# 明細書中にある文の表現のゆれ検出
## ― 翻訳前の文章表現の統一 ―

Effective detection of sentences having difference in expression

株式会社クレステック　IT ドキュメントセンター　システムコンサルタント
**楠本　浩二**

1986 年九州工業大学大学院修士課程情報工学専攻修了。構造化文書の研究開発に携わり、例規管理システム、法令審査支援システム、損害・生命保険約款チェックシステム、製造業マニュアルの文書編集・比較、精査システムの設計・開発に従事。

k-kusumoto@crestec.co.jp

## 1　はじめに

日本語は他の言語に比べて言語表現の多様性、すなわち「表記ゆれ」が顕著である。これが言語処理の様々な問題をより困難にし、文章の再利用、文書の検索に大きな支障を及ぼしている [1]。本稿ではこの表記ゆれを一文または複数の文のレベルに拡大した言語表現の多様性を考える。例えば、2 つの文「弁の(1)が下方向に押され、ポンプからの油がバルブに流れる。」と「弁(1)が ON になり、ポンプからの油をバルブに供給する。」は、伝えたい内容としては同一であるが、表現としては差異が発生している。このような文の間にある関係をここでは「表現のゆれ」と呼ぶ。

言語を計算機で処理すると、この表現のゆれが思わぬ結果をもたらすことがある。例えば、ゆれが発生している日本語文を原文として他の言語に翻訳した場合、両者の意味に重大な差が生じることがある。翻訳文が原文本来の意味と異なると、翻訳コストだけでなく、特に特許の場合、権利の範囲に影響を及ぼしかねない。このため、原文としての日本語文を計算機で処理する前に、このような表現のゆれを検出し、統一した正確な表現にしておくことが望まれる。

表現のゆれを検出する一手法として、まず対象となる文すべてを比較し、類似している文を検索することを考える。対象となる文の集合を $U$ とすると、集合 $U$ が単一の文書の範囲ならばその文書内に限定したゆれの検出になり、集合 $U$ が複数の文書要素を含むならば文書を横断したゆれの検出になる。検出した文と別の文の相違箇所を確認し、表現のゆれの修正が必要かを判断する。

## 2　課題と目標

類似検索は、対象の情報空間に任意のクエリを与えて類似した情報を取り出す処理が一般的である。文の間のゆれの検索では、ユーザーがクエリを作成するのではなく、個々の文自身がクエリとなって、別の文と順次比較する [2]。単一の文書または複数の文書にある文同士すべてを検索して比較する場合、全体の集合 $U$ の要素数を $N$ とすると $N\times(N-1)/2$ 回の比較が必要である。

しかし、この方法では文の数の増加とともに時間が指数関数的に増大する。また文のすべてを比較する場合、英数字だけの文と日本語文との比較や、日本語同士でも表題、品目、箇条書項目と叙述文との比較は無意味であり、無駄な比較計算となる。

本稿では、ゆれが発生している可能性がある文を網羅的に効率よく検出することと、検出した文の間で表現のゆれが発生している場合、修正の判断ができるようにその間の表現の差異を表示することを目的とする。

