形態素情報共起性による同形異義語認定方式の検討

1B−3

高木 伸一郎　小原 永　松岡 浩司　武石 英二　阿部 久子

ＮＴＴ情報通信網研究所

１．はじめに

各種情報案内や新聞記事などの文字情報を電話やファクシミリなどを通して利用者に配信する情報提供サービスが急増している。特に日本語の文章を合成音声で読み上げ、電話などで聞けるサービスは即時性の高い記事や長時間態勢の情報提供には非常に有効である。ＮＴＴでは、本サービス実現のために合成音声の品質の改善の他に文章を正しく解析し読み振り，韻律情報付与（アクセント／ポーズ等）を高い精度で実現する研究開発を進めている。ここでは、高精度の形態素解析技術と約43万語の日本語辞書を用いて99%以上の単語解析及び読み振り精度を達成している。[1]-[3] しかし、さらに高精度で読み振りを行うには、同形異義語（同形字面で意味属性及び読みが異なる単語のペア：以降、同形語と略する）の認定方式の検討が必要であった。[4]-[5]

本稿では、実際の新聞記事の単語解析や読み振りの不良を分析して単独名詞で出現する同形語の読み振り精度が低いことを確認するとともに、同形語周辺の形態素情報との共起関係を３タイプで定義して各要素の意味属性との連接強度を判定することにより同形語を認定する方式を提案し、実データを用いた机上評価を通して本方式の可能性を示す。

未知語（辞書未登録）関連		同形語関連	解析不良 辞書不良 関連	その他 口語 表現
固有名詞	一般語	８３件		
３１.３%	１５.７%	２３.４%	１８.６%	１１%

※：対象文章：新聞記事４９７５８文字（単語数３０２１６）

図１　形態素解析や読み振りの不良要因（総不良数３５５）

２．単語解析不良における同形語の出現状況

新聞記事約５万文字の形態素解析及び読み振り処理の結果から単語解析及び読み振り不良事例３５５件を収集してその不良要因を分析した結果を図１に示す。全不良事例のうち、「未知語」（辞書未登録に起因する不良）が46%、「解析不良／辞書不良」（辞書や解析規則に起因する不良）が18.6%であり、同形語に関する不良「同形語」は83件、23.4%と高い頻度を示している。さらに同形語認定の不良を形態的に以下に示す７項目に分類し（下線が不良読み）、それぞれ読み振り不良の比率と合わせて概観すると（表１参照）、「単独名詞」（複合語を形成しない同形語）は、数詞や連濁化とともに読み振り精度を低下させる要因であることが判った。

(1) 接辞関連　…複合語内の接辞認定ミスに関する不良
例：光素子（コー／ヒカリ），大まじめ（ダイ／オー），定年時（ドキ／ジ），文化村（ソン／ムラ），年換算（トシ／ネン）
(2) 単独名詞　…複合語を形成しない一般語に関する不良
例：金（キン／カネ）がかからない選挙，研究の足跡（アシアト／ソクセキ）が残る，寒気（カンキ／サムケ）と頭痛を覚えた
(3) 数詞　…助数詞属性の不備に起因する不良
例：一ケタ台（イチ／ヒト），十対一（ジュー／ジュッ）
(4) 連用形動詞／名詞　…動詞の連用形と連用形名詞の認定不良
例：工事に伴い休日が減り，数回にわたり株を（下線が名詞）

The Word Recognition Method for Kanji Homographs by Semantic Collocations

Shinichirou TAKAGI, Hisashi OHARA, Kouji MATSUOKA
Eiji TAKEISHI, Hisako ABE
NTT Network Information Systems Laboratories

(5) 用言関連　…同形動詞の認定ミスに関する不良
例：皇室に対して抱く（ダク／イダク）思い，匿名と断って（タッテ／コトワッテ），軽そうだ（ケイソー／カルソー）
(6) 固有／一般　…固有名詞と一般語の認定ミスに関する不良
例：神田（シンデン／カンダ）の町人は，仏ルノー（ホトケ／フツ）
(7) 連濁化　…連濁読み付与ミスに関する不良
例：二十一日付（ツケ／ヅケ），米価引き下げ路線（ビキ／ヒキ）

表１　同形語認定不良や読み振り不良の比率（認定不良数８３）

形式分類	件数	不良比率	読み振り不良の比率	
接辞関連	３８	４５.８	読み振り正解	３６.８%
単独名詞	１３	１５.７	読み振り不良	９２.３%
数詞	７	８.４		１００%
連用形／名詞	７	８.４		０%
用言	７	８.４		４２.８%
固有／一般	７	８.４		５７.１%
連濁化	４	４.９		１００%
全　体	８３	１００%	４７%	５３%

