確率的な手法による日本語文簡約†

福冨 諭　　高木 一幸　　尾関 和彦
電気通信大学
{fukutomi,takagi,ozeki}@ice.uec.ac.jp

1　はじめに

単文を要約するために，文から重要と思われる文節を抽出し，出現順に並べて要約文を生成する，文簡約と呼ばれる手法が提案されている．小黒らは文節の重要度と，文節間の係り受けの強さを数値化したときに，それらの合計を最大にするようなアルゴリズムを開発した [1]．ヒューリスティックスを用いてこれらの尺度を推定する手法が研究されている [2, 3].

一方，文簡約とは原文から生成される確率が最も大きい簡約文を求める問題と解釈することができる．Knight らは noisy-channel model を用い，“重要要素に冗長要素が加わって記事ができた” というアイデアを基に，重要要素の存在する確率と冗長要素の追加される確率から簡約文の確率を推定し，確率を最大にする簡約文を求める手法を報告した [4].

本稿では文簡約問題を “原文から生成される確率が最大になる簡約文を求める問題” と解釈し，かつ [1] のアルゴリズムを適用するため，[3] で用いられた文節間の係り受けの強さの推定法の 1 つに変更を加えたものと，[2, 3] で用いられた文節の重要度の推定法の 1 つを用いてそれぞれの評価尺度を推定する手法を提案する．

2　文簡約のための確率言語モデル

簡約文 $S = w_1 w_2 \cdots w_n$ を原文の長さ n の部分文節列と定義する．ここで w_i は原文の文節であって，原文の持つ素性 (原文の係り受け構造など) を保持している．S について考えられうる全ての係り受け構造のうちの 1 つを $D = d_1 d_2 \cdots d_{n-1}$ と定義する．ここで d_i は D において w_i が $d_i(w_i)$ を修飾する事象である．S と D を結合させたものを簡約構造木と呼び，ST と表現する．$n = 1$ のとき，ST は 1 文節 w_1 からなる．$n \geq 2$ のとき，

$$ST = (w_1, w_2, \ldots, w_n, d_1, d_2, \ldots, d_{n-1})$$

である．原文から簡約構造木 ST が生成される確率を $P(ST)$ とし，原文から簡約文が生成される確率を

$$P(S) = \max_D P(ST)$$

と定義する．そして文簡約問題を，$\arg\max_S P(S)$ を探索する問題と定義する．

2.1　生成確率の簡単化

$n = 1$ のとき $P(ST) = P(w_1)$ である．$P(w_1)$ は文節 w_1 が簡約文に選択される確率を表わす．$n \geq 2$ のとき

$$\begin{aligned} P(ST) &= P(w_1, w_2, \ldots, w_n, d_1, d_2, \ldots, d_{n-1}) \\ &= P(d_1 | w_1, w_2, \ldots, w_n, d_2, \ldots, d_{n-1}) \\ &\quad \cdot P(w_1 | w_2, \ldots, w_n, d_2, \ldots, d_{n-1}) \\ &\quad \cdot P(w_2, \ldots, w_n, d_2, \ldots, d_{n-1}) \end{aligned}$$

である．

2 文節間に係り受け関係が成立する確率は当該 2 文節のみに依存すると仮定すると

$$P(d_1 | w_1, \ldots, w_n, d_2, \ldots, d_{n-1}) = P(d_1 | w_1, d_1(w_1))$$

となる．また，ある文節が簡約文に採用されるかどうかはその文節の素性のみに依存し，他の文節が採用されるかどうかの影響を受けないと仮定すると

$$P(w_1 | w_2, \ldots, w_n, d_2, \ldots, d_{n-1}) = P(w_1)$$

となる．これらの仮定から簡約構文木の生成確率は

$$\begin{aligned} P(ST) &= P(w_1) P(d_1 | w_1, d_1(w_1)) \\ &\quad \cdot P(w_2, \ldots, w_n, d_2, \ldots, d_{n-1}) \end{aligned}$$

†Japanese Sentence Compression using Probabilistic Approach

のように簡単化できる.

