読みを求め　2）韻律を求め　3）読みと韻律情報から，音声波形を生成する　というステップからなる．この中の，読みを求める問題には，例えば「米（こめ／べい）」を読み分けるといった同形異音語の読み分けがある．この問題は，単語の多義性解消の研究と関連が深く，その流れに沿った研究として，Yarowskyはアクセント記号が欠落しているフランス語文から，アクセントを推定する問題に決定リストを用いている[1]．Liは決定リストにMDL原理に基づいた証拠の信頼度に関する指標を付加し，日本語の同形異音語の読み分けに応用して，EDRコーパスの中からランダムに選んだ50語と難易度の高いものを選んだ50語の計100語について実験している[2]．

本稿では，EDRコーパスを用いた決定リストによる読み分けにおいて，学習データを擬似的に増加させることによって，読み分け性能の向上を図る．

2 対象とする語の決定

日本語には数千語の同形異音語が存在する．例えば，形態素解析システムChaSen 1.0の形態素辞書から，同形異音語を機械的に抽出すると2,626語得られる．しかし，その中には，使用頻度のきわめて少ない特殊な語もかなり含まれている．そこで，EDRコーパスから，頻度が少ない方の読みが10回以上出現する同形異音語を抽出すると，362語得られる．

この中から，次の判断基準により，検討対象とする語を削減する．

1. EDRコーパスの読み付与の問題で同形異音語として抽出されたもの（例：「9月」の読みが「9がつ／くがつ」）
2. 連濁によって複数の読みが生じたものは，連濁規則を適用して，読みを求めればよいため[5]，省く
3. EDRコーパスの単語区切りでは同形異音語となるが，ChaSenでは同形異音語とならないもの（例：EDRコーパスでは「運ぶ」が「運（はこ）／ぶ」，ChaSenでは「運ぶ」の1語）
4. 微妙なニュアンスの違いはあるが，どちらの読みでもよいと思われるもの．　例：「この間（このかん／このあいだ）」

その結果，204語に絞られる．

次に，簡単なルールでほぼ読み分けできるものに関しては，決定リストを使うメリットが少ないとの観点から，対象としないことにする．簡単なルールとして，次のものが有効である．

【ルール】　・単独で現れれば訓読み
　　・接尾語か複合語の要素なら音読み

このルールを「塩（しお／えん）」に適用すると95%の正解が得られる．前記204語のうちのほぼ半数はこのルールが有効に働き，残る104語を決定リストの対象とする．

3 関連表現による学習データ増加法

3.1 決定リスト適用上の問題

例えば「米（べい／こめ）」の読み分けにおいて，「こめ」の共起語として容易に連想できる「農家」「水田」「自給率」「秋田」が学習データ中に何文含まれているかを調べると，EDRコーパスで形態素「米(こめ)」を含む全ての文113文に，「農家」3文，「水田」2文，「自給率」0文，「秋田」0文である．従って，学習データの量を増やせば，こういった共起語を獲得して，正解率が向上する可能性がある．

実際に，「米（べい／こめ）」の読み分けに関して，EDRコーパス中に含まれるデータを10段階に減少させたときの決定リストによる正解率の推移を調べ

ると，図1のようになり，学習データの増加が正解率向上につながるであろう事が予想される．

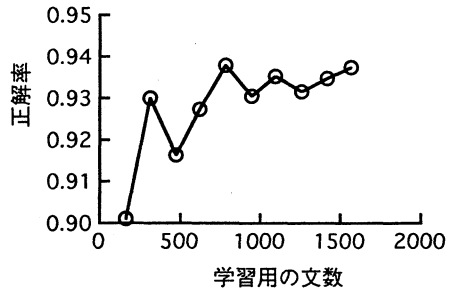

図1　学習データ量による正解率の変化

3.2　学習データ増加法

関連表現を基に，文を機械的に抽出し，それを本来の語に置き換えて学習データとする．例えば「米(こめ)」の読み分けのための学習データを獲得するために，「白米」「ヤミ米」「小麦粉」などの関連表現を使い，生コーパスから関連表現を持つ文を機械的に抽出する．例えば，「ヤミ米」による抽出で，

国の減反政策を無視している秋田県・大潟村の一部農民と日本消費者連盟などが提携し，今秋から過剰作付け分のヤミ米を産地直送することになった．

という文が得られる．この文の「ヤミ米」を本来の「米」に置き換える．置き換えられた文は，

国の減反政策を無視している秋田県・大潟村の一部農民と日本消費者連盟などが提携し，今秋から過剰作付け分の米を産地直送することになった．

となり，「米（こめ)」の学習用の文として用いても問題なさそうな文になっている．このような関連表現によって得られた文を学習データに追加してから，従来の方法で，決定リストを作成すると，例えば前述の「秋田」が共起語に採用される可能性がある．一方，「米（べい)」の関連表現としては，「米国」「アメリカ」を利用すればよい．

