

やさしい日本語ニュースのための自動文分割

美野 秀弥 田中 英輝

NHK 放送技術研究所 〒157-8510 東京都世田谷区砧 1-10-11

E-mail: {mino.h-gq, tanaka-h.ja}@nhk.or.jp

1. はじめに

著者らは、ニュースを国内の外国人住民のためにやさしく書き換えてサービスする研究を行っている[1]。本稿では、このような目的に有用な「自動文分割」に関する研究を報告する。

ニュースをやさしくする手段には、主に単語をやさしくすることと構文をやさしくすることがあり、作業者はこの2種類の手段を組み合わせでやさしく、かつニュースとして不自然にならないように書き換えている。

本稿では、構文をやさしくする主要な手段の1つである文分割に着目する。文を分割するには、分割が可能な点を判定した後、その分割点の前部の文末表現の修正、後部の接続表現や主語の補完などを行う必要がある。このとき、文書内の全ての分割点を分割すると、接続表現や主語の補完などにより文書内の文字数が増える場合や、分割した前後の文の文字数のバランスが悪くなる場合などがあり、必ずしも自然なやさしい日本語ニュースとはならない。

そこで、本稿では、やさしい日本語ニュースに書き換えるための文分割の実施の判定に関係する特徴を用いて、文分割を整数計画問題として定式化して解くことを提案する。分割点の対象は、頻度の高さから、用言を含む文節を修飾する従属節の直後とする。

2. 先行研究

文分割に関する研究として、江原ら[2]や武石ら[3]は規則を用いた文分割手法を提案している。彼らは、文の表層情報、分割点前後の形態素情報などを用いて分割規則を作成し、作成した規則を適用することで文分割を行っている。しかし、これらの研究では、分割した後の文書全体の評価については考慮していない。

一方、子供を対象に文書をやさしくするための研究として、De Belder ら[4]はやさしく書き換えるための規則を作成し、これらを用いて難易度を最低にする最適な文分割をすることを、整数計画法を用いて実現している。しかし、この研究では難易度の制御のみを考えており、文の自然性については必ずしも考慮していない。

著者らは、難易度、すなわち、やさしさの条件に加え、すでに提供されているやさしい日本語ニュースに近づけることで自然なやさしい日本語ニュースを実現できると考えた。本稿では、このような条件を満たした文分割をする手法を提案する。

3. 提案手法

著者らは、De Belder ら[4]の手法を参考にして、ニュースをやさしくするための文分割を下記の3つのステップで行うことにした。

- i) 対象のニュースに対して文節間の係り受け解析を行い、解析結果から分割候補点を抽出する
- ii) 分割候補点全てに分割規則を適用して下記の予備的な文分割処理を行う
 - (ア) 主語の補完
 - (イ) 文末表現の補完
 - (ウ) 接続詞の補完
- iii) 予備的な文分割処理の結果を用いて分割候補点の中から最適な分割点を求めて、これらの文分割処理のみを採用する

以下、3.1.節では i の分割候補点について、3.2.節では ii の分割規則について、3.3.節では iii の最適分割点の判定についてそれぞれ説明する。

3.1. 分割候補点

本稿では、用言を含む文節を修飾する従属節の直後を分割候補点とする。これらを、係り受け解析器 CaboCha¹を用いて抽出した。下記の例では、「続き、」が「あります。」にかかっており、「続き、」の直後が分割候補点となる。

- 関東では／きょうも／雨が／**続き、**／土砂災害の／危険性が／高くなっている／地域が／あります。

3.2. 分割規則

文分割処理に必要な分割規則は、江原ら[2]と武石ら[3]の手法を参考にして作成した。規則の一部には節境界情報を用いた。節境界情報とは、連用節、連体節、並列節など文節の種類を表す名称であり、節境界解析ツール CBAP[5]を用いて獲得した。