## 3　処理概要

処理の流れを図 1 に示す。

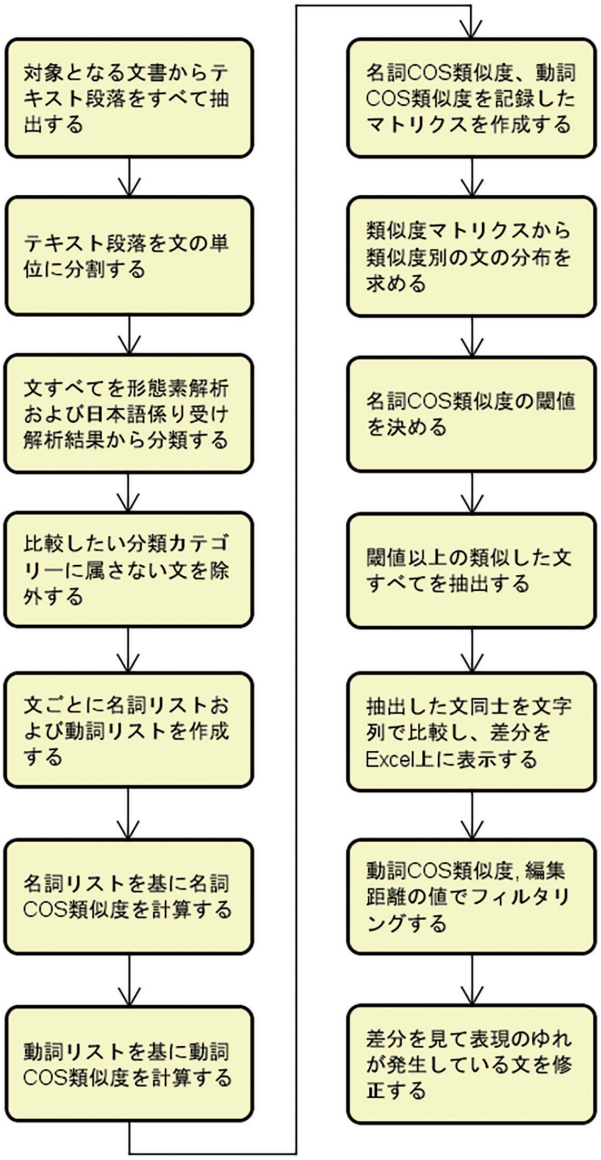


図 1　処理の流れ

## 3.1　対象の絞り込み

### 3.1.1　文に分割

まず文書からテキスト段落を抽出し、比較する単位に分割する。特許明細書のように、一段落中に多くの文を含むような場合は段落ではなく、句点で区切った文単位に分割する必要がある。したがって、段落の終端を表す改行と一文の終端を意味する句点「。」で分割する。

しかし、複数の文をまとめて比較単位とする方が表現のゆれの検出に適切な場合がある。特許ライティングマニュアル [3] の第 A 条に挙げられている連文のように、意味の上で関連性が強い文を連結するキーワードを含んでいる場合、分割せずに連文として処理する。キーワードの例を表 1 に示す。

表 1　文の連結キーワードの例（一部）

| | |
|---|---|
| そのため | このため |
| これにより | それにより |
| そのことにより | このことにより |
| その結果 | この結果 |

### 3.1.2　文の分類

次に文書から抽出した文すべてを 3 つのカテゴリーに分類する。

$U_1$ = { 英数字だけで構成される文の集合 }

$U_2$ = { 述語としての動詞を持たない文の集合 }

例えば「システムに異常が発生」、「ただし標準大気条件下。」のような文がこれに当てはまる。

$U_3$ = { 述語としての動詞を持つ文の集合 }

$U_2$ と $U_3$ の分類をするためには、形態素解析 [4] および日本語係り受け解析 [5] を実行し、結果として述語としての動詞が 1 つも存在しない場合は $U_2$、1 つ以上存在する場合を $U_3$ と判定する。文の表現に差異が生じる可能性が高い文は、述語を持つ集合である。そのため、$U_1$ および $U_2$ に属する文は比較対象から除外する。

このように文を 3 つのカテゴリーに分け、それぞれに属する文の数をそれぞれ $N_1$, $N_2$, $N_3$（$N=N_1+N_2+N_3$）とすると、計算回数は、分類しない場合の $N\times(N-1)/2$ 回よりも $N_1\times N_2\times N_3$（$N_1$, $N_2$, $N_3\neq 0$）回分節約できる。

## 3.2　類似度の計算

本稿では、検証する技術文書として、ある輸送機械の特許明細書を対象とした。さらに参考として輸送機械の保守マニュアルも例とした。双方とも 500 頁を超える分量の情報を集めた。ただし、特許明細書の「特許請求の範囲」は特許特有の文として記載され、一文が長くかつ複雑なため対象外とした。

文 $S_1$ と文 $S_2$ の類似性を測るために COS（コサイン）類似度を計算する。文それぞれを特徴づける構成要素として特に名詞の COS 類似度および動詞の COS 類似度の双方に着目する。

### 3.2.1　名詞リストおよび動詞リストの作成

$U_3$ は述語としての動詞を持つ文の集合なので、文単位の形態素解析および日本語係り受け解析結果から名詞のリストおよび動詞のリストを抽出できる。このとき動詞は原形でリストする。

