5A-1

仕様書特有表現の表記揺れを検出するツールの試作と評価

久野綾子† 平尾英司† 田村一樹‡ 吉川大弘‡ 古橋武‡
日本電気株式会社 情報・ナレッジ研究所† 名古屋大学大学院 工学研究科‡

1. はじめに

システム開発に関わる仕様書は、その作成に複数の人間が関わることが多いため、表記揺れが発生しやすい。表記揺れは言葉の解釈に混乱を生じるため、事前に抽出して統一する必要がある。文書中の表記揺れを検出する手法として、Word の校正機能のように一般的な表記揺れパターンを登録した辞書を用いる手法や、大規模コーパスを用いて文字列の連続確率等を学習する手法[1]、同じくコーパスを用いて文脈の類似度を算出する手法[2]が提案されている。しかし、一般的な文書と異なり、仕様書には個別の案件に特有の複合語を大量に含むという特徴がある。案件に特有の複合語が多い文書では、例えば「振込み金額」が正解である文書における「振込み額」のように、仕様書内でのみ成り立つ表記揺れが多数発生する。このような表記揺れは、形態素単位での語の正しさや文字列の並びの普遍性を解析しても検出できないため、既存のツールや手法だけでは仕様書の表記揺れを十分に検出できない。そこで、本稿では個別の案件毎の仕様書特有表現の表記揺れを検出する手法について提案した。さらに試作ツールによる有効性の評価結果を報告した。

2. 仕様書特有表現の表記揺れ検出指標

仕様書特有表現の表記揺れを検出するため、これらの表記揺れ状態にある複合語の特徴を調査した結果、以下の 3 パターンが抽出された。

A) 「正しい記載の複合語」と「表記揺れの状態である複合語」間の文字列は類似しやすい
B) 複合語単位では、文字列が長い複合語ほど表記揺れとなる可能性が高い
C) 「正しい記載の複合語」に比べ「表記揺れの状態である複合語」の出現数は極端に少ない

この結果から、複合語間の文字列のズレが小さく、出現数の偏りが大きい組み合せを抽出することで、表記揺れが抽出されると期待できる。

Detection of Term Variation in Specifications
†Knowledge Discovery Research Laboratories, NEC Corporation
‡Graduate School of Engineering, Nagoya University

そこで、これらのパターンに基づき、複合語 a,b の組み合せの一方が表記揺れである可能性を定量化する指標として $score_A$ (a,b) 、$score_B$ (a,b) 、$score_C$ (a,b) という 3 つの指標を定義した。以下に各指標の詳細を解説する。

$score_A$ (a,b) は、A)の特徴を反映し、複合語 a と複合語 b 間の文字列の相違を表す編集距離 x と、表記揺れである可能性の関係を定量化した指標である。表記揺れである確率は編集距離 x が大きいほど下がり、正規分布に従うと仮定し、(1) 式のような関数で指標化した。

$$score_A = \exp(-(x-1)^2) \quad \cdots \quad (1)$$

$score_B$ (a,b) は、B) の特徴を反映し、複合語 a と複合語 b 間の文字列長に対する編集距離 x の割合と、表記揺れである可能性の関係を定量化した指標である。表記揺れである確率は編集距離 x が大きいほど下がり、文字列が長い方の文字数 L が大きいほど上がるよう、(2) 式のような関数で指標化した。

$$score_B = 1 - \frac{x}{L} \quad \cdots \quad (2)$$

$score_C$ (a,b) は、C) の特徴を反映し、複合語 a と複合語 b 間の出現頻度の偏りと、表記揺れである可能性の関係を定量化した指標である。表記揺れの発生しやすさは複合語 a の出現回数 N_a と複合語 b の出現回数 N_b の比率の偏りと正比例の関係にあると仮定として、(3) 式のような関数で指標化した。

$$score_C = 2 \times \left| \frac{N_a}{N_a + N_b} - 0.5 \right| \quad \cdots \quad (3)$$

すなわち、(3) 式の指標では複合語 a の出現回数 N_a と複合語 b の出現回数 N_b の比率が、均等であった場合 (0.5) と比べ、どの程度離れているかを求めている。

表記揺れの抽出に置いて指標は一元化されていることが望ましいため、これらの 3 指標の線形和を取った表記揺れ指標 S (a,b) を以下の (4) 式で定義し、これを表記揺れの可能性を示す指標とした。α、β、γ は各指標の重みである。

$$S(a,b) = \alpha \cdot score_A(a,b) + \beta \cdot score_B(a,b) + \gamma \cdot score_C(a,b) \quad \cdots (4)$$