この変形を再帰的に適用することにより

$$P(ST) = \begin{cases} P(w_1), & \text{if } n = 1, \\ \prod_{i=1}^{n-1} P(d_i|w_i, d_i(w_i)) \prod_{i=1}^{n} P(w_i), & \\ & \text{if } n \geq 2 \end{cases}$$

および

$$P(S) = \begin{cases} P(w_1), & \text{if } n = 1, \\ \max_D \prod_{i=1}^{n-1} P(d_i|w_i, d_i(w_i)) & (1) \\ \quad \cdot \prod_{i=1}^{n} P(w_i), & \text{if } n \geq 2 \end{cases}$$

を得る.

2.2 文簡約アルゴリズムの適用

式 (1) の両辺の対数をとると

$$\log P(S) = \begin{cases} \log P(w_1) & \text{if } n = 1, \\ \max_D \sum_{i=1}^{n-1} \log P(d_i|w_i, d_i(w_i)) & \\ \quad + \sum_{i=1}^{n} \log P(w_i), & \text{if } n \geq 2 \end{cases}$$

となり，$\log P(S)$ を最大にする S は [1] のアルゴリズムによって求めることができる.

このアルゴリズムでは $\log P(d_i|w_i, d_i(w_i))$ に相当する値は係り受けの強さと呼ばれる．2 文節が簡約文に採用された際に係り受け関係を持つ確率の対数なので，これを係り受けの強さと解釈するのは不自然ではない．同様に $\log P(w_i)$ に相当する値は文節重要度と呼ばれる．文節が簡約文に採用される確率の対数は文節の重要性を表現すると考えても不自然ではない.

3 確率の推定

3.1 コーパス

毎日新聞全文記事および 54 文字データベース [5] の 2002 年 5 月から 2003 年 3 月までの 28423 記事を学習セットとし，2002 年 4 月の 50 記事を抽出して評価セットとした．このコーパスは新聞記事とその人手による要約文が対になっているものである．文節重要度の学習では記事と要約文のそれぞれ全文を用い，係り受けの強さの学習では記事の第 1 文と要約文の全文を用いた．各文は JUMAN[6] と KNP[7] によって係り受けを解析した.

3.2 文節間の係り受けの強さの推定

原文中の文節 $w_i, w_j (i < j)$ についてそれぞれの文節のクラス $C_k(w_i), C_u(w_j)$ と，文節間の関係 $C_r(w_i, w_j)$ ごとに，係り受けの強さの推定を行なう.

$C_k(w), C_u(w)$ は文節 w を係り文節と考えたときと受け文節と考えたときのクラスであり，次のような形態的な特徴によって分類する [3]．係り文節をクラス分けするための素性は次の 3 つである:

- 主辞 (最後の自立語) の品詞.
- 主辞が活用語ならばその活用形.
- 主辞が活用語でなく，付属語があれば，その付属語の文字列表現.

受け文節をクラス分けするための素性は次の 4 つである:

- 主辞の品詞.
- その文節が文末であるかどうか.
- 主辞が名詞の場合，決定詞 (“だ”, “である”) を含むかどうか.
- 主辞が形容詞の場合，その活用形.

係り文節のクラスは約 200，受け文節のクラスは約 100 となった.

文節間の関係 $C_r(w_i, w_j)$ は次の 3 つに分類される:

1. w_i が w_j を修飾する場合．図 1 の例では “議員は” と “述べた．” との関係など.
2. 係り受けを木構造で表わしたとき，2 文節がともにルート (文末) から終端 (修飾されない文節) までの経路の上にある場合．図 1 の例では “責任を” と “辞任すると” との関係など.
3. それ以外の場合．図 1 の例では “議員は” と “辞任すると” との関係など.