特許明細書中の以下 2 つの文を例に説明する。

> 燃料噴射量制御は、燃料噴射量を最適に制御する。

を文 $S_1$ とする。

> (1)は、エンジン回転速度と燃料噴射量から燃料噴射時期を算出する。

を文 $S_2$ とする。

このとき文 $S_1$ の名詞リストは〈燃料 | 噴射 | 量 | 制御 | 燃料 | 噴射 | 量 | 最適 | 制御〉となり、動詞リストは〈制御する〉となる。動詞の「制御する」の「制御」はサ変名詞なので名詞リストにも追加する。他方、文 $S_2$ の名詞のリストは、〈エンジン | 回転 | 速度 | 燃料 | 噴射 | 量 | 燃料 | 噴射 | 時期 | 算出〉となり、動詞リストは、〈算出する〉となる。

### 3.2.2　文の間の COS 類似度

名詞の類似度、すなわち名詞をベクトル要素としたときの COS 類似度は、式 1 で求められる。「・」はベクトルの内積を示す。

$$\text{名詞 COS}(S_1, S_2) = \frac{S_1 \text{の名詞ベクトル} \cdot S_2 \text{の名詞ベクトル}}{|S_1 \text{の名詞ベクトル}| \times |S_2 \text{の名詞ベクトル}|}$$

**式 1**

式 1 で求められる COS($S_1$, $S_2$) を「名詞 COS 類似度」と表す。上記例の $S_1$ と $S_2$ の名詞ベクトルは（燃料、噴射、量、制御、最適、エンジン、回転、速度、時期、算出）なので、これらの単語すべてを正規化して式 1 に代入すると名詞 COS 類似度として 0.648 が求められる。

同様に動詞の類似度、すなわち動詞をベクトル要素としたときの COS 類似度は式 2 で求められる。

$$\text{動詞 COS}(S_1, S_2) = \frac{S_1 \text{の動詞ベクトル} \cdot S_2 \text{の動詞ベクトル}}{|S_1 \text{の動詞ベクトル}| \times |S_2 \text{の動詞ベクトル}|}$$

**式 2**

式 2 で求められる COS($S_1$, $S_2$) を「動詞 COS 類似度」と表す。上記の例では、動詞ベクトルは（制御する, 算出する）なので、正規化して式 2 に代入すると、動詞 COS 類似度として 0 が求められる。

### 3.2.3　ベクトルに重み付けをした TF-IDF COS 類似度

特定の技術を詳細に記載する特許明細書などの場合、語句の出現頻度に偏りがある。そこで名詞ベクトル、動詞ベクトルに出現頻度と重要度を考慮した重み付けをする TF-IDF COS 類似度を求め、重み付けのない COS 類似度による検出結果と比較する。

本例の特許明細書を対象にすると、表 2 のような TF-

表 2　名詞の TF-IDF 例

| 名詞 | 名詞の出現数 $n$ | TF ($n$ / $N$) | 名詞が出現した文の数 $s$ | IDF $-\log(s/S)$ | TF × IDF |
|---|---|---|---|---|---|
| 制御 | 387 | 0.4 | 249 | 0.65 | 0.26 |
| エンジン | 265 | 0.28 | 172 | 0.82 | 0.2296 |
| 量 | 234 | 0.24 | 156 | 0.86 | 0.2064 |
| 回転 | 225 | 0.23 | 130 | 0.94 | 0.2162 |
| 速度 | 206 | 0.21 | 134 | 0.92 | 0.1932 |
| 燃料 | 63 | 0.07 | 37 | 1.48 | 0.1036 |
| 噴射 | 56 | 0.06 | 40 | 1.45 | 0.087 |
| 最適 | 17 | 0.02 | 17 | 1.82 | 0.0364 |
| 時期 | 8 | 0.01 | 7 | 2.21 | 0.0221 |
| 算出 | 5 | 0.01 | 4 | 2.45 | 0.0245 |

