自然言語処理 90－6
（1992．7．17）

文中における語句の『近さ』について

脇田早紀子、奥村薫、金子宏
日本アイ・ビー・エム(株)東京基礎研究所

「校正支援システム」を実務で使う場合は、ユーザー自身の手で校正知識を表現できるようにしておくことが大切である。その際、直感に沿う形で簡単に語句の『近さ』を表現できれば、「“再び”と“～し直す”が『近く』に出てくると重言」「“ように”と“ない”が『近く』に出てくるとあいまい」などと幅広く利用できて便利である。本研究では、『近さ』判定アルゴリズムを提案した後、助詞の『近さ』を題材にしてその有効性を確かめた。

"Distance" between two words in a sentence

Sakiko Wakita, Kaoru Okumura, Hiroshi Kaneko
IBM Research,
Tokyo Research Laboratory

A Knowledge of critiquing often takes a form like this; 'if a word A is 'near' a word B, then …'. In this paper, we offerr a simple way to discribe 'distance' between two words. Users of critiquing system can write down their knowledge of critiquing themselves in this way.

1　はじめに

現在、新聞社の実務に役立つ校正支援を目標に「日本語校正支援システム　ＦｌｅＣＳ」を開発中である。対象を新聞記事に絞って考えると、表記の基準がはっきりしていること、文体に特徴があり捉えやすいことなど、比較的実用化しやすくなる。我々はこれまで、校正知識の記述を効率よく行う方法について研究してきた。

日本語の文章の中から修正すべき箇所を捜す方法には、

- 解析に失敗した箇所を誤りとして警告する。
- 複数の解釈が成り立つ箇所を警告する。

などがある。ただし、それぞれ

- 誤った表記・読みにくい文などでも何らかの解釈はできてしまうことが多い。
- よほど正確な（しかも柔軟な）意味情報を使わない限り、たいてい複数の解釈がのこってしまう。

という問題があるため、深い解析をしても確かな結果を得るのは難しい。

そこで我々が開発している「ＦｌｅＣＳ」では、形態素解析を行うほか、誤りの可能性が高い文の特徴を登録しておいて発見させるという方法を採っている。[6)]

このような知識を蓄えていこうとすると、「ある２つの単語が（感覚的に）『近い』」とき警告を出したいということがよくあるのに気付く。例えば以下のような文である。

・同じ助詞が近くに出てくると、重なっていて読みにくい。

> 次に顕著なのは、これは、生成文法の初期から関心が持たれていたことだが、…(略)。
>
> 経企庁もベテンにかけられたというのですが、こうした役人のいい加減な国民無視の行政態度の責任はきびしく追及されてしかるべきだと思いますが、…(略)。[1)]
>
> 私は小林が中村が鈴木が死んだ現場にいたと証言したのかと思った。[1)]

図１－１　読みにくい

・「数詞(＋数詞接尾辞)＋を(格助詞)」と「越す」が近くに出てくると、「超す」のまちがい。

> 五万人を越す人出となった。

図１－２　「超す」のまちがい

・「ように」と「ない」が近くに出てくると、あいまい。

> ローマ字の場合には、フランス語のようにスペルと発音のへだたりがないから問題はない。[2)]

図１－３　あいまい

・「再び」と「～し直す」が近くに出てくると、重言。

> 再び検討し直した。

図１－４　重言

このような文の検出は係り受け解析などでもできるはずだが、本研究ではより直感的な『近い』という概念を導入して検出を試みた。この方法には、

- 処理が簡単。
- わかりやすい。校正支援システムのユーザー自身で校正ルールを書ける。メンテナンスもしやすい。
- 多様な対象を扱える。

などの利点がある。

2　目的

本研究の目的は、記述と処理の両方が簡単な方法で、直感的な『近さ』を判定するアルゴリズムを与えることである。

実際の場面で校正システムを快適に使用するためには、

- 検出率が高い
- 過検出率が低い
- 使いやすい
- 処理が速い

ことが重要になる。

3　『近さ』判定アルゴリズム

文は一本調子で流れているのではなく、いく筋かの流れがすっとまとまるような部分がところどころにあるものだ。

昼過ぎから雨に なった ので、明日のことが心配だ。

図３－１　文のまとまり

この例では、「昼過ぎから」という句の落ち着く先を探しているとすぐに「なった」で見つかり、「ので、」と一段落ついたときにはもうすっきりしている、自然な文である。

野次馬根性からノーベル賞を 授与された 日本人に ついて、文化勲章授与とのタイムラグを調べてみた。[3)]