¹ <http://code.google.com/p/cabocha/>

そして、規則の拡張のために、作成した規則を訓練データ 200 文に適用し、誤った箇所については規則の追加、修正を加えた。

この結果、主語の補完に関する規則数は 2、文末表現の補完に関する規則数は 23、接続詞の補完に関する規則数は 26 となった。

分割規則の例と適用例を以下に示す。ただし、「/」は分割点を、A,B,C は文節を表している。

| |
|--|
| (「主語の補完」の規則の例) |
| A(CBAP:主題ハ)B/C。 → AB。 AC。 |
| (「文末表現の補完」の規則の例) |
| ～、/～ました。 → ～ました。～ました。 |
| ～ず、/～ました。 → ～ませんでした。～ました。 |
| ～ており、/～ました。 → ～ました。～ました。 |
| (「接続詞の補完」の規則の例) |
| A(CBAP:並列節ガ)/B → A。しかし、B |
| A(CBAP:ナド節)/B → A。このように、B |
| A(CBAP:理由節ノデ)/B → A。このため、B |
| (ニュースへの規則の適用例) |
| このつがいは、4個の卵を産んだとみられていましたが、観察のカメラが故障して撮影できなくなり、22日、改めてカメラを設置して映像を確認したところ、ひなの姿が映っているのが確認されました。 |
| ↓ |
| このつがいは、4個の卵を産んだとみられていました。しかし、観察のカメラが故障して撮影できなくなりました。そして、22日、改めてカメラを設置しました。そして、映像を確認しました。すると、ひなの姿が映っているのが確認されました。 |

3.3. 最適分割点の判定

本節では、3.1.節で抽出した分割候補点の中から最適な分割点を求める手法について説明する。提案手法は De Belder ら [4] の手法を拡張したものである。

彼らは、文書中の単語数 W と文数 S を用いて文書全体の難易度を式(1)で表している。式(1)の係数 α と β は、難易度が付与された正解データを使い、線形回帰を用いて求めている。

$$v = \alpha W + \beta S \quad (1)$$

そして、難易度の変化「(書き換え前の難易度) - (書き換え後の難易度)」を表した式(2)と、分割点の決定を表している式(3),(4)を使い、式(2)が最大となる分割点を、整数計画法を用いて求めることを提案している。以下、変数の定義を説明する。

i は文書中の文の位置を、 j はその一つの分割を表す。文 $i(i \geq 1)$ が複数の分割点を持つ場合はさまざまな分割が可能となる。例えば、文 i に分割点 a, b があれば、 (a) 、 (a, b) 、 (b) の 3 通りの分割が可能となる。これらが「一つの分割」であり、それぞれを $j(j_1 \geq j \geq 1)$ で表す。

a_{ij} は分割の採用を表す変数で、 a_{ij} が 1 ならば文 i に対して分割 j を適用し、0 ならば適用しないことを示す。文 i に対して分割を行わない場合は、 $j=0$ で表し、 $a_{i0}=1$ とする。

Δw_{ij} は単語数の変化を示し、 Δs_{ij} は文数の変化を示す。変化とは「(分割前の値) - (分割後の値)」のことである。

$$\max : \sum_{ij} (\alpha \Delta w_{ij} + \beta \Delta s_{ij}) a_{ij} \quad (2)$$

$$\text{subj.to} : a_{ij} \in \{0, 1\} \quad (3)$$

$$\sum_{j=0}^{n_i} a_{ij} = 1, \forall i \geq 1 \quad (4)$$

以上見てきたように、De Belder らは難易度の変化が最大となる分割点を求める問題を解いている。

著者らは、全体としてはこの方式に従って難易度の変化が最大となる分割点を求める。ただし、De Belder らが使用した単語数と文数の他に係り受け間の距離を使う。さらに、すでに提供されているやさしい日本語ニュースに近づけるための特徴も使う。

3.3.1.節では提案手法で用いた特徴について説明し、3.3.2.節では式(1)から(4)の拡張について説明する。

3.3.1. 分割に使う特徴

本稿で用いる、難易度を表す特徴と、やさしい日本語ニュースに近づけるための特徴を説明する。

i) 難易度を表す特徴

- 単語数の変化 (Δw) と文数の変化 (Δs)

これらは De Belder ら [4] が用いた特徴と同じである。

- 係り受け間の距離の変化 (Δd)