表３　動詞の TF-IDF 例

| 動詞 | 動詞の出現数 $v$ | TF $(v/V)$ | 動詞が出現した文の数 $s$ | IDF $-\log(s/S)$ | TF×IDF |
|---|---|---|---|---|---|
| 制御する | 69 | 0.2 | 64 | 1.24 | 0.248 |
| 算出する | 5 | 0.01 | 4 | 2.45 | 0.0245 |

表４　TF-IDF COS 類似度マトリクスの例

| 行番号 | 文の内容 | 名詞リスト | 動詞リスト | 160 | 219 | 221 | 232 | 256 | 455 | 523 | 556 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 160 | 燃料噴射量制御は、燃料噴射量を最適に制御する。 | 燃料\|噴射\|量\|制御\|燃料\|噴射\|量\|最適\|制御 | 制御する | NA | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| 219 | 燃料噴射時期制御は、燃料噴射時期を最適に制御する。 | 燃料\|噴射\|時期\|制御\|燃料\|噴射\|時期\|最適\|制御 | 制御する | 0.817 | NA | 0 | 0 | 1 | 0 | 0 | 0 |
| 221 | (1)は、エンジン回転速度と燃料噴射量から燃料噴射時期を算出する。 | エンジン\|回転\|速度\|燃料\|噴射\|量\|燃料\|噴射\|時期\|算出 | 算出する | 0.379 | 0.142 | NA | 0 | 0 | 0 | 0 | 0 |
| 232 | インジェクタのノズル(6)には、常に燃料圧力が加わっている。 | インジェクタ\|ノズル\|燃料\|圧力 | 加わる | 0.144 | 0.176 | 0.117 | NA | 0 | 0 | 0 | 0 |
| 256 | 燃料噴射圧制御は、燃料噴射圧を最適に制御する。 | 燃料\|噴射\|圧\|制御\|燃料\|噴射\|圧\|最適\|制御 | 制御する | 0.594 | 0.722 | 0.101 | 0.127 | NA | 0 | 0 | 0 |
| 455 | NOx センサ(5)は排出ガス中の NOx 濃度を検出する。 | センサ\|排出\|ガス\|中\|濃度\|検出 | 検出する | 0 | 0 | 0 | 0 | 0 | NA | 0 | 0 |
| 523 | 冷却水を循環させ、尿素水を解凍する。 | 冷却\|水\|循環\|尿素\|水\|解凍 | 解凍する\|循環する | 0 | 0 | 0 | 0 | 0 | 0 | NA | 0.65 |
| 556 | このようにして冷却水を循環させ、尿素水を保温する。 | 冷却\|水\|循環\|尿素\|水\|保温 | 保温する\|循環する | 0 | 0 | 0 | 0 | 0 | 0 | 1 | NA |

行番号

右上の橙色のセル部分は、動詞 TF-IDF COS 類似度
左下の黄色のセル部分は、名詞 TF-IDF COS 類似度

IDF 値が得られる。このとき $N$ は特許明細書で検出した名詞の総数を表しており、値は 963 である。$S$ は文の総数を表しており、値は 1,124 である。

同様に動詞の TF-IDF を表３に示す。$V$ は特許明細書で検出した動詞の総数を表しており、338 である。

表３から TF×IDF の重み付けをしたベクトルを式１に代入すると名詞 TF-IDF COS 類似度として 0.379 が求められる。この場合も動詞 TF-IDF COS 類似度は 0 である。

### 3.2.4　TF-IDF COS 類似度マトリクスの生成

COS 類似度や TF-IDF COS 類似度はマトリクスデータとして作成する。TF-IDF COS 類似度マトリクスの例の一部を表４に示す。

マトリクスの作成では、文 $S_1$ と文 $S_2$ のペアごとに名詞 COS 類似度と同時に動詞 COS 類似度の計算も実行する。このとき、名詞 COS 類似度は $N_3\times(N_3-1)/2$ 回の計算を実行するが、動詞 COS 類似度は、名詞 COS 類似度＝0 のときは類似性がないので計算しない。そうすると動詞 COS 類似度を求める計算回数は $N_3\times(N_3-1)/2$ よりも少なくなる。

こうして $N_3\times(N_3-1)$ のマトリクスの領域を半分に分けて名詞類似度および動詞類似度を記録し、効率的にメモリーを使用する。また文の識別のために、集合 $U_3$ のファイルにおける文の行番号を記録する。

