単語正規化による固有表現の同義性判定

高橋いづみ，浅野久子，松尾義博，菊井玄一郎

日本電信電話株式会社　ＮＴＴサイバースペース研究所
{takahashi.izumi,asano.hisako,matuo.yoshihiro,kikui.genichiro}@lab.ntt.co.jp

1. はじめに

近年 Web 上の膨大な情報から特定の商品や人物などの情報や評判を獲得する技術の必要性が高まっている．しかし様々な書き手により作成された Web の文書では，同種の情報が複数の異なった表現（＝同義語）で記述されている．例えば「安藤美姫」には「安藤」「ミキティ」等の同義語が存在し，安藤美姫に関する情報をまとめるためには，用いられる可能性がある表現全てを収集し，辞書を作成しておく必要がある．本研究ではその第一歩として実世界の事物のテキスト表現である固有表現を対象とし，Web 上から収集した固有表現を用いて作成した同義語候補ペアに対し，同義性判定を行う手法の確立を行った．本研究では２つの固有表現が１度でも同じ意味で使われていることがあれば「同義である」と定義する．これは収集した同義語から辞書を作成し，実体の情報をもれなく収集することを最終目的としているためで，実際にそのテキスト中において同義か否かについては多義解消問題として切り分ける．

同義性判定は従来研究においては必ず同義語候補の収集とセットで扱われ，任意の手法で作成された同義語候補ペアに対して同義性判定を行う研究はない．これは同義語には，「ミキ，ミキティ」のように表記が類似しているほど同義の可能性が高い場合と，「木村拓哉，キムタク」のように表記の類似度だけでは同義性が計れない場合が混在しており，特定の同義語の種類に特化して同義性判定をさせざるをえないためである．

主な同義語獲得のための従来手法は識別手法と生成手法の２つで，識別手法は任意のテキスト集合から同義語候補ペアを収集して同義か否か判定する．ペアの作成時に表構造，特殊な表現（○○こと××）などのメタ情報を利用したり [1]，文字種や略語関係等の制約を設ける [2] ことで一定の精度を担保している．しかし対応する同義語候補ペア以外の同義性判定は行えず，カバー範囲が狭いという問題があった．また同義性判定時には文脈情報を用いる場合が多く，本研究の「同義である可能性があれば同義とする」タスク設定には合わない．

生成手法には，確率モデルを用いてある文字列表現の同義語候補として考えられる表現を全て作成したのち，Web を用いて実在を確認する手法 [3] がある．しかし，生成できる同義語候補が略語やカタカナ異表記など特定の種類に限定される，大量に無関係な同義語候補を生成してしまうため同義性判定の処理量が多い，精度が低いなど，未だ不十分な点が多い．

以上から獲得可能な同義語のカバー範囲を広げるためには，同義語の種類に特化せずに同義性判定が行えるようにする必要がある．また，入力に多様性のある同義語ペアを多く含むよう，緩い制約により同義語候補ペアを収集しなければならない．しかしその際には無関係な同義語候補ペアを極力排除し，精度低下や処理量の増加を防ぐことも必要となる．よって我々は最低限の制約により収集した同義語候補ペアに対して，従来手法と比較してなるべく精度を下げずに，より広範囲の同義性判定が可能な手法の確立を目指す．

2. 同義語の分類

どのような同義性判定手法を取るべきかを念頭に，同義語を表記とその読みの知識のみで同義性の推測が可能なもの（「安藤美姫，ミキティ」）と，それ以外の知識が必要なもの（「松井秀喜，ゴジラ」「オランダ，Netherlands」）の２つに分けた．その割合は約９：１であり，前者が圧倒的に多いといえる（4.1 評価用データ参照）．その表記又は読みから推測が可能な同義語を派生型同義語と定義し，本報告における同義性判定対象とした．よって同義性判定に文脈情報は用いない．