図３－２　文のまとまり

一方、この例では「野次馬根性から」という句を抱えてその落ち着く先を探していると、うっかり「授与された」につなげてしまいそうになるが、うまくいかない。「ついて、」と大きな切れ目にさしかかってもまだ抱えたままなのでどうも落ち着かない。結局、最後の「調べてみた」まで持っていくことになるが、「野次馬根性から」という句は、そんなに『遠い』道のりを抱えていくには不向きなのだ。

今、「野次馬根性から」と「調べてみた」が『遠い』という言葉を使った。本研究では、人がなんとなく感じるこの『近さ』『遠さ』を、語と語の間に『区切り＝抱えているものを下ろしたくなるところ』があるかどうかで判定する方法をとった。どのくらいなら『近い』と感じるかは語によって当然違ってくるので、区切りのほうも、切れ方の強さに応じて数レベルに分類しておく。

一般的な区切りとして、４レベルの『区切り』を導入した。切れ方の強い順に、文区切り・引用区切り・重文区切り・述語区切りである。

文区切り ： 句点
　　空行
　　行頭のスペース

最も大きな区切り。文は普通「。」で切れるが、「。」がない部分（箇条書きなど）のために少し付け足した。

引用区切り ： 括弧　「」『』（）
　　箇条書きの頭になる部分　①②など

重文区切り ： 連用中止
　　接続詞，接続助詞

引用区切り・重文区切りは文の中に文があるようなものなので当然といえる。

述語区切り ： 動詞のうち体言化していないもの
　　形容詞
　　形容動詞
　　断定の助動詞
　　名詞＋読点

述語区切りは、正確にいうと述語となる可能性のある品詞列である。別の言い方をすれば種々の語句の係り先候補である。この区切りを用いると例えば

彼女がバラが好きだと言った。
　　↑
　　└── 述語区切り

図３－３　間に述語区切りがない

のような入れ子の文の「が」と「が」は近く、

バラが好きだと彼女が言った。
　　↑
　　└── 述語区切り

図３－４　間に述語区切りがある

のように入れ子をはずした文の「が」と「が」は遠いと判定できる。

「名詞＋、」を述語候補に入れるかどうかは難しい問題だが、新聞の文体を考慮すると、これを区切りに含めない場合は過検出が多くなると考えられるので含めることにした。

４　検証の方法

上述のアルゴリズムの有効性を検証するため、以下のような実験を行った。

材料：新聞記事（産経新聞）政治面
　　約1Mbyte（５０万文字分）。
　校正・校閲前のものも交じっているので、今回の目的に適している。

抽出ツール：ＦｌｅＣＳ
　ＦｌｅＣＳは本来、「日本語校正支援システム」である。しかし、校正用ルールの代わりに、字面または品詞列を用いて検索条件を記述すれば、その条件に適合する部分を抽出するツールとして使用できる。

対象：同じ助詞が重ねて出てくる文。

は…は(係助詞)
が…が(接続助詞)
が…が(格助詞)
に…に(格助詞)
を…を(格助詞)

具体的には、同じ助詞が二度『近く』にでてくる箇所を発見する。ここで『近い』というのは、前のものの影響が十分残っているうちに次のものが出てくることを指す。「十分残っているうち」かどうかは、たいていの人が受ける「感じ」で決める。

助詞が『近い』という状況の中には、強調し過ぎてくどい場合、助詞の使い方を間違えたため(たまたま)重なっている場合、入れ子の構文になっている場合などが交じっている。いずれにせよ読みにくい文である可能性が高いので検出したい。

作業は以下のように行う。

(1) 『近さ』の条件を設定する。
(2) ＦｌｅＣＳで抽出する。
(3) 検索されたものの中に、近くないものが交じっているかを調べる。
(4) 交じっていれば(1)に戻る。

以上に述べた作業が終わった後、文中に同じ助詞が出ているが『遠い』と判定されたものから100例程度に目を通し、検出されるべきものを落としていないか調べた。

疑わしいものをもれなく検出することより、余計な警告が少なくなることの方にやや重点を置いて条件を調整した。これは、見かけの精度（正警告数／全警告数）が使い勝手に最もクリティカルであるという経験上の事実に基づく。