原文の全ての文節対 w_i, w_j について，3 つ組

$$(C_k(w_i), C_w(w_j), C_r(w_i, w_j))$$

を考え，文節 w_i, w_j がともに要約文に表われる回数と，要約文中で w_i が w_j を修飾する回数を記録する．最終的に

$$\begin{aligned} & P(d_i|w_i, d_i(w_i)) \\ & \approx P(d_i|(C_k(w_i), C_w(d_i(w_i)), C_r(w_i, d_i(w_i)))) \\ & = \frac{w_j \text{ が } w_k \text{ を修飾する回数}}{(w_j, w_k) \text{ が要約文に出現する回数}} \end{aligned}$$

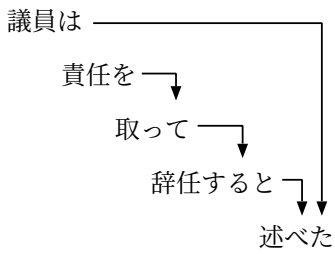

図 1: 例文 “議員は責任をとって辞任すると述べた.” の係り受け構造.

のように推定する．ここで (w_j, w_k) とは 3 つ組

$$(C_k(w_i), C_w(d_i(w_i)), C_r(w_i, d_i(w_i)))$$

の性質を持つ学習データ中の文節対である．この値の対数 $\log P(d_i|w_i, w_j)$ が係り受けの強さである．

3.3 文節重要度の推定

文節重要度は [2] で報告された手法によって推定する．文節は次のような形態的な特徴でクラス分けされる:

- 主辞の品詞.
- 主辞が名詞ならば付属語を持つかどうか.
- 主辞が名詞で助詞を持つならば，その助詞の詳細な品詞.

文節のクラスは約 60 となった.

ある原文に対して文節が簡約文に採用される確率 $P(w)$ は生成しようとする簡約文の文節数に依存する．ここでは次のように仮定する．原文から文節数 m の簡約文が生成されるときに文節 w が簡約文に採用される確率を $P_m(w)$ とすると，文節数 am の簡約文が生成されるときに w が採用される確率は $P_{am}(w) = aP_m(w)$ である.

この仮定は原文からどんな文節数の簡約文が生成されるとしても，2 文節 w_i, w_j が簡約文に採用される確率の比 $P_m(w_i)/P_m(w_j)$ が一定であることを導く．簡約アルゴリズムは文節重要度と係り受けの強さの総和を最大にする部分文節列を求めるものであるから，$\log P_m(w_i) - \log P_m(w_j)$ が一定であれば，簡約文の文節数によって個々の $\log P_m(w)$ が変化したとしても影響を受けない.

以上の考えに基き，文節 w について，そのクラス $C(w)$ ごとに，原文に含まれる $C(w)$ の数と要約文に含まれる $C(w)$ の数の比 (文節残存率) から文節が簡約文に採用される確率 $P(w) \approx P(C(w))$ を推定し，その対数 $\log P(w)$ を文節重要度とする.

4 主観評価実験

4.1 実験条件

次の 3 種類の手法で簡約を行い，比較した:

- 従来手法
  文節残存率と TF-IDF 値 [8] の積の対数を文節重要度とし，原文における 2 文節の係り受けの強さ [2] と 2 文節がともに要約文に採用されて係り受け関係を持つ確率 (文節対残存率) [3] の積の対数を係り受けの強さとするもの．[3] では “結合手法” と呼ばれている.
- 単純化手法
  文節残存率から文節重要度を推定し，文節対残存率により係り受けの強さを推定するもの.
- 提案手法
  3 節の手法によって文節重要度と係り受けの強さを推定するもの.

原文の文節数と簡約文の文節数の比を簡約率と呼ぶ，実験では 1 つの原文につき 70%，50%，30% の 3 種類の簡約率で簡約を行った.

評価セットの 50 の記事のそれぞれについて 3 つの手法，3 つの簡約率によって簡約を行なった．合計で 450 の簡約文を 11 名の被験者が重要情報の保持度，自然さ，総合評価の 3 つの尺度によって 0(最も悪い) から 5(最も良い) の 6 段階で評価した．それぞれの評価点には意味が設定されており，被験者はその意味についての説明を受けた.