３．共起情報を用いた同形語認定方式

同形語の周辺の単語を用いて同形語の読み振りを行う考えは、既に示されており、[6] 感覚的にも十分可能性があることは分かる。しかし、単独名詞型同形語において認定のための共起情報の収集方法や判定方法は具体的に言及されていない。本検討では、不良事例の分析から読み振り精度の低下の一要因となっている単独名詞型の同形語を対象に具体事例を通して認定方式の可能性を示した。

３．１認定対象とする単独名詞型同形語

検討単純化のため以下の条件を全て満たすものを検討対象とした。
①形態素情報を利用した読み振りが可能な条件
・読み振りが文脈の影響を受けにくいもの
（例）「今日：コンニチ／キョウ」は文脈に依存することが多い
②読み振りを必要とする条件
・読みの異なる出現分布が比較的均等なもの
（例）足跡：ソクセキ／アシアト，方：ホー／カタ
・読み振り不良が聞き手の理解にとって重大なもの
（例）心中：シンチュー／シンジュー，寒気：カンキ／サムケ

３．２単独名詞型同形語の認定に用いる形態素情報

同形語周辺に位置する以下の３タイプの共起関係を持つ単語に着目して、同形語と共起するこれらの単語を収集する。さらに同形語ごと及びタイプごとに、共起するこれらの単語に必要な意味属性に関する制約（共起情報）を規則化した。

a：修飾／並列要素：同形語の属する名詞句内において「ａのｂ」形式の修飾単語や並列する単語（「の」「と」「や」で接続）
b：述部用言要素　：同形語を格要素とする述部用言や連体修飾する用言（体言止めの名詞，だ文述部の名詞を含む）
c：格要素　：同形語の述部用言が有する他の格要素（述部用言が出現するまでの格助詞を伴う名詞）

例えば、同形語「寒気」の場合、「体に吐き気と寒気を覚える」の例文で、収集／規則化される意味属性は、a＝『健康』（吐き気）b＝『感覚』『学習』（覚える）c＝『身体』（体）となる。ここで、『』内が意味属性である。

3．3共起情報による同形語判定

同形語の判定は、原文上の同形語周辺から上記の3タイプの単語の意味属性を抽出して、これと3．2に示した規則化した共起情報とを照合して行う。同形語の判定値Tは、タイプa，b，cごとの意味属性の一致度PaPbPcと各タイプの判定重みkakbkcの積和より以下で線形式で表現し、異なる読みの同形語W1，W2ごとに算定した判定値T1，T2を比較して同形語を認定する。ただしTが有効値δ（同形語ごとの固有の値）を超えない場合あるいはT1とT2が同じ値の場合は判定値の比較は信頼できないので認定不可とする。

$$T = F(k, P) = kaPa + kbPb + kcPc > \delta$$

ここで、意味属性の一致度Pは、抽出した共起する単語の意味属性が予め規則化した意味属性と一致する場合P＝1，包含関係の場合P＝0．5，一致しない場合P＝0とする。また、判定重みkは、判定に寄与する共起情報にタイプ別の片寄りがあることから設定する指標で『規則化した共起情報が判定に有効である比率』を示す。今回は実例からタイプごとに収集した意味属性について同形語の判定に十分かを机上で判断しこの比率を判定重みとした。具体的に、同形語「足跡」（ソクセキ／アシアト）について大量の新聞記事文章より収集した規則用共起情報から作成した判定重みkを表2に示す。また、タイプ別の判定重みを図2で比較すると述部用言のタイプが他の2タイプに比べて著しく低いことが判る。

上記の判定方式を用いると、例えば「研究の足跡が残る」の例文での判定では、同形語＝「足跡（ソクセキ）」の場合、タイプaの意味属性『思考』（研究）の一致度Pa＝1、タイプbの意味属性『存在』（残る）の一致度Pb＝1、また、「足跡（アシアト）」の場合、タイプaとタイプbの一致度はPa＝0、Pb＝1となり次の判定値Tの比較により足跡＝ソクセキと判定できる。

T1（ソクセキ）＝0.62＊1＋0.31＊1＝0.93
T2（アシアト）＝0.67＊0＋0.23＊1＝0.23
→ T1＞T2 → 認定成功（足跡＝ソクセキ）

表2　同形語認定評価のための規則作成用データ

見出し		足跡	
読み		ソクセキ	アシアト
出現数		24	22
共起情報数	修飾／並列	17	9
	述語用言	30	28
	格要素	8	19
判定重み＊	修飾／並列	0．62	0．67
	述語用言	0．31	0．23
	格要素	0．71	0．67