### 3.2.5　類似した文の抽出

サンプルとした特許明細書の類似度マトリクスから集計した TF-IDF COS 類似度別の文のペアの分布を表５に示す。

表５の例から、文の数 $N_3$ の組み合わせ全体の約 67% にあたる 0≦COS 類似度＜0.1 は類似性がないと判断できる。

この分布を基に名詞 COS 類似度に関して閾値を設定し、類似度マトリクス中から閾値以上の類似関係にある文の内容を抽出すると同時に文の動詞 COS 類似度も取

表５　TF-IDF COS 類似度別の文のペアの分布

| 名詞 TF-IDF COS 類似度の範囲 | 左記類似度範囲にある文のペア数 | 割合 |
|---|---|---|
| 0 ≦ COS < 0.1 | 422,654 | 66.97 % |
| 0.1 ≦ COS < 0.2 | 40,629 | 6.44 % |
| 0.2 ≦ COS < 0.3 | 49,878 | 7.90 % |
| 0.3 ≦ COS < 0.4 | 42,216 | 6.69 % |
| 0.4 ≦ COS < 0.5 | 32,533 | 5.15 % |
| 0.5 ≦ COS < 0.6 | 21,579 | 3.42 % |
| 0.6 ≦ COS < 0.7 | 12,192 | 1.93 % |
| 0.7 ≦ COS < 0.8 | 5,996 | 0.95 % |
| 0.8 ≦ COS < 0.9 | 2,276 | 0.36 % |
| 0.9 ≦ COS < 1.0 | 994 | 0.16 % |
| COS = 1.0 | 179 | 0.03 % |

得する。

本処理では、COS 類似度マトリクスを作成する時間が最大となる。そのため、このマトリクスを CSV ファイルとして保存し、閾値を任意に変更可能とする。例えば、閾値を 0.7 以上にすると、表から合計 9,445 個の文のペアを抽出できる。この例では、文すべての組み合わせの数が 631,126 個なので、表現のゆれの対象を全体の 1.5% まで絞ることができる。

## 3.3　文の差分表示

### 3.3.1　抽出した文すべての差分表示

上記閾値以上の文のペアについて、名詞 COS 類似度、動詞 COS 類似度とともに編集距離による類似度を Microsoft® Excel® 上に表示する。この Excel への出力では、文書比較／日本語精査モジュール [6] を利用する。3.2.1 のサンプルの文の間の差分表示を表６に示す。

### 3.3.2　類似度でのフィルタリング

Excel のフィルタリング機能を利用し、編集距離による類似度が 100% となる文を除いた文のペアを表示する。その後、名詞 COS 類似度、動詞 COS 類似度および編集距離による類似度の値でフィルタリングする。

# 4　結果と考察

## 4.1　COS 類似度と TF-IDF COS 類似度

### 4.1.1　速度比

COS 類似度計算は R 言語を使って下記 PC 上で求め、そのときの計算速度を実測した。

PC：　Windows10 Pro　RAM：8.00GB

CPU：　Intel(R) Core(TM) i5-6200U CPU @ 2.30GHz 2.40GHZ

処理言語：R 言語 x64 3.4.1

開発環境：RStudio Version 1.0.153

表７に実測値を示す。文書ごとに名詞ベクトル、動詞ベクトルの大きさが異なるため、毎秒当たりの計算回数も文書ごとに異なる。TF-IDF COS 類似度の計算速度は、重み付けの処理のため、本例の２つの文書では COS 類似度の計算速度の５割前後となった。

### 4.1.2　検出した文のペアの分布

名詞 TF-IDF COS 類似度≧0.1 の文のペア数の分布を図２に示し、動詞 TF-IDF COS 類似度≧0.1 の分布を図３に示す。参考のため、重み付けをしない名詞 COS 類似度≧0.1 の文のペア数の分布を図４に示し、動詞 COS 類似度≧0.1 の分布を図５に示す。

表現のゆれの可能性がある名詞 COS 類似度 0.5 以上の文のペア数に着目する。特許明細書、保守マニュアル双方とも図２の名詞 TF-IDF COS 類似度の方が、図

