

翻訳文における表記揺れの検出・訂正

5 J-3

日本アイ・ビー・エム(株)

東京基礎研究所

脇田 早紀子

1. はじめに

日本語校正支援システムFleCSは、92年末頃より新聞社ユーザーの実務に用いられてきた。その中で表記揺れを扱うことはよくあったが、例えば「コミュニケーション」と「コミュニケイション」、「行う」と「行なう」のように、正表記と誤表記が単語単位で決まっているようなものがほとんどであった。つまり、後者を発見したら前者に修正を求めるように禁止語辞書を作っておけばだいたい用が足りていた。

ところが、昨年より弊社の作成するマニュアル文書の校正に用いるようになって、それだけでは不便なことが増えてきた。マニュアル文書の多くは英語からの翻訳文である。一つの英単語（「target」）の訳として複数の日本語（「宛先」「目標」「ターゲット」）がある場合がよくあるが、そのうちのどれかだけが正しいのではなくて、ひとまとまりの複合語や分野などの状況によって決まる（「target database name」は「宛先データベース名」、「target application name」は「ターゲット・アプリケーション名」）。複合語は次々新しく作られ、正表記の変更もある。なにより、誤表記は翻訳者によりまちまち（「target database name」に対して「ターゲット・データベース名」「宛先データベースの名前」など）なことから、とても禁止語辞書のような形では抑えきれない。

そこで、比較的良好にメンテナンスされている『翻訳者用辞書』（注2）を元にして、なるべく手間をかけずに翻訳文における表記揺れを検出・訂正する目的で、『翻訳揺れ検出』を作ったので報告する。

2. 従来の方法

これまでFleCSが採用していたのは以下の3方式である。

未登録語警告方式

正表記（「ネットワーク」）に対して誤表記（「ネッットワーク」「ネッットワアク」）は登録されていないので、形態素解析が失敗する部分を未登録語として検出することができる。この方式は予想していない誤り検出できるのが利点だが、形態素解析でできてしまう誤表記（「ネッット・ワーク」）や、場合により使い分ける必要がある語に対しては無力である。

禁止語辞書方式

誤表記を「禁止語辞書」に登録し、置き換え語として正表記を登録すれば、簡便かつ確実に検出でき、置き換えるべき正表記の提示もできる。ただし、登録された通りの誤表記以外には無効である。

カタカナ語登録方式

カタカナ語が出現するたびに、発音に直してメモリーに蓄積しておく。発音が似ていて（「バイオリン／ヴァイオリン」、「インタフェース／インターフェース」）、違う表記のものが現れると検出する。正表記、誤表記ともに未知のものに対して有効であるのが強みだが、発音が似たカタカナ語限定の方式。

3. 目的

今回は、文字列として一致する部分が大いものだけでなく、見た目が似ていないもの（「ストラクチャー」と「構成」「構造」）も似ているとして扱う必要があることを考慮して、以下の条件を満たす『翻訳揺れ検出』を作成することを目的とする。

- ・同じ英語表記から翻訳されたい語列は似ていると判定する。
- ・可能性のある誤表記をあらかじめ列挙する必要があるようにする。
- ・「翻訳揺れ検出」独自の辞書を人手でメンテナンスする必要があるようにする。

4. 翻訳揺れ検出の仕組み

『翻訳者用辞書』にある訳語を「正表記」として、それに「似ていて」異なる表記を探す。

『翻訳者用辞書』例：

global resource serialization complex 大域資源逐次化システム <ES>
 global search グローバル・サーチ、一括検索
 global search 広域検索、グローバル・サーチ <RS>
 global selection 大域選択 <ES>
 global service グローバル・サービス、大域サービス
 global shared resources (GSR) グローバル共用資源、大域共用資源 (GSR)
 global sharing グローバル共用 <PS>
 global stack グローバル・スタック、大域スタック
 global storage グローバル記憶域、大域記憶域
 global symbol グローバル記号、大域記号
 global system lock グローバル・システム・ロック、大域システム・ロック