※新聞記事240日分より抽出　＊実データより机上で収集

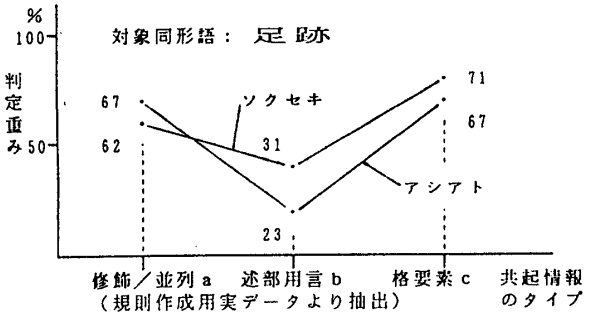

図2　実例における共起情報タイプ別の判定重み

4．共起情報を用いた同形語認定方式の評価と考察

3．1の条件に合う「足跡：ソクセキ／アシアト」を評価対象に選び、規則用データ作成とは異なる新聞記事より収集した表3に示す評価用データの「足跡」44件（内訳ソクセキ＝11件，アシアト＝33件）に対して以下の判定手段で同形語認定率を算定し判定方式の可能性について考察を行った。

表3　同形語認定評価のための評価用データ

読み		ソクセキ	アシアト
出現数		11	33
共起情報数	修飾／並列	9	13
	述語用言	7	33
	格要素	8	21

※新聞記事180日分より抽出

<判定法候補>

①タイプ別　：　A−1：T＝Paのみ
　　A−2：T＝Pbのみ
　　A−3：T＝Pcのみ
②重み付き　：　B　：T＝kaPa＋kbPb＋kcPc

<判定手段>

①有効値δ＝0
②判定値が同点の場合は認定不正解とする
③同一タイプの共起単語が複数ある場合、最も高い一致度を使用
④同形語認定率＝認定正解数／総評価数（＝44）

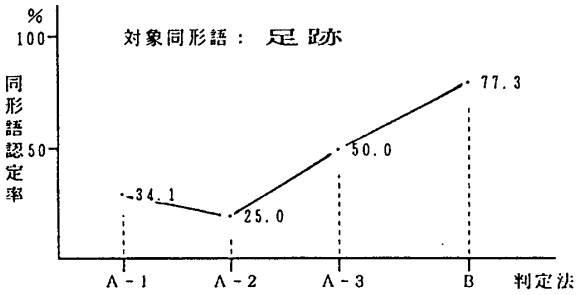

図3　共起情報による同形語認定率

<考察>

図3に同形語＝「足跡：ソクセキ／アシアト」の評価用データにおける同形語認定率を示す。これより、以下のことが判った。

①3タイプの判定重みを考慮した本方式の判定式は77.3%と有効
②原文中での共起情報の充足率が低いため一つのタイプの共起情報だけでは同形語認定は不十分（格要素での50%が最大）
③述部用言の共起情報は充足性は高いが判定し切れない場合が多いこれはタイプbでは、構文構造が埋め込み状の場合、周辺に位置する用言を誤って収集する可能性があるためで、以下の例では、「足跡」の述部用言として「開いた」を誤って収集する。

（例）足跡が　ヒヅメを　開いた　形に　変化する
… 正確な共起関係

5．おわりに

実際の新聞記事での単語解析や読み振りの不良から単独名詞型同形語の認定精度が低いことを分析するとともに、同形語周辺の3タイプの単語の意味属性の共起制約を判定することにより同形語を認定する方式を提案し、実データを用いて本方式の可能性を示した。今後は、本方式の実用性を高めるために、対象同形語数を拡大して判定規則を積み上げるほか、誤認定（同形語認定誤り）の調査と対策を検討する。

<参考文献>

[1]池原、安田、島崎、高木：日本文訂正支援システム（REVISE）、NTT研究実用化報告、Vol.36,No.9,pp.1159-1167 (1987)
[2] 宮崎：係り受け解析を用いた複合語の自動分割法、情処論、Vol.25,No.6,pp.970-979 (1984)
[3] 小原、高木他：日本文推敲支援技術、NTTR&D、Vol.40,No.7,pp.905-914 (1991)
[4] 吉田：自然言語理解（知能情報処理とロボット特集）、信学誌、Vol.65,No.4,pp.365-367 (1982)
[5] 宮崎、大山：階層的単語属性を用いた同形語の自動読み分け法、信学論、Vol.J68-D,No.3,pp.392-399 (1985)