3. 表記揺れ検出方法

表記揺れ検出手法は「複合語抽出」、「表記揺れ指標の算出」、「誤検出パターンの除外」の3ステップから構成される。

複合語抽出

仕様書に多く含まれる案件特有の造語として作られた複合語に対応するため、まず入力テキストから複合語を抽出する処理を行う。具体的には入力テキストを形態素解析し、特定の品詞（主に名詞）をつなぎ合わせる。

表記揺れ指標の算出

次に、抽出した全複合語ペアを一対比較し、2節で述べた表記揺れ指標Sを算出する。

誤検出パターンの除外

表記揺れ指標 S（a,b）には、S の値が高い複合語ペアであっても表記揺れではない誤検出パターンとして、①複合語間が接辞（接頭辞及び接尾辞）のみ異なる（例：“経理部門”-“経理部門内、“運用受託者”-“運用受託後”）、②複合語間が「対で使われやすい語」のみ異なる（例：“入力情報”-“出力情報”、“内部設計”-“外部設計”）、という 2 パターンが存在する。そこで、①の誤検出パターンに対しては、接頭辞 162 語・接尾辞 196 語を登録した接辞辞書を用意し、②の誤検出パターンに対しては「対で使われやすい語」（“入力”-“出力”等）を1258 組登録したペア語辞書を用意し、各誤検出パターンに該当するケースを候補から除外するようにした。なおペア語辞書の作成にあたっては、意味上類似している単語ペア（“輸送”-“移送”等）は、実際の表記揺れの可能性があるため、登録しないように配慮する必要がある。

4. 表記揺れ検出ツールの試作と評価

3 節で述べた手法の有効性を評価するためツールを試作した。複合語抽出の際の形態素解析には「MeCab」[3]を利用した。試作したツールを実案件の仕様書に適用し、誤検出対策の有無による表記揺れ検出精度の違いを比較した結果を表 1 に示す。(4)式の重みは、α=0.3、β=0.3、γ=0.4 と設定し、指標 S を算出した。指標 S が 0.7 を下回ると、正解数が極端に少なくなることから、0.7 以上の候補ペアを目視で確認し、実際に表記揺れと思われるペアを正解としてカウントした。網羅率は、誤検出対策なしの正解数を100 とした場合の正解数の割合である。

表 1 誤検出対策による表記揺れ検出精度の比較

誤検出対策	指標 S≧0.7 の候補数	正解数	正解率	網羅率
なし	969	133	14%	100%
接辞考慮	401	123	31%	92%
接辞＋ペア語を考慮	308	123	40%	92%

接辞、ペア語の辞書を導入することで、網羅率の低減は 8%に抑えつつ、候補数を 68%削減し、正解率を約 3 倍向上させることができた。なお、誤検出対策を行ったことで、網羅率が下がった原因は、接辞辞書に登録した文字が接辞以外の使われ方だった場合も除外してしまったためである。これは、形態素解析で実際に接辞として使われているケースのみ除外することで、改善可能と考えられる。また、誤検出対策で除外しきれなかった誤検出の原因は主に 3 パターンに分類され、①誤検出パターンの把握不足（例：“想定”-“想定額”）、②複合語の内部にある接辞の考慮不足（例：“計算回数”-“計算総回数”）、③文字列は似ているが案件内で区別して使い分けている語の解析ミス（例：“郵便番号”-“先郵便番号”）、であった。①に関しては辞書の充実、②に関しては複合語の内部にある接辞の除外、③に関しては案件特有の用語登録にて改善可能と考えられる。

5. まとめ

本稿では、編集距離と出現頻度の偏りに基づき、仕様書特有表現の表記揺れの指標を算出する手法を提案した。提案手法を用いることで仕様書に特有の表記揺れを検出できることを確認した。また、誤検出パターンを除外することで、高い網羅率を維持しつつ、正解率を約 3 倍向上させた。今後、明らかになった誤検出例への対応や各指標の重みの最適値の把握などを検討してゆく。

参考文献

[1] 河田他, “両方向 N-gram 確率を用いた誤り文字検出法”, 電子情報通信学会論文誌 Vol. J88-D-II No.3, 2005.

[2] 増山他, “大規模コーパスからのカタカナ語の表記揺れリストの自動構築”, 第 10 回言語処理学会年次大会発表論文集, 2004.

[3] MeCab, http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html