「似ている」ことの定義を以下に示す。

まず、「正表記」を元に、複合語は単語に区切り、「の」「な」「・」を除いたものを「標準表記」と呼ぶ。以下、区切りはピリオドで示す。

例：「グローバル・システム・ロック」→「グローバル.システム.ロック」

「広域検索」→「広域.検索」

また、「翻訳者用辞書」において、複合語を除く一つの英単語に対して併記されている訳語のリストにあるものを「同義語」とする。

例：「グローバル」「大域」「広域」

そして、「標準表記」の区切りごとに、

1. 平仮名一文字または中黒を挟む

2. 「同義語」に置き換える

の2種類の操作を行うことにより一致させ得るものを「似ている」と判定する。

このようにして、「似ている」と判定されたもので、かつ、「標準表記」の元となった「正表記」と一致しないものに対して警告し、置き換えるべき語として対応する「正表記」を提示する。

例：「グローバル.システム.ロック」を「標準表記」として「グローバルなシステム・ロック」「広域システムロック」などを検出し、「グローバル・システム・ロック」が提示される。

5. 翻訳揺れ検出の実装

FleCSのパターンの一つとして実装した。形態素解析用の辞書とは別に「標準表記辞書」「同義語辞書」「正表記辞書」を用意した。この3つの辞書は、前節の定義に基づき、「翻訳者用辞書」をもとにして自動で生成する。

「標準表記辞書」により、頭の2文字を鍵として、「似ている」可能性のある「標準表記」をすべて得ることができる。

例：「グロ」→「グローバル.システム.ロック」

「同義語辞書」により、区切り単位の文字列を鍵として、それと置き換えられる語を得ることができる。

例：「グローバル」→「大域」「広域」

「正表記辞書」により、「標準表記」を鍵として、「正表記」を得ることができる。

FleCSは文章を解析する際、まず現在位置からの2文字を鍵として「標準表記」を得て、そのひとつづつについて「似ている」かどうか調べる。すなわち、区切り単位ごとに「同義語辞書」を利用して置き換え、または平仮名一文字か中黒を挟むことで一致させることができるか調べる。成功すると、「正表記辞書」を用いて「正表記」を得て、一致しない場合は警告し、置き換える語として「正表記」を提示する。

6. 結果の出力例

文章「…グローバル資源の逐次化システム…」

→出力「似た正表記あり:大域資源逐次化システム」

文章「…グローバル検索…」

→出力「似た正表記あり:グローバル・サーチ,一括検索,広域検索」

7. 問題点

「正表記」が複数あるものもある。そうすると、いずれの「正表記」についても他方の「正表記」に「似ていて」異なる語として検出されてしまう。このようなものの中には、分野ごとに「正表記」が定められているものと、一冊の本の中でどれかに統一されていればよいものがある。いずれも後処理によって不要な検出を削除することを考えている。

8. まとめ

「翻訳者用辞書」をもとに、「正表記」と「似ていて」違う（同義語で置き換え、「の」「な」「・」のあるなしなど）表記の語を検出し、「正表記」を提示する仕組み「翻訳揺れ検出」を作成した。実際の使い勝手については、発表の際に報告する。

謝辞

ユーザーとして要望を出し、材料を提供し、いつも研究に協力して下さっている日本アイ・ビー・エム(株)NLS滝北さんに感謝します。

参考文献

[1]奥村ほか：日本語校正支援システムFleCSの新聞社における実用化, 情報処理学会自然言語処理研究会92-NL-91, (1992)

[2]奥村ほか：日本語校正支援システムFleCS-新聞社における実用化報告, 情報処理学会第45回全国大会3F-5, (1992)

(注1)表記は「」、本報告で定義した用語は「」で示している。()は例である。

(注2)「翻訳者用辞書」は、人間の翻訳者が参考にするため用いている対訳辞書である。