5 結果

テキスト1Mbyteの解析にかかった時間は２時間弱であった。使用したマシンはＩＢＭパーソナルシステム／５５*)（ＰＳ／５５７１Ｖモデル）である。

各種区切りを利用して『近さ』条件を設定すると、修正を必要とする箇所の候補を劇的に減らせる。その様子を表１に示した。たとえば格助詞の「が」が重複している文を見つけたいとき、二千以上の文から探さなくてはいけなかったら（←文区切りのみで抽出した場合）見る気もしないが、数十に絞ってあれば（←述語区切りまで全て使って抽出した場合）調べる気になるだろう。

表１　『近さ』による絞り込み効果

対象	使用した区切り			
	文	引用	重文	述語
が…が(接助)	22	6	-	-
は…は(係助)	2982	1091	347	-
が…が(格助)	2089	1097	372	43
に…に(格助)	4048	2276	954	322
を…を(格助)	5404	3174	1293	140

以下、個々の対象について検討する。

5．1 が…が(接続助詞)

文区切り・引用区切りにより『近い』と判定された６文の中には、このようなものがあった。

> 調査会は一国平和主義を批判しているが、一国だろうが二国だろうが、（憲法の平和主義の精神を）世界に訴えていくべきだ。

図５－１　が(接助)：近いが気にならない

「雨が降ろうが槍が降ろうが」式のこのような文は「が」が近くても読みにくくない。そこで、

「が(接助)」特別修正
> 未然形に接続する「が」は除く

という変更を加えると、抽出されるのは以下の２文だけになった。

*) パーソナルシステム／５５はＩＢＭの登録商標です。

焦点となるのは具体的な人選だが、最終的には、宮沢首相と、党役員会で証人喚問問題への対応を一任された綿貫幹事長ら執行部の判断に委ねられることになりそうだが、週明けの十七日から予算委審議が再開できるかどうか予断を許さない状況になっている。

結果について共和から報告はなかったが、一カ月共和の森口副社長か大川常務かはっきりしないが、議員会館にきて…(略)

図５－２　が(接助)：近くて気になる

これらは、つながりのはっきりしない「が」でだらだらと続いていて読みにくい。[1]一方、遠いと判定された文を全て調べたところ、このような読みにくさを持つ文はなかった。

①平成元年十一月ごろ、阿部文男元長官から、共和が計画していた会員制レジャークラブ「麹町倶楽部」の発起人代表になることを依頼されたが、はっきり断った②平成二年四月ごろ、再び阿部元長官を通じて名誉理事長への就任を依頼され、即答を避けたが、その後、「立派に完成したあかつきに、理事会で満場一致で推薦を受けることを見極める」ことを条件に内諾を与えたーと一連の経過を説明した。

図５－３　が(接助)：遠くて気にならない

５．２　が…が(格助詞)

