2J-3

誤りやすい英単語を認識させる手法

菅野啓*，金子美和**，青木和夫***
日本アイ・ビー・エム株式会社　ソフトウェア開発研究所

1. はじめに

現在，英文を作成する際の補助ツールとして，翻訳ソフト，英文スペルチェッカー，文法チェッカー，電子辞書がある．しかし，翻訳ソフトは，今の訳文の品質そのままでは一般業務には使用が難しい．英文スペルチェッカーは，単語の使用誤りは検出しない．文法チェッカーは，活用形などの単純な使用誤りしかチェックしない．電子辞書は，曖昧な語がある時に意識的に引くものであるので，誤使用を見つけてくれない．

英文を作成した後で，私たちは通常スペルチェッカーで英単語の綴り誤りをチェックする．しかし，綴りが正しい場合は何も処理されない．それは既存のスペルチェッカーは，辞書引き後の未知語に対してのみを処理しているからである[1]．筆者は，未知語以外の単語を対象にして，誤りやすい英単語に母国語を対訳する事で，直感的に誤使用を発見し易くさせる手法を実験した．本稿では，その手法の基礎となる「誤りやすい語」の定義とその抽出アルゴリズムを 2 章で，「誤りやすい語」の辞書について 3 章で，評価結果を 4 章で説明する．

2. 誤りやすい語の定義と抽出アルゴリズム

「誤りやすい語」とは何かを明確にするために，実際の使用誤りを調査した．以下は英文メールの中での誤使用の例である．(括弧の中が正しい)

- *The words are extracted from the whole content. (context)*
- *If you have a compliance about a software…　(complaint)*
- *The hotel serves complementary services: (complimentary)*
- *The register on the planar should be changed. (resister)*

これらの例から，「誤りやすい語」の特性の1つは，語の綴りや発音が似ていることではないかと推測できた．そこで"Common Errors in English"に示されている誤りやすい単語の組を分析した[2]．その結果，全ての組が綴りや発音が似ていることが分かった．また他の要因としては，経験則から使用頻度が少ないイディオムや単語の誤使用，語義の記憶違いによる誤使用等がある．今回は，「誤りやすい語」を，綴りや発音が「似ている語」と，非頻出のイディオムと単語として定義し検討した．図１の①は全単語を似ている/似ていないで分類した図で，点線の文書図は通常の英語文章からのビューを示す．似ている/似ていないの閾値は，定値でなく各人の知識･経験･性格などにより揺れる．②は更に「似ていない語」を頻出/非頻出で分類した．この閾値も，定値でなく各人の知識･経験･性格などにより揺れる．結果として，文書ビューは頻出語を包含する位置に移動される．③の文書ビューの中の「似ている語」と非頻出語を合わせた斜線の部分が，私たちが「誤りやすい語」と見なす候補になり得る．

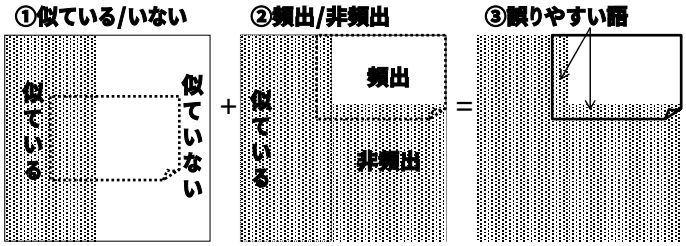

図 1. 誤りやすい語の定義

図 2 は，英文から「誤りやすい語」の候補を抽出する実行時のアルゴリズムである．与えられた英文を形態素解析し，未知語を除いた単語に対して，「誤りやすい語」辞書に定義した単語と照合して対訳付きで抽出する．ここで抽出されなかった単語に対して，ストップワードでフィルタリングした非頻出のイディオムと単語を英日辞書で対訳を付けて抽出する．

1. 英文を入力する
2. 形態素解析を行い，語にストップワードや未知語などの属性を付ける．
3. 未知語以外の全ての語に対して 4.と 5.を繰り返す．(未知語は従来のスペルチェックの処理対象)
4. 「誤りやすい語」辞書引きし，綴りや発音が「似ている語」を対訳付きで抽出する
5. ストップワードでフィルタリングし，非頻出のイディオムと単語だけを英日辞書引きし対訳を付ける
6. 対訳付きの語だけを出力する

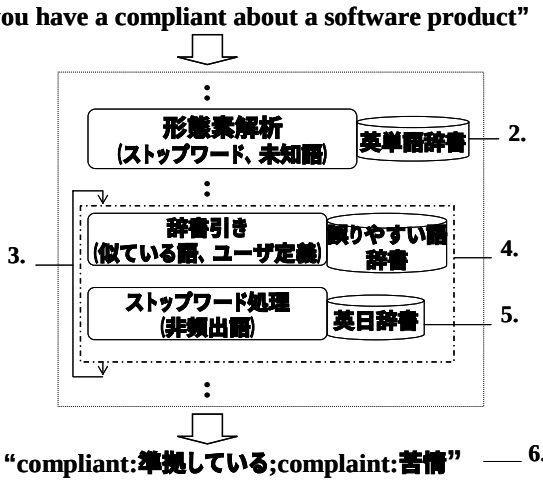

図 2. 抽出アルゴリズム

この抽出アルゴリズムは，構文解析や意味解析や文脈解析といった複雑な解析を必要としないので，実装が比較的簡単でパフォーマンスの良いものになっている．

Method for non-native English users to detect misusage of English words
* Kei Sugano　** Miwa Kaneko　*** Kazuo Aoki
Software Development Laboratory – Yamato (YSL), IBM Japan, Ltd.