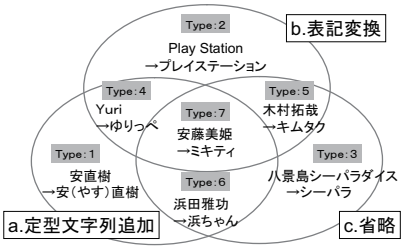

図 1: 派生型同義語の分類

対象とした派生型同義語は，標準的な表記に対して何かしらの操作を受けて生成されたものと定義し，例えば「安藤美姫」からは「安藤」「美姫」「ミキ」「ミキティ」等が生成される．この派生型同義語を，どういった操作が行われたかという観点で図１のように分類した．操作は次に挙げる３つで，生成される同義語は７種類となる．

a. 定型文字列の追加 ：接頭/接尾辞等の文字列を追加

b. 表記変換 ：読みを保存して表記を変換

c. 省略 ：文字順を保存して文字を削除

a. は実体を呼ぶ際にニュアンスや説明を付加する意味で「安直樹，安（やす）直樹」の“（やす）”や，「ミキ，ミキティ」の“ティ”等を追加する操作で，この操作単体で生成される同義語を Type1 とする．追加する文字列は，接頭辞や接尾辞，読み等定型的な文字列とし，操作後の表記から操作前の

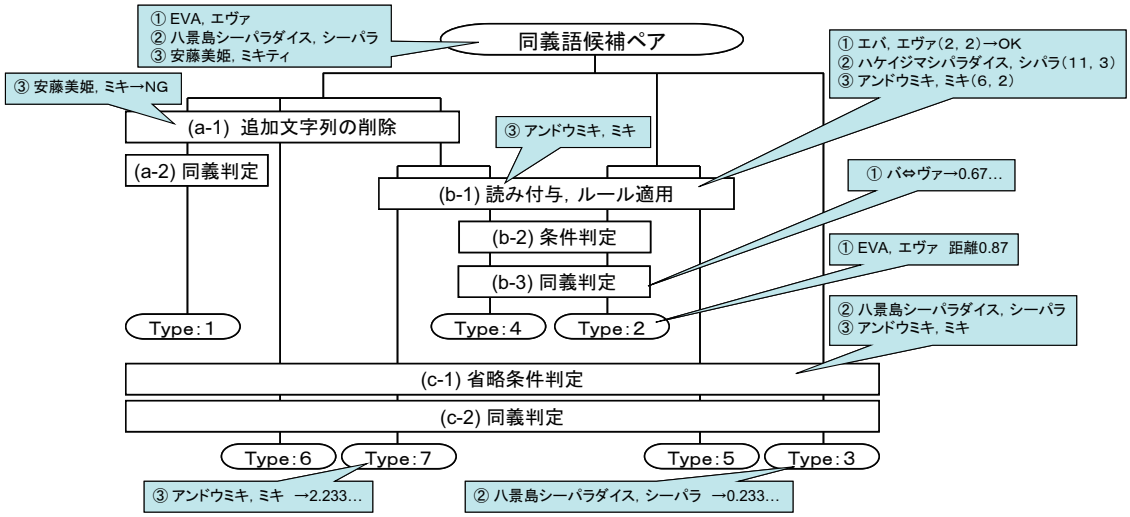

図 2: 同義性判定の処理フロー

表記が容易に推定可能なものと定義する．**b.** は文字種のゆれ「Play Station，プレイステーション」や，カタカナ表記ゆれ「プレイステーション，プレーステーション」等，読みを保存したまま表記を変換する操作で，単体で Type2 を生成する．**c.** は「八景島シーパラダイス，シーパラ」,「North Atlantic Treaty Organization，NATO」等の省略語を作成する操作で，Type3 を生成する．操作前後の同義語は包含関係になると定義した．

a.b.c. の組み合わせで生成される同義語は Type4〜7 で，「安藤美姫，ミキティ」を例にとると「安藤美姫」が **b.** により「アンドウミキ」になり，**c.** で「ミキ」と省略され，**a.** で“ティ”が挿入されたと考える．

3. 提案手法

3.1 着眼点

派生型同義語には 1 章でも説明したように，(i) 表記及び表記から推測される読みをヒントに同義性判定が行える場合（前章の **a.** や **b.** による）と，(ii) それだけでは同義性が計れない場合（**c.** による），さらにその両方の特徴を備えている場合が混在している．そこで，同義語候補ペアがどの種類の派生型同義語なのか条件判定を行って分離し，ペアの種類ごとに適した同義性判定手法を適用する．

(i) の同義性判定には表記や読みの一致度を用いる．**a.** の操作により追加されたと思われる文字列を推定して削除し（＝表記の正規化），ペアが全く同じ表記になれば Type1 と判定する（文字列追加判定）．次に，同義語候補ペアに読みを付与して正規化を行った後，その類似度で Type2 の判定を行う（表記変換判定）．また，表記正規化後の同義語候補ペアに読み正規化を行えば Type4 が判定できる．