文区切り・引用区切り・重文区切り・述語区切りにより『近い』と判定された43文の中には、助詞の使い方が不適切なもの、例えば

本来ならアメリカが経済が国際競争力においても一番になってもらわなければならない。

図５－４　が(格助)：近くて気になる

と、構文が入れ子になっているもの、例えば

それがロシアが主張する「法と正義」にかなう最低線との考えを前面に押し出す構え。

一方、宮沢首相をはじめ官邸サイドは、喚問などの対象者が宮沢派関係者が中心だけに対応に苦慮している。

図５－５　が(格助)：近くて気になる

のように、修正を要する文・書き替えると読みやすくなる文が多く見つかった。

一方、入れ子になっていてもあまり気にならないものもある。

それにしても、外交に強いはずの首相が、米側の反日感情が強い中でなぜそのリアクションを予想できなかったのか、首をかしげる向きは多い。

図５－６　が(格助)：近いがあまり気にならない

あえて入れ子の文を書くときは「が」のあとに「、」を打てとよくいわれる。「が」のあとに「、」が打ってあれば検出しないことにしてもよいと考えられる。

また、36文中２文、述語区切りの趣旨からいうと切れるべきなのに切れなかったことがあった。

また夜勤が月八回以上の病院が六割以上（平成二年）を占め、週休二日制の病院は四％。

「宮沢派の人は他のムラ（派閥）の人達に比べ政治改革に対する危機感が希薄のような気がする」との思いは“加藤ルート”を通しての感触か？

図５－７　が(格助)：遠いはずだが

これらの場合、「～が～の」は「～が～である」と同じ意味で用いられている。

以上の考察をもとに、

「が(格助)」特別修正

１つめの「が」が「が、」となっているものは除く。

「が(格助)」特別区切り

格助詞「が」の影響は「の」で切れる。

と変更した。これにより24例に絞ることができて、過検出はなくなった。今回のテキストの場合、この修正によって検出すべきものがされなくなってしまうことはなかった。

一方、遠いと判断されたものから100例調べたが、検出し損ねているものは１例だけだった。

さらに、野党側が阿部代議士の秘書や共和関係者、さらに金銭授受が報道されている鈴木元首相、塩崎元総務庁長官らの証人喚問を迫ることも予想され、自民党は対応に苦慮しそうだ。

図５－８　が(格助)：未検出例

これは、「名詞＋、」を述語候補扱いしたため述語区切りが挟まることになり、『遠い』と判断されたものである（この文の場合は述語区切りがなくても「が」の特殊区切りでやっぱり切れてしまうが）。「名詞＋、」を述語区切りから除くと、正検出24に過検出が７交じってしまうので、このままにしておくことにした。

５．３　を…を(格助詞)

文区切り・引用区切り・重文区切り・述語区切りにより『近い』と判定された140文を見てみると、過検出の割合が多かった。

最初に質問に立った社会党の山花貞夫書記長は外交・防衛問題や政治倫理問題を中心に宮沢首相の政治姿勢を追及。

外務省内には「国連創設五十周年に当たる九五年をめどに常任理事国入りをめざす」（波多野国連大使）との声もあるが、…(略)

図５－９　を(格助)：遠いはずだが

これらの場合、「～を～に」は「～を～にして」と同じ意味で用いられている。これは述語区切りの意図からすれば切るべき部分である。

(略)…総裁を初め党四役はすべてわれわれを支えてくれた。

図５－１０　を(格助)：遠いはずだが

この場合も、「～を」の影響は「初め」で切れている。

また、入れ子でも「を、」になっていると気にならないのは「が(格助)」の場合と同様である。

(略)…当初から「そう問題にならない」（国対幹部）塩崎、鈴木両氏という「根幹ではない」人物の喚問・招致を、阿部代議士喚問を見送る“取引材料”としたふしもある。

図５－１１　を(格助)：近いがあまり気にならない

そこで、

「を(格助)」特別修正

１つめの「を」が「を、」となっているものは除く

「を(格助)」特別区切り

格助詞「を」の影響は「に」で切れる。
「を初(始)め」のときもそこで影響が切れる。

と変更すると過検出はなくなり、29文まで絞られた。除かれた文を調べたところ、「特別区切り」の悪影響はなかった。

残ったものには、文法的におかしい場合、

金銭の授受については「進学の相談や合否を早く結果を出してほしいという場合もあり、浄財をいただいたこともあるかもしれないが、…(略)

図５－１２　を(格助)：近くて気になる

タイプミスの場合、

(略)…と述べ、野党結束を最優先ことを強調した。

図５－１３　を(格助)：近くて気になる

入れ子の場合、

同長官はまた、こうした日本側の認識をすでに外交ルートを通じ先方に説明していることを明らかにした。

図５－１４　を(格助)：近くて気になる

などがある。

５．４　に…に(格助詞)

文区切り・引用区切り・重文区切り・述語候補区切りを用いて『近い』ものを抽出したところ、322文になったが、実際に読みにくい文の割合は少ない。

景気浮揚のためには四年度予算案をいつまでも人質にしておくわけにはいかない事情も…(略)

野党にも国民の疑問に答える責任があることを自覚すべきではないか。

図５－１５　に(格助)：近いが気にならない

入れ子の構文にはなっていても、「には」「にも」「に、」が入れ子の始まりを予想させるため、読みにくくならない。

五八年に安保理の非常任理事国に選出され…(略)

