文節の順序による可読性向上手法の提案

情報科学科　原 源記　　指導教員：山村 毅

1　はじめに

現代社会では，Facebook や LINE のような SNS や Twitter といったコンテンツを利用する人が多くなった．そのため，短い文を書くことや，砕けた文を書いてしまうことが多く，長い文を書く機会が少なくなっていると思われる．それによって，物事を他人に伝える際に，誤った表現や分かりにくい表現を使ってしまうことが考えられる．

そこで，本研究では，分かり難い文と理解しやすい文の特徴について考え，文の可読性を向上させるルールにはどのようなものがあるか調査する．また，“文節の順序” に焦点を置き，機械学習によって文の可読性の定量化を行い，その指標値を利用して，その文が理解しやすい文であるか否かの判定を行うシステムを考案することを目的としている．

2　関連研究

松尾 [1] は，文の分かりやすさの尺度について，①文節の係り受けの距離，②なじみのある語順，という 2 つの尺度を挙げている．そこで，尺度①については，係り受け距離の合計を最小にした例文とそれ以外の例文のどちらが分かりやすいかというアンケートを，尺度②については，「いつ」「どこで」「誰（何）が」「何を」という部分の並べ替えを用いて，どの順番が最も分かりやすい例文であるかというアンケートをそれぞれ実施した．その結果，尺度①については，分かりやすいと支持した割合は 24.2% に留まる一方で，主語が先頭であるときに分かりやすいと支持した割合は，全体で 84.5% と高く，これから，係り受け距離の最小化は文章の改善法としての影響力はあるものの，決定的な要因ではないということを示した．また，尺度②については，「誰（何）が」「いつ」「どこで」「何を」（「どうした」）という語順が最も馴染みのある語順であるということを示した．

3　文節タグ

本研究では “文節の順序” に焦点を置き，文節の並べ替えによって文の可読性の変化を観察する．しかし，文節をそのままの形で扱うのでは処理がしにくいと思われる．そこで，文節をある特定のタグに置き換えることを考える．これを文節タグと呼ぶことにする．表 1，2 に，用いた文節タグの例を示す．（〜節の節の部分を省略）

表 1　文節タグの例（1 単語からなるもの）

が [格助詞]	が [接続助詞]
から [格助詞]	から [接続助詞]
で [格助詞]	で [助動詞]
で [接続助詞]	でも [副助詞]
に [格助詞]	の [連体化]
は [係助詞]	を [格助詞]
形容詞	接続
短文	動詞
副詞	名詞

表 2　文節タグの例（2 単語からなるもの）

だ [助動詞] ＋が [接続助詞]	だ [助動詞] ＋から [接続助詞]
だ [助動詞] ＋と [接続助詞]	だけ [副助詞] ＋に [格助詞]
て [接続助詞] ＋は [係助詞]	て [接続助詞] ＋も [係助詞]
で [格助詞] ＋は [係助詞]	で [助動詞] ＋は [係助詞]
で [格助詞] ＋も [係助詞]	と [格助詞] ＋は [係助詞]
と [格助詞] ＋も [係助詞]	に [格助詞] ＋は [係助詞]
に [格助詞] ＋も [係助詞]	

4　実験と評価

まずはじめに，理解しやすい文として **1995 年の毎日新聞の記事中の文**を仮定し，それを CaboCha によって依存構造解析し，ルート節に直接掛かっている文節をもとに，文節に分ける．そして，ルート節以外の文節を並べ替えた文を用意し，それらに人手によって理解しやすい文か分かり難い文かの判定を付与した後，各文節を文節タグに置き換える．

次に，文節タグに置き換えた各例文のうち**理解しやすいと判定した文について**各文節タグの平均距離を求め，それを用いて，式 (1) に示す文の分かりやすさの指標値 f を計算する．

$$f = (D_1 - d_1)^2 + (D_2 - d_2)^2 + \cdots + (D_n - d_n)^2 \quad (1)$$

ここで，n はルート節に直接掛かっている文節数，$D_1, D_2, \ldots, D_n$ はその文に含まれる各文節とルート節との距離，$d_1, d_2, \ldots, d_n$ はそれらの平均距離である．

最後に，計算された指標値に対して閾値を設定し，指標値が閾値以下なら理解しやすい文であり，閾値以上なら分かり難い文である，のように判定を付け，それが人手による判定とどれだけ一致しているのか，精度を計算する（閾値を変化させながら，精度が高くなる閾値を調べる）．

5　まとめ

例文 1,000 文について処理した結果，文節を並べ替えた文は 20,630 文となった．そのデータに対し，考案したシステムによって精度を計算したところ，閾値が **0.39** のときに精度が最も高くなり，**90.3%** を得ることができた．

本研究において，理解しやすいか分かり難いかの判断を自分のみで行ったが，今後，アンケートなどを通して，その結果を指標値に反映することが可能ならば，さらに性能を向上できるのはないかと考えられる．また，システムとして完成させるために，文節タグ付けの自動化を実装することが今後の課題であると考えられる．

参考文献

[1] 松尾 樹：“依存構造解析を利用した単語の入れ替えによるわかりやすい文への変換法”，愛知県立大学情報科学部卒業論文，2011