(ii) は省略語であり，**c.** の操作にはある程度法則性があるとされる [3] ため，機械学習により Type3 を判定する（省略判定）．

(i)(ii) 両方の性質を持つものに関しては，表記や読み両方の正規化を行うことにより同義語ペア同士が包含関係となり，Type5〜Type7 に対しても省略判定をすることが可能となる．これにより全ての種類の派生型同義語に対応でき，カバー範囲を広げることが可能になると考えた．

以下からはその手順を図 2 を用いて説明していく．入力された固有表現の同義語候補ペアに対して，文字列追加判定，表記変換判定，省略判定およびその組み合わせの全ての判定を行い，同義と判定された時点で Type が確定する．全部で 7 通りある判定の中で，1 回も同義と判定されなければ同義語ではないと判定する．

入力となる同義固有表現ペアは，1 テキスト内の固有表現総当りで作成することとした．総当りであれば，テキストの記述方式に依存せず同義の可能性のある同義語候補ペアを全て作成することが可能である．また 1 テキスト内という制約によって，テキストを跨いで存在する無関係な同義語候補ペアを作成すること，組み合わせ可能性の爆発を防ぐことが出来る．1 テキスト内に同義固有表現のペアが存在するかについては，予備実験を行い大量の文書を集めれば 98 %の確率で存在することを確認した．

3.2 文字列追加判定（表記正規化による判定）

追加文字列を削除する表記正規化には，ヒューリスティックに記述した 4 7 パターンの文字列正規表現ルールを使用する．削除するのは接尾辞，接頭辞，読み，記号等など定型的なものとし，ある特定の固有表現にのみ起こる事象はルール作成の対象外とした．例えば「安（やす）直樹，安直樹」のペアは，まず表記正規化により読み仮名の（やす）が削除されるとペア同士が全く同じ表記になるため，同義と判定する．

3.3 表記変換判定（読みの正規化による判定）

処理の流れは以下のようになる．

（**b-1**）同義語候補ペアに形態素解析器を用いて読みを付与

（**b-2**）ルールを用いて読みを正規化し，適用後の音節数が一致するかの条件判定を行う

（**b-3**）条件に一致したものに対して距離計算

（b-2）で用いるルールとは，読みの長音，促音，母音連続に対して適用するヒューリスティックなルールで，母音連続時に長音化や拗音の直音化をしたり，長音，促音の削除を行う．例えば「ウィンブルドン」は「ウインブルドン」,「ヨウコ」は「ヨコ」となる．

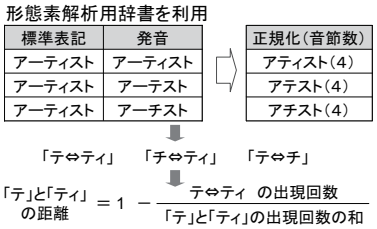

図 3: 読みの近さの距離

（b-3）の計算には，統計的に求めた「読みの近さ」を利用した．読みの近さの求め方は，まず形態素解析用辞書から標準表記が等しく読みが異なる単語を収集する（図 3 参照）．そして辞書に登録されている読みに対して上記の（b-2）を行い，音節数が等しい場合に同じ音節位置であるが表記が異なる音節を抜き出し，出現回数を調べて類似度を計算する．

例を図 2 の (1)「ＥＶＡ，エヴァ」のペアを用いて説明する．まずカタカナに変換すると,「エバ，エヴァ」となり，それぞれの音節数は共に 2 であるため，距離計算を行う．距離計算は，同じ位置にある音節間の読みの類似度の和とする．類似度は 0 〜 1 の間の値となっており，全く同じ表記である「エ」同士では距離 0，表記の異なる「バとヴァ」ではこのペアの類似度である 0.87 である．よってこのペア間の距離は 0.87 となり，あらかじめ決めた閾値により同義か否かを判定する．

3.4 省略判定（機械学習による判定）

包含関係にあるという条件を満たす同義語候補ペアにのみ SVM（TinySVM *を利用）を使って略語か否かの 2 値判定を行った．処理の流れは以下になる．

（**c-1**）同義語候補ペア間で DP マッチを用いてアラインメントをとり，包含関係にあるかを確認

（**c-2**）省略前後の表記の差異から素性を抽出し， SVM を用いて判定

判定のためのモデルの学習に利用したデータは，評価用データと同じドメイン，収集方法で集めた文書セット（ブログ 2,200 テキスト，ニュース 1,600 テキスト）である．これに固有表現抽出器を用いて固有表現を抽出，1 テキスト内総当りで作成した同義語候補ペアのうちの包含関係にある 4,002 対に対し人手で正解を付け，学習セットとして用いた．正否の割合は 3 ： 2 であった．