図５－１６　に(格助)：近いが気にならない

また、「～日に」「～月までに」など時を表す「～に」の場合も、ほかの「～に」と競合しにくい。

そこで、

「に(格助)」特別修正

名詞と接尾語（数詞接尾を除く）につく「に」のみを検索する。
「には」「にも」「に、」は対象から除く。

と変更すると、97文に絞ることができた。

すると読みにくい文が見つけやすくなる。

(略)…既存の体制へのチャレンジに日本がその先頭に立つことがいいのか。

一方、米沢氏は、政府が今秋から始めるプルトニウム輸送に、現在、建造中の海上保安庁の巡視船の代わりに海上自衛隊の護衛艦を護衛にあてるよう要求。

図５－１７　に(格助)：近くて少しおかしい

その他、一つの動詞に二つの「～に」が係るもの

最初に質問に立った社会党の山花貞夫書記長は…(略)

図５－１８　に(格助)：近いがおかしくない

と、入れ子の構文のもの

政府、国民の間に世界平和のために特別な責任と義務を負う覚悟が生まれ、…(略)

図５－１９　に(格助)：近い；入れ子

とがある。前者は全く修正する必要がないし、後者も「～を」「～が」の入れ子に比べれば自然なものが多い。

以上のように、「～に」の場合は過検出率が高くなるので、厳密性を要求されるときには校正ルールとして不適切かもしれない。

書き替えて読みやすくなるものは半分程度だった。

５．５　は…は(係助詞)

文区切り・引用区切り・重文区切りを用いて『近い』ものを抽出したところ、347文になったが、読みにくい文の割合は少ない。

(略)…本格支援は現段階では難しいことを強調した。

図５－２０　は(係助)：近いが読みやすい

「では」「には」「については」「とはいえ」などの「は」は他の「は」と競合しにくいと考え、

「は(係助)」特別修正

名詞と接尾語（数詞接尾を除く）につく「は」のみを検索する。

のように変更すると、70例に絞ることができた。

また、斎藤氏は日本政府がこれまで北方四島は日本固有の領土である根拠として主張してきた一八五五年の「日露通好条約」に触れ、…(略)

図５－２１　は(係助)：近くて少し気になる

ただしまだ書き直しを必要としない文がほとんどだった。「は」は日本語の中でもとりわけ興味深い（難しい）題材で、これまでにも多くの研究がある。「は」がどう使われていると読みにくいかなどということも、一筋なわではいかないのだろう。

６　まとめ

区切りを用いた『近さ』判定アルゴリズムの妥当性を検討するため、同じ助詞が重なっている文の検出を行った。その結果、このアルゴリズムを用いると、

- 『近い』ものの検出は実用になる精度で行える。
- 若干の条件を加えることによって、「が～が(接助)」「が～が(格助)」「を～を」など有効な校正ルールを簡単に作成できる。

ことがわかった。

7　今後の発展

この『近さ』判定アルゴリズムを用いて二重否定・重言なども検出できる。また、かな漢字変換の誤りを発見するとき共起の範囲を限定するために利用するなど幅広い応用が考えられる。

8　おわりに

２つの単語が『近く』にある箇所を発見する目的で、４種の区切りを導入し、発見対象に応じたレベルで『近い』ものを抽出することに成功した。この方法を用いれば、直感に沿った形で簡単に校正知識を表現することができる。

修正個所の候補をこの『近さ』条件のみで十分絞れるものもあり、個別に条件を付け加えることで実用になる量まで絞ることのできるものもあった。

謝辞

本研究を進めるにあたって、産經新聞社の製作局システム管制部および校閲センターの方々には、新聞記事データを提供していただくなど大変なご協力を頂きました。ここに感謝の意を表します。

参考文献

1) 本多勝一著「日本語の作文技術」朝日文庫
2) 岩淵悦太郎編著「第三版　悪文」日本評論社
3) 千早耿一郎著「悪文の構造」木耳社
4) 牛島ほか：日本語文章推敲支援ツールのプロトタイピング；コンピュータソフトウェア Vol3-1,pp.35-46 (1986)
5) 鈴木・武田：日本語文書校正支援システムの設計と評価；情報処理学会論文誌 Vol30,No.11,pp.1402-1412 (1989)
6) 奥村ほか：日本語校正支援システム「ＦｌｅＣＳ」；92-NL-87,情処 自然言語処理研究会(1992)