張ら [6] は、外国人を対象に日本語の難易度の主観評価値と相関の高い特徴を調査し、係り受け間の距離を有効な特徴の 1 つに挙げている。本稿ではこれの特徴に用いた。距離は文節間の文節数を示す。例えば、文節列 $g_1 \cdots g_i \cdots g_j \cdots$ の g_i と g_j の距離 d は $d=j-i$ となる。

ii) やさしい日本語ニュースに近づけるための特徴

NHK では「NEWS WEB EASY¹」を開始している。

著者らは、このニュースに合わせた文分割を行いたいと考え、以下の特徴を使った。

- 1 文の平均長との差の変化 ($\Delta p = |l - 35|$)

1 文の文字数 l と、1 次編集² ニュース (以下、NWE ニュース) 30 記事の 1 文の平均長 35 文字との差の絶対値 $|l - 35|$ を特徴に用いた。

¹ <http://www3.nhk.or.jp/news/easy/>

NHK の報道局、放送文化研究所、放送技術研究所が共同で実施しているやさしい日本語ニュースの実験サービス

² 「NEWS WEB EASY」では日本語教師と記者が共同で一般のニュースを書き換えており、本稿では、やさしい日本語を理解した日本語教師が書き換える作業を 1 次編集と呼ぶ

● 節境界の相対頻度 ($c = n_r / N_r$)

元ニュースと NWE ニュースの 30 対のデータを対象に、元ニュース中の分割候補点に節境界 r が現れる頻度 N_r と実際にこれが NWE ニュースで分割されている頻度 n_r の相対頻度 n_r / N_r (以下、節境界の相対頻度) を調べたところ、節境界の種類によってこの値に大きな差があった(表 1)。そこで、この値を特徴に用いた。

表 1：節境界の相対頻度(一部)

| 並列節ガ | 連用節 | 並列節デ | テ節 | ナド節 |
|------|------|------|------|------|
| 0.72 | 0.53 | 0.47 | 0.24 | 0.20 |

3.3.2. 整数計画法の利用

3.3.1. 節にある特徴を用いて、De Belder ら[4]の式(1)から(4)を拡張した。

式(1)の拡張は式(5),(6)である。式(5)は、文書中の単語数 W と文数 S の他に、文書中の全単語の係り受け間の距離の合計値 D を用いて新たに定義した難易度を表す式である。式(6)は、式(5)にやさしい日本語ニュースに近づけるための特徴 P, C を加えた式である。 P は文書中の全文を対象に計算した Δp の合計値である。 C は文書中の全分割候補点における節境界の相対頻度 c の合計値である。式(5),(6)の係数 $\beta_1 \sim \beta_3, \gamma_1 \sim \gamma_5$ は、正解 1 を付与した元ニュースと正解 0 を付与した NWE ニュースの 30 対のデータを使い、線形回帰を用いて求めた。

$$v_n = \beta_1 W + \beta_2 S + \beta_3 D \quad (5)$$

$$v_l = \gamma_1 W + \gamma_2 S + \gamma_3 D + \gamma_4 P + \gamma_5 C \quad (6)$$

本稿では、式(6)の変化を表す式(7)が最大となる分割点を求める。このとき、式(8)から(12)の制約条件を使った。

式(8),(9)は式(3),(4)と同様に分割点の決定を表している。

式(10)は各分割点の係り受け間の距離が 1 以上あることを表している。例えば、「歩いて／帰る」など係り受け間の距離が 0 の場合、「歩いて」は分割点から除外される。