処理の例を，図 2 の (2)「八景島シーパラダイス，シーパラ」のペアについて，図 4 を用いて説明する．ペアは包含関係にあり省略判定条件に一致するため，SVM で判定を行う．省略前後の固有表現，削除された形態素と文字，残った形態素と文字それぞれについて，図 4 に示したような素性を用いた．SVM による判定結果が定めた閾値以上であれば同義であると判定する．

* TinySVM : http://chasen.org/ taku/software/TinySVM/

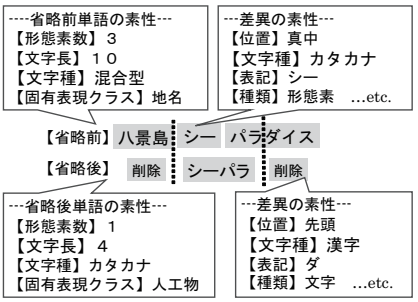

図 4: 省略判定に用いた素性の例

3.5 組み合わせによる判定

それぞれの判定単体では同義とならない場合について図 2 の (3)「安藤美姫，ミキティ」のペアで説明する．まず文字列追加判定時の表記正規化により挿入部分が削除され,「安藤美姫，ミキ」となるが，この時点では表記が等しくないため，同義と判定されない．次に表記変換判定で読みの正規化を行うと,「アンドウミキ，ミキ」となるが，音節数が異なるため，条件判定を通らない．しかしこの時点でペアが包含関係になるため省略判定の条件と一致し，SVM により同義か否かの 2 値判定が行われる．同義と判定されれば，このペアは **a.b.c** 全ての操作で生成された Type7 に確定する．以上のように，３つの判定部を組み合わせることで全ての派生型同義語を判定対象とすることを可能とした．

4. 評価実験

4.1 評価方法及び評価用データ

提案手法の有効性を調査するために，同義語候補ペアを用いた同義性判定実験を行った．実験の評価用データとしてブログと新聞記事 2,030 テキストに対し，固有表現を人手で抽出後，1 テキスト内の固有表現総当りでペアを作成したものに同義か否かの情報を人手で付与した．総ペア数 314,661 ペアのうち，同義語数は 4,562 ペア，派生型同義語は 3,703 ペアであった．この同義語中約 9 割をしめる派生型同義語を，今回の実験の正解とする．また，正解である同義語ペアの種類は表 1 のようになった．本来同義性判定の評価には，判定に成功したペア数を評価用データの全ペア数で割った正解率を用いるべきだが，評価用データの不正解の割合が 99 % と非常に高く，正解率は性能評価に不適切であった．そこで式 1 の精度と式 2 の再現率を用いて評価を行った．

$$\text{精度} = \frac{\text{システムが出力した正解同義語ペア数}}{\text{システムが出力した同義判定ペア数}} \quad (1)$$

$$\text{再現率} = \frac{\text{システムが出力した正解同義語ペア数}}{\text{正解同義ペア数}} \quad (2)$$

4.2 ベースライン手法

比較のためのベースライン手法として，同義語候補ペア間の編集距離を用いて同義性判定を行った．編集距離とは２つの文字列の一方を操作し，もう一方の文字列に変換するための最小コストであり，従来研究においても２つの文字列の距離をあらわす尺度として広く用いられている．編

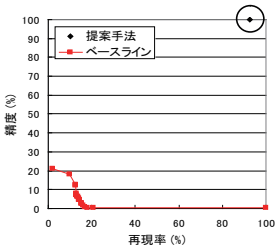

図 5: 文字列追加判定結果

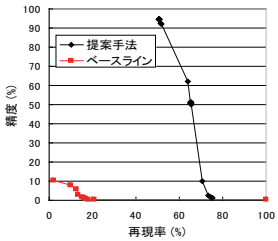

図 6: 表記変換判定結果

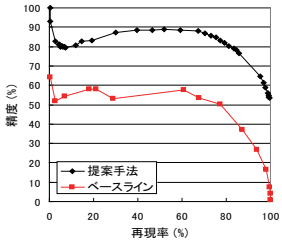

図 7: 省略判定結果

表 1: 評価用データの割合

同義語分類	ペア数	割合 (%)
追加（Type1）	41	1
変換（Type2）	668	17
省略（Type3）	2262	61
追加＋変換（Type4）	13	1
変換＋省略（Type5）	539	14
追加＋省略（Type6）	194	5
追加＋変換＋省略（Type7）	75	2