表６　文の間の差分表示の例

| 行番号 | 文の内容１ | 行番号 | 文の内容２ | 名詞 TF-IDF COS 類似度 | 動詞 TF-IDF COS 類似度 | 編集距離に基づく類似度 |
|---|---|---|---|---|---|---|
| 160 | 燃料噴射量制御は、―燃料噴射量を最適に制御する。 | 221 | (1)は、エンジン回転速度と燃料噴射量から燃料噴射時期を算出する。 | 0.379 | 0 | 0.333 |

表７　類似度の計算速度

| 対象文書 | 計算回数（合計） | COS 類似度計算（回/秒） | TF-IDF COS 類似度計算（回/秒） | TF-IDF COS 計算速度 / COS 計算速度 |
|---|---|---|---|---|
| 特許明細書 | 1,262,252 | 912 | 470 | 0.52 |
| 保守マニュアル | 3,777,192 | 643 | 312 | 0.49 |

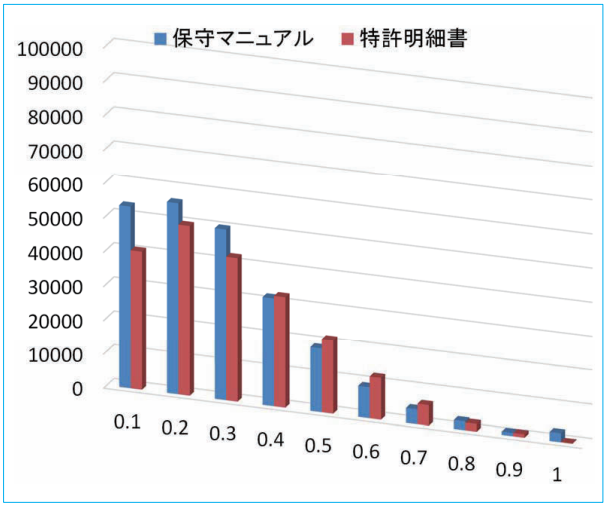


図２　名詞 TF-IDF COS 類似度別の文のペア数

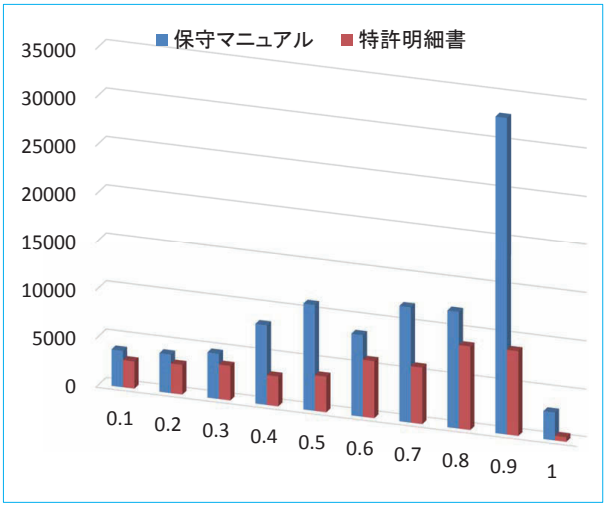


図３　動詞 TF-IDF COS 類似度別の文のペア数

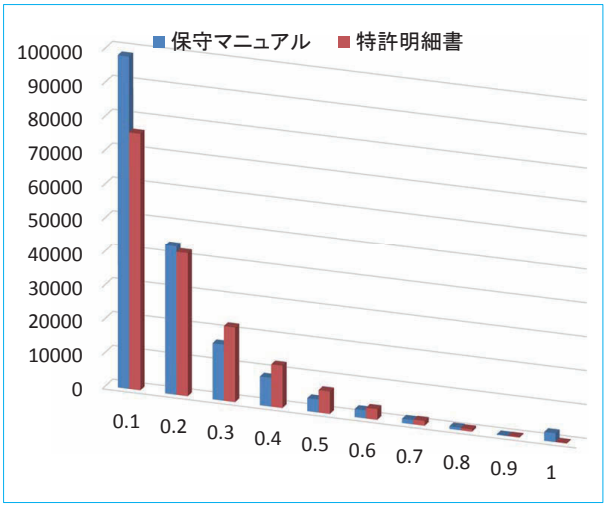


図４　名詞 COS 類似度別の文のペア数

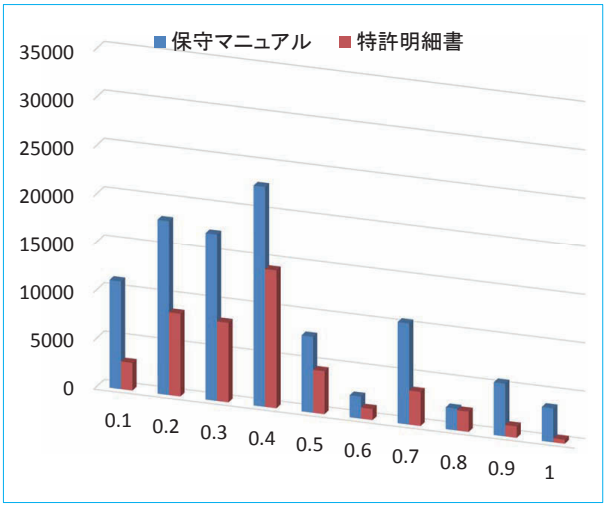


図５　動詞 COS 類似度別の文のペア数

４の重み付けしない名詞 COS 類似度よりも多く検出されている。