式(11),(12)は NWE ニュースに近づけるための式である。式(11)は式(5)の難易度を下げすぎないために作成した。難易度の変化が \max_d 以下であることを表している。 \max_d は、元ニュースと NWE ニュースの 30 対のデータを用いて式(5)の難易度の変化を計算した際の上限值とした。式(12)は過度に文分割が行われないための条件である。分割後の文数の増加「(分割後の文数) - (分割前の文数)」が \max_s 以下であることを表している。 \max_s は、元ニュースと NWE ニュースの 30 対のデータを用いて文数の増加を計算した際の上限值とした。

$$\max : \sum_{ij} (\gamma_1 \Delta w_{ij} + \gamma_2 \Delta s_{ij} + \gamma_3 \Delta d_{ij} + \gamma_4 \Delta p_{ij} + \gamma_5 c_{ij}) a_{ij} \quad (7)$$

$$\text{subject to : } a_{ij} \in \{0, 1\} \quad (8)$$

$$\sum_{j=0}^{n_i} a_{ij} = 1, \forall i \geq 1 \quad (9)$$

$$\Delta d_{ij} \geq 1, \forall i \geq 1, \forall j \geq 1 \quad (10)$$

$$(\beta_1 \Delta w_{ij} + \beta_2 \Delta s_{ij} + \beta_3 \Delta d_{ij}) a_{ij} \leq \max_d \quad (11)$$

$$\sum_{ij} a_{ij} \Delta s_{ij} \leq \max_s \quad (12)$$

以上、難易度を表す特徴と NWE ニュースに近づけるための特徴を定め、それらを用いた目的関数を提案した。また、NWE ニュースに近づけるための制約条件も追加した。これらの式を使い、整数計画法を用いて最適分割点を求める。

4. 評価実験

3.2. 節で作成した分割規則と 3.3. 節で提案した最適分割点の判定手法を評価した。

4.1. 分割規則

4.1.1. 実験概要

2000 年から 2011 年までの NHK ニュースからランダムに抽出した 450 文を用いて、分割規則の評価を行った。評価者は日本語教師 1 名で、評価手順は下記の通りである。

- 1 文ごとに文中の全ての分割候補点に 3.2. 節の分割規則を適用し、文分割処理を行う。
- 2 分割した箇所 1 つごとに「分割点」、「主語の補完」、「文末の補完」、「接続詞の補完」の 4 つの評価項目について適切・不適切の 2 値で評価した。ただし、補完の項目については分割が適切と評価した場合のみ評価した。

4.1.2. 分割規則の実験結果

実験結果を表 2 に示す。450 文のうち 325 の分割候補点で分割規則が適用され、これらが「分割点」の評価対象となった。そして、「分割点」で適切と評価された 237 の分割候補点で「文末の補完」、「接続詞の補完」、「主語の補完」を行い、それぞれ評価を行った。「主語の補完」は、補完できない場合があるために評価数が他と異なっている。以下、評価結果が不適切となった分割点を調べた。

「分割点」で不適切と評価されたものは、係り受け解析誤りを除けば下記の例のような場合がほとんどであった。

- 混乱の收拾に向け、(分割点)／与野党が支持した。
- ダルビッシュ投手はユニフォームを着て、(分割点)／チームメートと 2 時間練習を行いました。

これらは、分割点の用言が副詞的な役割を担っているため、分割すると不適切になったと考えられる。分割点の用言の役割を明確にし、上記の例のような場合

表 2：文分割処理の実験結果

| 評価項目 | 評価数 | 適切 | 不適切 |
|--------|-----|------------|------------|
| 分割点 | 325 | 237(72.9%) | 88(27.1%) |
| 主語の補完 | 64 | 47(73.4%) | 17(26.6%) |
| 文末の補完 | 237 | 156(65.8%) | 81(34.2%) |
| 接続詞の補完 | 237 | 118(49.8%) | 119(50.2%) |

を分割点の対象外とする必要がある。

「主語の補完」、「文末の補完」で不適切と評価されたものは、係り受け解析誤りによるものがほとんどであった。

「接続詞の補完」で不適切と評価されたものは、適切な規則がないために不適切となったものが多かった。20 種類の節境界情報を用いて作成した 26 の規則で適切な接続詞を付与するには限界があり、節境界情報の細分化を行って規則を追加する必要がある。

4.2. 最適分割点の判定手法

4.2.1. 実験概要

2012 年 4 月から 12 月までの元ニュースと NWE ニュースの 264 対のデータを用いて、最適分割点の判定手法の評価を行った。評価手順は下記の通りである。