※追加：文字列追加，変換：表記変換

集距離に基づき式 (3) を定式化して類似度とし，閾値を設けて同義性判定を行った．

$$\text{類似度} = 1 - \frac{\text{編集距離}}{\text{ペアの総文字数}} \tag{3}$$

4.3 実験結果

提案手法およびベースライン手法を用い，同義語候補ペアに対して同義性判定実験を行った結果，再現率-精度曲線は図 8 となった．文字列追加判定，表記変換判定，省略判定の個別の結果はそれぞれ図 5，図 6, 図 7 である．３つの判定処理のうち，文字列追加判定と省略判定においては判定結果が値として出力されるため閾値が存在するが，文字列追加判定は同義か否かで判定を行うため，結果は曲線とならない．組み合わせた結果に関しては，文字列追加判定の閾値を単体で最も精度がよかった 0.6 に固定して，省略判定の閾値を代えて結果を出してある．

実験の結果，ベースライン手法では精度，再現率共に 60 %を超えなかったため，同義性判定は簡単な問題でないことがわかる．提案手法は精度 67.45 %，再現率 71.57 %と，ベースライン手法より精度 13.59 %，再現率が 19.87 %向上した．表記類似，読み類似，省略それぞれの判定においては表記類似度による手法と比較して精度，再現率ともに大幅に向上した．判定を組み合わせた場合に関しては，単独で判断を行った場合よりも精度は下がった．

4.4 考察

それぞれの判定手法について考察を行う．文字列追加判定結果の再現率が 100 %でないのは，文字列追加判定のためのルールを作成する際に 1 実体に特化したものを対象外としたためである．表記変換判定については，ベースライン手法と比較して精度，再現率ともに大幅に向上した．判定誤りには３種類あり，(1) 読み付与と正規化が失敗したために，距離計算まで達しない（５２%），(2) 距離計算による誤判定（４６%），(3) 読み正規化失敗による誤判定（２%）となっている．(1) の主な原因は，韓国語や中国語，記号など形態素解析器の読み付与機能が未対応なものへ正しい読みが付与できない場合がほとんどであった．省略判定についても精度，再現率共に向上した．グラフの左側に向けて精度が下がる原因は，略語の同義語候補ペアのうち形態素単位でいくつも削除が起きているようなペアは，省略の起きる場所によって単語の意味が全く異なってしまい，判定が難しいためである．

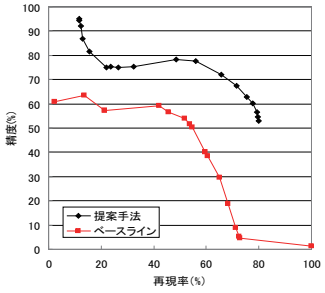

図 8: 提案手法とベースライン

5. おわりに

本テーマではテキストからの同義固有表現ペアの同義性判定に取り組んだ．文書中に存在する同義語ペアの約９割を占める派生型同義語を対象として，その生成過程において生じた表記と読みの変化を正規化することで，様々な種類の派生型同義語全ての同義性判定を可能とする方法を提案した． 1 テキスト中の固有表現総当りで作成した同義候補ペアに対する同義性判定実験では，提案手法は精度 67.45 %，再現率 71.57 %と，単語の表記類似度を用いたベースライン手法より精度 13.59 %，再現率が 19.87 %向上し，提案手法の有効性が確認できた．今後は考察で述べた問題点を改善すると共に，同義と判定された固有表現ペアを実体ごとに集約して同義固有表現辞書を作成する予定である．

参考文献

[1] 関恒仁, 嶋田和孝, 遠藤勉. 表の構造を利用した類義語抽出. 言語処理学会第１１回年次大会発表論文集, pp. C1–6, 2005.

[2] 酒井浩之, 増山繁. 略語とその原型語との対応関係のコーパスからの自動獲得手法の改良. 自然言語処理 Vol12, No4,, pp. pp.207–231, 2005.

[3] 村山紀文, 奥山学. Noisy-channel model を用いた略語自動推定. 言語処理学会第１２回年次大会発表論文集, pp. pp.763–766, 2006.