3.3　関連表現の決定のガイドライン

関連表現を見つけるには，基本的には，分類語彙表[2]や角川類語新辞典[3]などのシソーラス辞書で類語を調べる．例えば「米」の場合，分類語彙表の「1.4320　米・糠・小麦粉など」には「米（こめ)」を含む29語が掲載されている．これらを関連表現として用いるが，注意すべき点がある．それは，多義性のある語を省くことである．分類語彙表1.4320の第3項目に掲載されている「ふすま」は，「小麦を粉にするときできる皮のくず[4]」の意味の他に，扉の意味があるため，これを関連表現に採用すると，誤って扉に関連する文が抽出される可能性がある．

4　実験

4.1　決定リストのパラメータ

李らの方法[2]をベースに，処理のパラメータを以下のように定めた．

形態素解析：　ChaSen Ver. 1.0を使用した．以下に記載の語や品詞は，ChaSenの体系に従う．
EDRコーパスにおける単語の分割とChaSenによる分割に食い違いがあり，例えば「対米政策が...」という文章では，EDRコーパスの場合，「対／米／政策／が...」と分割されるのに対し，ChaSenでは「対米／政策／が...」と分割される．従って，この文は「米」を含む文ではないとした

証拠の種類：　直前の1文字，直後の1文字，直前の品詞，直後の品詞，直前の語，直後の語，ウインドウ内の自立語

ウインドウ幅：対象とする語の，前に10自立語，後ろ10自立語の範囲とする（李らの文献[2]で，前後5，前後10，前後20を評価し，ほとんど差がなかった）

自立語：　自立語は，ChaSenにおける品詞区分のうちの次のものとした．　動詞，普通名詞，サ変名詞，固有名詞，地名，人名，数詞，形式名詞，副詞的名詞，時相名詞，形容詞，様態副詞，名詞形態指示詞，連体詞形態指示詞，副詞形態指示詞，程度副詞，量副詞，頻度副詞，時制相副詞，陳述副詞，評価副詞，発言副詞

複文の処理：　複数の文にまたがる処理は行わない．例えば，対象とする語が文頭にあった場合，直前の文字として前文の文末を使うことはない

4.2　実験の対象語と関連表現

第2章において，決定リストの対象語として絞った104語から，関連表現が有効と思われる次の10語を選ぶ．　－　米（こめ／べい)，仏（ふつ／ほとけ)，表（ひょう／おもて)，額（がく／ひたい)，訳（やく／わけ)，今日（きょう／こんにち)，種（しゅ／たね)，床（とこ／ゆか)，金（きん／かね)，縁（え

ん／ふち）－　これらはいずれも単語の多義性が読みの違いに現れているものである．それに対して，例えば「方（ほう／かた）」の場合には，単語の意味の違いが読みに対応するというよりも，慣用的な表現毎に，読みが異なると考えられるため，関連表現による方法は不向きである．

以上の10語に関して，前述の関連表現決定のガイドラインに沿って，関連表現を決定した．その内容を表1に示す．

4.3　実験結果

対象とする10語に関する，もともとEDRコーパ

表 1　読み分け対象の10語に関する関連表現

表記	読み	関連表現
米	こめ	ポップコーン，コーン，ピーナッツ，南京豆，炒り豆，豆，押麦，精麦，麦，麹，小米，屑米，上米，地米，産米，古米，新穀，新米，半搗き米，神米，粳，糯，糯米，米麦，穀，穀類，穀物，五穀，米穀，雑穀，飯米，もち米，玄米，白米，半搗米，早場米，外米，ヤミ米，米粒，粉米，糠，米糠，ふすま，押し麦，ひきわり麦，麦粉，とり粉，ふくらし粉，きな粉，かたくり粉，メリケン粉，小麦粉
	べい	（関連表現による追加せず）
仏	ほとけ	神仏，神，神明，氏神，明神，鬼神，女神，天神，武神，八幡，福の神，七福神，布袋，大黒，恵比寿，疫病神，死神，山の神，水神，荒神，大仏，如来，釈迦，釈尊，阿弥陀，薬師，菩薩，観音，地蔵，閻魔，仁王，明王，天主，救世主，救い主，メシア，キリスト，造物主
	ふつ	フランス
表	おもて	裏，表側，表裏，裏表
	ひょう	図表
額	ひたい	頬，ほっぺた，ほおげた，ひたい，おでこ，眉間，脳天，前額，富士額，額際，髪際，眉宇，頭かみ，こめかみ，頬っぺた，豊頬，顎，あご，上顎，うわあご，下顎，したあご，頤，おとがい，目鼻，頭部，後頭部，前頭，後頭，脳天，首筋，眼球，瞳孔，耳たぶ，顔面，胴体，胴体，背筋，肩先，胸部，胸郭，腹部，下腹部，四肢，肢体，二の腕，手首，ふくらはぎ，太もも，足首，かかと，足の裏，ひじ，親指，人差指，食指，中指，薬指，小指
	がく	金額，全額，半額，同額，多額，定額，倍額
訳	やく	直訳，逐語訳，意訳，全訳，抄訳，新訳，旧訳，初訳，改訳，名訳，定訳，適訳，誤訳，和訳，邦訳，英訳，監訳　（品詞を名詞に限定）
	わけ	理由
今日	きょう	昨日，明日，あす，きのう，明後日，あさって
	こんにち	「最近の」の文を抽出し「今日の」に置き換える
種	たね	種子
	しゅ	種類
床	とこ	病床 寝台 寝床
	ゆか	天井
金	かね	¦金銭，現金，大金，小遣い，義捐金，寄付金，賜金，賞金，償金，慰謝料，手切れ金，持参金，礼金¦　＋　¦は，を，の，に，が，も¦　（名詞＋助詞の形で関連表現とした．主旨は，もし，助詞を伴わないと「現金強盗」「現金収入」など「金強盗」「金収入」に置き換えることができないものが抽出されてしまうため，それを避けるためである）
	きん	金銀，純金，こがね，砂金，金塊，洋銀，プラチナ，白金，しろがね，金無垢
縁	ふち	端，右端，左端，両端，突端，上端，下端
	えん	由縁，機縁，ゆかり，縁起，きずな，奇縁，宿縁，腐れ縁，旧縁，悪縁，良縁