一方、動詞 COS 類似度の分布を見ると、図３の特許明細書、保守マニュアル双方ともに類似度 0.5 以上の文のペア数が多い。これに対し、重み付けしない動詞 COS 類似度の図５では類似度 0.5 以上が逆に少なく、対照的な結果となっている。

このように名詞 TF-IDF COS 類似度によって検出される文のペアは、動詞の類似度も高いものが多いことから、類似していることが期待できるため、TF-IDF COS 類似度を基にして類似した文を抽出する。

## 4.2　各類似度から見る表現のゆれ

表現のゆれを測るための指標として、名詞 TF-IDF COS 類似度、動詞 TF-IDF COS 類似度および編集距離による類似度を利用する。これらの組み合わせから表８に示す８パターンが考えられるが、実際にはⅤおよびⅧの組み合わせはない。またⅥおよびⅧの場合、文の間に類似性はないと判断できる。したがってⅠ～Ⅳの４パターンそれぞれから検出できた表現のゆれの特徴を列挙する。これらの差分表示から正誤を判断し、必要があれば原文を修正する。

表８　表現のゆれを測る指標とパターン

| 指標 | 名詞 COS 類似度 | 動詞 COS 類似度 | 編集距離による類似度 | 表現のゆれ可能性 |
|---|---|---|---|---|
| Ⅰ | 高 | 高 | 高 | 有 |
| Ⅱ | 高 | 高 | 低 | 有 |
| Ⅲ | 高 | 低 | 高 | 有 |
| Ⅳ | 高 | 低 | 低 | 有 |
| Ⅴ | 低 | 高 | 高 | ― |
| Ⅵ | 低 | 高 | 低 | 無 |
| Ⅶ | 低 | 低 | 高 | ― |
| Ⅷ | 低 | 低 | 低 | 無 |