4.2 結果

主観評価実験の結果を表 1 に示す.

簡約率が同じ場合は，従来手法が提案手法の評価値よりも高い評価値を得ているが，その差は小さい．従来手法と単純化手法がほぼ同じ評価を得ていることから，文簡約においては TF-IDF 法の効果が小さいといえる．提案手法が他の 2 種類の手法と同等の結果を得

表 1: 主観評価実験の結果．被験者 11 名の平均．括弧内は標準偏差を表わす．

評価尺度	簡約率	従来手法	単純化手法	提案手法
重要情報の保持度	70%	3.91(±0.92)	3.89(±0.90)	3.80(±0.93)
	50%	3.23(±0.95)	3.17(±0.95)	2.97(±0.91)
	30%	2.03(±1.02)	1.93(±1.10)	1.89(±1.01)
自然さ	70%	3.86(±1.19)	3.66(±1.30)	3.69(±1.32)
	50%	3.58(±1.26)	3.47(±1.39)	3.30(±1.43)
	30%	3.15(±1.61)	2.79(±1.66)	2.83(±1.63)
総合評価	70%	3.69(±1.12)	3.57(±1.12)	3.53(±1.14)
	50%	3.06(±1.13)	2.87(±1.31)	2.66(±1.20)
	30%	2.03(±1.07)	1.84(±1.09)	1.83(±1.02)

ていることは，提案手法が従来手法と同程度に有効であることを示している．

簡約率が小さくなるにつれて 3 種類の尺度による評価値はどれも小さくなるが，自然さの変化は他の 2 つに比べて小さい．実験に用いた係り受けの強さの推定法の持つ，原文の係り受け構造を保持する簡約文を生成する性質によって簡約文の自然さが保たれやすいと考える．

5 まとめ

文簡約を “原文から生成される確率が最も大きい簡約文を求める問題” と考え，簡単化のための仮定を用いて文節重要度と文節間の係り受けの強さを導いた．原文と要約文が対になったコーパスからこれらの尺度を学習し，簡約アルゴリズムを用いて文を簡約した．主観評価実験の結果，ヒューリスティックスを用いて文節重要度と文節間の係り受けの強さを推定した従来手法と同等の評価を得た．

謝辞

本研究の一部は日本学術振興会科学研究費補助金基盤研究 (C)16500077 の支援を受けた．

参考文献

[1] Rei Oguro, Kazuhiko Ozeki, Kazuyuki Takagi, and Yujie Zhang. An efficient algorithm for Japanese sentence compaction based on phrase importance and inter-phrase dependency. In *TSD 2000 Proceedings*, pp. 103–108. Springer, 2000.

[2] 諸岡祐平, 江嵜誠, 高木一幸, 尾関和彦. 重要文抽出と文簡約を併用した新聞記事の自動要約．言語処理学会第 10 回年次大会発表論文集, pp. 436–439, March 2004.

[3] Kiwamu Yamagata, Satoshi Fukutomi, Kazuyuki Takagi, and Kazuhiko Ozeki. Sentence compression using statistical information about dependency path length. In *TSD 2006 Proceedings*, pp. 127–134. Springer, 2006.

[4] Kevin Knight and Daniel Marcu. Summarization beyond sentence extraction: a probabilistic approach to sentence compression. *Artificial Intelligence*, Vol. 139, pp. 91–107, July 2002.

[5] 毎日新聞. 毎日新聞全文記事および 54 文字データベース（2002 年度版）, 2002.

[6] 黒橋禎夫, 河原大輔. 日本語形態素解析システム JUMAN version 3.61, May 1999.

[7] 黒橋禎夫, 河原大輔. 日本語構文解析システム KNP version 2.0 β6, June 1998.

[8] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, Vol. 24, No. 5, pp. 513–523, 1988.