- 1) 元ニュースに 3.3 節の提案手法を適用して最適な分割点を判定する。
- 2) NWE ニュースで分割されている箇所を最適な分割点として、1) で求めた分割点と比較を行い(表 3)、その結果を用いて分割精度(式(13))、分割再現率(式(14))、全体精度(式(15))を求める。
- 3) De Belder ら[4]の提案手法を適用して求めた分割点と比較した場合、および、全分割候補点を分割点として比較した場合(ベースライン)についても、2)と同様に分割精度、分割再現率、全体精度を求める。

3.3 節の提案手法にある整数計画問題は lp_solve パッケージ¹を用いて解いた。

$$\begin{aligned} \bullet \quad & \text{分割精度} && \frac{A}{A+C} && (13) \\ \bullet \quad & \text{分割再現率} && \frac{A}{A+B} && (14) \\ \bullet \quad & \text{全体精度} && \frac{A+D}{A+B+C+D} && (15) \end{aligned}$$

4.2.2. 実験結果

実験結果を表 4 に示す。分割精度、分割再現率ともに提案手法が De Belder ら[4]の手法を上回った。

一方、全体精度は De Belder らの手法が上回った。これは、全ての分割候補点を分割しない場合の全体精度が 0.67(1-0.33(ベースラインの全体精度))と高い値であること、De Belder らの手法の分割再現率が低いことが原因と思われる。

表 3：正解との比較

| | 提案手法適用結果 | |
|----------|----------|---------|
| | 分割した | 分割しなかった |
| 正解：分割する | A | B |
| 正解：分割しない | C | D |

表 4：最適分割点の判定の実験結果

| | 分割精度 | 分割再現率 | 全体精度 |
|----------------|------|-------|------|
| ベースライン | 0.33 | 1.00 | 0.33 |
| De Belder ら[4] | 0.28 | 0.24 | 0.55 |
| 提案手法 | 0.37 | 0.64 | 0.52 |

以上の結果から、提案手法がやさしい日本語ニュースのための最適分割点の判定に有効に働いているといえる。しかし、分割精度、全体精度ともに十分ではなく、難易度を表す特徴、やさしい日本語ニュースの特徴をさらに増やして精度を向上させる必要がある。

また、本実験では、1 人の作業者が書き換えた NWE ニュースを用いた。今後、複数の作業者が書き換えたニュースを用意して、作業者間の分割点の一致度を調べる必要がある。

5. まとめ

本稿では、ニュースを外国人住民にとってやさしくするために、De Belder ら[4]の手法を拡張して文分割を行う手法を提案した。

難易度を表す特徴とやさしい日本語ニュースに近づけるための特徴を追加し、これらの特徴を用いて整数計画問題として定式化した。そして、評価実験を行い既存手法と比較した。

今後、さらに特徴を追加して精度を向上させたい。また、本稿では用言を修飾する従属節のみを対象としたが、体言を修飾する従属節も対象にしたい。

文 献

- [1] 田中 英輝, 美野, やさしい日本語によるニュースの書き換え実験. 自然言語処理研究会, Vol.2010-NL-199 No.11, 2010
- [2] 江原暉将, 福島孝博, 和田裕二, 白井克彦. 聴覚障害者向け字幕放送のためのニュース文自動短文分割. 情報処理学会自然言語処理研究会, NL-183-3, 2000.
- [3] 武石英二, 林良彦. 接続構造解析に基づく日本語複文の分割. 情報処理学会論文誌, Vol.33, No.5, 1992.
- [4] Jan DE BELDER, Marie-Francine Moens. Text Simplification for Children. In Proceedings of the SIGIR 2010 Workshop on Accessible Search Systems.
- [5] 丸山, 柏岡, 熊野, 田中. 日本語節境界検出プログラム CBAP の開発と評価. 自然言語処理, Vol.11, No.3, pp.39-68, 2004.
- [6] 張萌, 伊藤彰則, 佐藤和之. 「やさしい日本語」作成支援のための難易度自動推定の検討. 自然言語処理研究会, Vol.2012-NL-206 No.6, 2012.

¹ <http://lpsolve.sourceforge.net/5.5/>