3.「誤りやすい語」の辞書

「誤りやすい語」辞書には,「似ている語」候補とその対訳を登録する. 辞書はカスタマイズが可能で, ユーザが語の追加･削除･修正が出来る. 以下は辞書の構造である.

＜似ている語のレコード形式＞

見出し語:訳語;分類;似ている語:訳語

例) compliant:準拠している;RST:complaint:苦情,苦痛

＜誤りやすい語のレコード形式(ユーザ定義)＞

見出し語:訳語;分類(;似ている語:訳語)

例) convene:会合する;U

「分類」には, 見出し語が,「綴りが似ている語」,「発音が似ている語」,「ユーザが登録した語」等を指定する.

「似ている語」は, 形態素解析用の英単語辞書から規則で抽出する. この規則は, "Common Errors in English"の誤りやすい212 組の分析を行なって定義した. 2つのストリングの類似度の距離を計算する方法は, 昔から良く研究されており, Edit distance, Longest Common Subsequence distance, Hamming distance などがある[3]. 簡単で解り易い Hamming distance 手法を用いて, 単語間の距離と類似度を定義した.

$$\text{距離 } d(x,y)=\sum_{i=0}^{n} a(x_i,y_i) \quad a(x,y)=\begin{cases}0 & (x=y)\\ 1 & (x\neq y)\end{cases} \quad \text{類似度 } 1-\frac{d(x,y)}{n}$$

表 1 が分析した結果であり, 類似度が 0.5 以上の組は 201 組で全体の 94.8%カバーしている(①を参照). 残りの 11 組 (accede/exceed, bare/bear, cite/sight など) は全て発音記号の類似度が 0.5 以上であった.

表 1. 212 組の分析

類似度	距離 1	距離 2	距離 3	距離 4	距離 5	
0.0 - 0.09	0	0	0	0	1	
0.1 - 0.19	0	0	0	0	0	
0.2 - 0.29	0	0	1	2	0	
0.3 - 0.39	0	1	0	1	1	
0.4 - 0.49	0	0	1	3	0	
0.5 - 0.59	1	10	6	2	1	①
0.6 - 0.69	2	28	4	0	0	
0.7 - 0.79	18	30	6	0	0	
0.8 - 0.89	82	6	0	0	0	
0.9 - 0.99	5	0	0	0	0	
total	108	75	18	8	3	

以下は, 類似度の算出方法である.

- 単語長が同じ場合:
  (例) adapt/adopt (類似度 0.8)
- 単語長が異なる場合:開始位置を合わせて距離を測る
  - *先頭の文字が一致する場合 : 先頭から測る*
    (例) continual/continuous (類似度 0.7)
  - *先頭の文字が一致しないで, 最後尾の文字が一致する場合 : 最後尾から測る*
    (例) aural/oral (類似度 0.6)

注) 分析した 212 組には, 先頭の文字も最後尾の文字も一致しない組は無かった.

この規則を英単語辞書に適用し, 比較検討した結果, 単語の長さに応じて,「似ている語」とみなす類似度の閾値を変更する必要があった. 例えば, 閾値を低く設定した場合では, compliance の候補として/complete /compline/confluence/ compliant/comedienne/…と多く抽出される. このような接辞が同じ単語(comp-,-ness など) を除く為に, 閾値を高く設定してしまうと, 4 文字長程度の単語には似ている単語が存在しなくなる. このため, 4 文字長以下の単語は先頭が 2 文字以上または先頭と末尾の両方が一致するものを抽出することとした. これらの規則を英単語辞書に適応したところ, 7 万単語中 3 万語が「綴りが似ている語」として抽出された.

4. シミュレーションによる評価

2章のアルゴリズムや3章の辞書の有効性を検証するために, 実際の英文でシミュレーションを行った. その結果は, 標準的なストップワードでは, 明らかに誤りそうもない単語が「誤りやすい語」の候補となってしまった. そこで, ストップワードの見直しを行い, 最終的に以下の単語もストップワードに分類した.

- *特定の品詞(名詞, 形容詞, 動詞以外)の全ての単語*
- *固有名詞*
- *カタカナ訳される語*
- *頻出語*

図 3 はストップワード見直し後の結果である. 全体の 64 単語中「誤りやすい語」の対訳表示の候補が延べ5単語(8%)に絞り込まれた. 5 単語の内訳は, "compliant (complaint が正解)"と"complimentary (complimentary)" は「似ている語」で, "supervise"と"abide by"は非頻出の単語とイディオムである.

If you have a compliant (準拠している;) about a software product or other services, the ABC System might be able to help you. The ABC System has experiences to handle a compliant of consumer software. The Board of ABC System, located in Tokyo, works with the twelve consumer groups which supervise (監督する; 管理する) abide by (に従う) the consumer laws. We can help individual consumers by the following complimentary (賞賛の; 無料の) services:

図 3. 最適化後のシミュレーション

他の英語の文章でもシミュレーションを行った結果, 平均 431 単語中 54 語(12%)の誤り候補数となり良い結果を示した.

5. まとめ

英語を母国語としない人にとって, 英作文で「誤りやすい語」とは何かを定義して「誤りやすい語」を抽出するアルゴリズムを提案した. そのアルゴリズムの中の「誤りやすい語」辞書を説明し, 綴りが「似ている語」の抽出方法を提案した. シミュレーションを行って評価しストップワードの最適化を行なった結果は良好であった.

参考文献

[1] Fred J. Damerau, “A technique for Computer Detection and Correction of Spelling Errors”, Communications of the ACM, Volume 7 Issue 3 (1964).

[2] Paul Brians, “Common Errors in English”, http://www.wsu.edu:8080/~brians/errors/errors.html

[3] Gonzalo Navarro, “A guided tour to approximate string matching”, ACM Computing Surveys, Volume 33 Issue 1 (2001).