### Ⅰ　指標の類似度すべてが高い

すべての類似度が高いため、この場合の文のペアの差異は一部に限られ、誤字や脱字も検出しやすい。検出例を表９に示す。

以下の検出ができた場合、特許ライティングマニュアル[3]の各条の規則を適用し、修正可能である。

・主語、目的語などの省略の検出（第Ｃ条）
・読点箇所の不統一の検出（第Ｅ条）
・非均質並立表現「ばらつき」の検出（第Ｆ条）

### Ⅱ　編集距離による類似度が低い

文字の並びとしての差異が大きい場合である。一例として以下の検出が当てはまる。

・一方が長文（複数の短文）、他方が短文のような文の検出（第Ａ条）

検出例を表 10 に示す。

### Ⅲ　動詞 COS 類似度が低い

動詞が一致していない場合である。一例として以下の検出が当てはまる。

・非均質並立表現「ばらつき」の検出（第Ｆ条）

検出例を表 11 に示す。

### Ⅳ　動詞 COS 類似度、編集距離による類似度ともに低い

動詞が一致していない文であって、文字の並びとしての差異も大きい場合である。この検出例を表 12 に示す。

## 5　おわりに

対象とする文の数が多くなると、計算時間も表現のゆれの候補の数も大きくなる。表現のゆれを「網羅的に、効率的に、精度よく」検出するという課題に対し、まず文章すべてを比較すべき文に分割して分類した。そして述語としての動詞を持つ文同士の名詞 TF-IDF COS 類似度すべてを計算した。さらに名詞 TF-IDF COS 類似度が 0 より大きい文のペアを対象に動詞 TF-IDF COS 類似度を求めた。次に名詞 TF-IDF COS 類似度が一定の閾値以上の文のペアに対し、動詞 TF-IDF COS 類似

表９　類似度すべてが高い例

| | |
|---|---|
| 水温または吸気温が規定値以下であると、装置が、コントロールバルブを開く。 | 水温または吸気温が低いとき、装置が、コントロールバルブを開く |
| 引き操作時に、他の操作条件や作業内容に応じて再生弁を切り替えることにより、ポンプの負荷に応じた最適な動作になるように制御する。 | 再生制御は、引き操作時に、他の操作条件や作業内容に応じて、再生弁を切り替える。これにより、この制御はポンプの負荷に応じた最適な操作になるように制御する。 |
| この結果、ポンプ１とポンプ２の油が合流し、アタッチメントの操作速度が早くなる。 | この結果、ポンプ１と２の油が合流し、アタッチメントの操作速度が速くなる。 |

表 10　編集距離による類似度が低い例

| | |
|---|---|
| 単独操作の結果、ポンプ１と２の油が合流し、装置２の操作速度が速くなる。 | 装置１の単独操作時には、ポンプ１と２の油を合流する。これにより、装置２の動作速度を速くする。 |
| エンジンが停止すると、（１）がリレーを OFF にし、その後、（２）の電源が OFF になる。 | エンジンが停止すると、（１）がリレーを OFF にする。 |

表 11　動詞 COS 類似度が低い例

| | |
|---|---|
| 弁の（１）が下方向に押され、ポンプからの油がバルブに流れる。 | 弁（１）が ON になり、ポンプからの油をバルブに供給する。 |
| 操作レバーを戻すと制御弁内にある流量制御弁が戻り、コントロール圧 P が下がる。 | 操作レバーを操作すると制御弁内にある流量制御弁が切り替わり、コントロール圧 P が上がる。 |

表 12　動詞 COS 類似度、編集距離による類似度が低い例

| | |
|---|---|
| 下記条件をすべて満たしたとき、コントローラ(2)は、モニタコントローラに信号を送る。 | 下記条件が一定時間継続すると、コントローラは、通信により、モニタコントローラに信号を送信する。 |
| コントローラが、ドアが開いていると認識し、内部で A 端子を 30 秒間アースに接続する。 | コントローラが、ドアが閉じていると認識し、内部で A 端子のアース接続を遮断する。この結果、ライトは点灯しない。 |

度または編集距離による類似度の値でフィルタリングした。そうすると、表現のゆれに特徴が生じる。こうしてフィルタリングによって絞り込んだ文のペアの差分表示から表現のゆれの修正の要不要を判定できる。

今後、文の類似度以外の情報も合わせて、表現のゆれだけでなく、日本語の誤りの検出精度を向上させ、検出された文の修正を支援する方法を提供する予定である。

## 参考文献


[1] 山本和英：“日本語の表記ゆれ問題に関する考察と対処”
Japio YEAR BOOK 2015, pp. 202-205, 2015.

[2] 楠本浩二：“文書比較結果の多目的利用”
Japio YEAR BOOK 2014, pp. 306-313, 2014.

[3] 一般財団法人日本特許情報機構　特許情報研究所
特許ライティングマニュアル「産業日本語」初版

[4] 形態素解析エンジン MeCab
http://taku910.github.io/mecab/

[5] 日本語係り受け解析器 CaboCha
http://taku910.github.io/cabocha/

[6] 文書比較／日本語精査モジュール やまと歌
http://www.crestec.co.jp/yamatouta/