表 1　決定リストと関連表現による読み分けの誤り率

語	読みとコーパス中の頻度	関連表現の文数	決定リストによる誤り率[%]	関連表現による誤り率[%]	誤りの減少率[%]
米	こめ 113, べい 1464	こめ 152	6.5	4.2	35.0
仏	ふつ 166, ほとけ 12	ふつ 528, ほとけ 168	9.0	6.7	25.0
表	ひょう 240, おもて 68	ひょう 404, おもて 182	19.2	10.6	21.4
額	がく 307, ひたい 28	がく 263, ひたい 207	13.0	11.9	11.4
訳	わけ 36, やく 11	わけ 53, やく 15	25.5	23.4	8.3
今日	こんにち 295, きょう 212	こんにち 371, きょう 190	18.7	17.8	5.3
種	しゅ 430, たね 59	しゅ 181, たね 29	18.8	18.0	4.3
床	ゆか 123, とこ 20	ゆか 90, とこ 25	13.4	13.4	0.0
金	かね 350, きん 130	かね 148, きん 41	13.5	13.8	-1.5
縁	えん 92, ふち 10	えん 47, ふち 37	11.8	18.6	-58.3
全体	誤りの減少 16.8% (「縁」を除くと 19.7%)				

スに現れる文のみを使っての決定リストによる読み分けの正解率と，前述の関連表現によって，学習データを増加させて決定リストによる読み分けを行った場合の正解率を計算する．正解率算出にあたって 10 fold cross validation を用いた．結果を表 1に示す．10語全体で，関連表現の手法により，読みの誤りが17%減少している．

「額」に関しては，「額に汗して...」といった「汗」を伴った慣用表現の文が，全体の26%を占め，「汗」を含む文全ての読みは「ひたい」であった．慣用表現は，関連表現によって置き換えができないため，この手法が逆効果になる．「汗」が「額（ひたい)」と強く共起し，元来の決定リストが有効に働く．そこで今回，前処理によって，「汗」を含む文を取り除き，残りの難易度の高い文に関してのみ扱った．

5　まとめ

同形異音語の読み分けアルゴリズムを検討した．決定リストを用いた読み分けにおいて，正解率向上のために，学習データの増加が有効と考え，それを機械処理によって獲得する方法として，関連表現を用いる方法を提案した．その有効性を評価するために，同形異音語の中から，読み分けの難易度の高いと考えられる意味の多義性に起因する同形異音語10語について，この方法を適用した．その結果，読み分けの誤りを17%減少させた．今回扱った10語のうち，関連表現に適さない語が1語あり（縁)，それを除いた9語に関しては，読み分け誤りを20%減少させた．

現在，関連表現の決定方法と，本法の適不適の判断には，単語の多義性と慣用句の有無に基づくガイドラインを設けたものの人手が介在しているため，基準の明確化ないし自動化が今後の課題である．

謝辞

日頃より御指導いただく機械認識研究室原田義久室長に深謝します．貴重なご助言をいただいた奈良先端科学技術大学院大学松本裕治教授に深謝します．形態素解析に，松本研究室開発の ChaSen1.0 を使用させていただいた．

参考文献

[1] Yarowsky: Decision lists for lexical ambiguity resolution: Application to Accent Resroration in Spaiish and French, Proceedings of the 32th Anual Meeting of the Association for Computational Linguistics(1994)

[2] 李航，竹内純一：　証拠の強さと信頼度を考慮した日本語同形異音語の読み分け，情報処理学会自然言語処理研究会資料，97-NL-119(1997)

[3] 国立国語研究所編：　フロッピー版　分類語彙表，秀英出版(1994)

[4] 大野晋，浜西正人：　角川類語新辞典 CD-ROM 版，角川書店(1989)

[5] 佐藤大和：　連濁の分析と規則化の検討，日本音響学会講演論文集，1-2-10(1983)