概念距離と係り受けを利用した要約文の文節対応付け†

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

WWW の普及を始めとする情報技術の発展により，我々は大量の情報に触れることが可能になった．RSS (Rich Site Summary) のようなウェブサイトの要約を配信するための仕様が開発されるなど，要約技術が脚光を浴びている．

よりよい要約を作るためには，人間が作る要約を分析することが必要である [1]．その基礎となるのは原文の一部と要約の一部を対応付けることである．

原文と要約が対になったコーパスを利用し，原文中の単語で要約に採用されるものの特徴や，採用されないものの特徴を学習して，要約を生成する手法が報告されている [2]．その基礎となっているのが単語の対応付けである．ただしコーパス中の要約は原文中の単語を原文と同じ順序で使ったものであるという制限があった．

本研究では 1 つの文をより短い表現で言い換える文簡約に注目する．文節を対応付けの単位とし，原文と要約の文節間の概念的な距離と，係り受け構造の保持度を定量化したものを用いて対応付けを行う．原文と要約で文節の出現順序が異なったり，表現が変わったりしても対応付けが可能である．

2　文節の対応

文中での意味の単位として文節を用いる．原文と要約の文節について，文中での意味が同じであれば，それらの文節同士は対応しているという．図 1 に示す例文の文節対応付けを図 2 に示す．ϕ は，相当する文節がないことを表わす．

例では要約中の文節 “東京地裁から” に対応する原文中の文節は “東京地裁に” ではなく “同地裁から” とした．開始決定をした主体を対応付けるためである．

† Aligning Phrases in Original Text and its Summary Using Concept Distance and Inter-phrase Dependency

原文: 東京地裁に会社更生法適用を申請した○○工業は，同地裁から会社更生手続きの開始決定を受けたと発表した．
要約: ○○工業が東京地裁から会社更生手続きの開始決定を受ける．

図 1: 原文と要約の例

原文の文節	要約の文節
東京地裁に	ϕ
会社更生法適用を	ϕ
申請した	ϕ
○○工業は	○○工業が
同地裁から	東京地裁から
会社更生手続きの	会社更生手続きの
開始決定を	開始決定を
受けたと	受ける
発表した	ϕ

図 2: 文節の対応の例

本研究では文節の意味によって対応付けを行うため，一部の自立語を付属語とみなし，前の文節に連結する．具体的には形式名詞 “こと” や動詞 “する”，“せる” “なる” である．

3　評価関数

原文 $O = o_1, o_2, \ldots, o_n$ と要約 $S = s_1, s_2, \ldots, s_m$ 中の文節間の対応 $R : \{O, \phi\} \rightarrow \{S, \phi\}$ の評価関数を $f(R) := -\mathrm{SCD}(R) + w\mathrm{SDD}(R)$ と定義する．ここで o_i, s_j は文節，$\mathrm{SCD}(R)$ は文節間の概念距離，$\mathrm{SDD}(R)$ は係り受け構造の保持度，w は重みである．

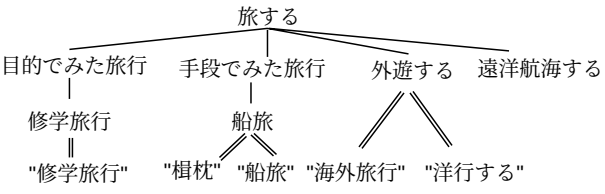

図 3: 概念の木構造

3.1 文節間の概念距離

文節を対応付けるとき，原文と要約とで表現が異なる場合であっても対応付けを行うため，文節に概念を対応させ，文節間の類似度を評価する．例えば “開催する” と “開く” は表現が異なるが，対応する可能性がある．“開く” には “(扉が) 開く” や “(会議を) 開催する” のような概念が対応するため，それを文節の対応付けに利用する．

3.1.1 文節と概念の対応付け

概念辞書によって文節と概念を対応させる．文節中の自立語からなる形態素列を主辞と呼ぶ．主辞の部分列について，長いものを優先し，後方から辞書を検索する．辞書に対応する概念があれば，それを文節の概念に採用し，検索を打ち切る．このとき複数の概念があれば，その全ての概念を採用する．

例えば “経営戦略会議” に対応する辞書の検索は “経営戦略会議”, “戦略会議”, “経営戦略” “会議”, “戦略”, “経営” の順に行われる．

3.1.2 基本的な概念距離の定義

上位概念と下位概念を結ぶと木構造を作ることができる．ある概念から別の概念までの枝の数を概念距離と呼び，$\mathrm{CD}(\cdot,\cdot)$ と表現する．木構造の例を図 3 に示す．各ノードについて，親が上位概念にあたる．ただし 2 重線は概念と文節の対応を表わす．文節は引用符をつけて表記した．概念距離は例えば次のようになる．

$\mathrm{CD}(\text{“修学旅行”}, \text{“旅する”}) = 2$

文節に対応する概念が複数ある場合には，全ての概念間の距離を求め，最小のものを文節間の概念距離とする．例えば “開く” には “(扉が) 開く” や “(会議を) 開催する” などの概念が対応するが，これと “開催する” との概念距離を求めると次のようになる．

$\mathrm{CD}(\text{“開く”}, \text{“開催する”}) = 0$

3.1.3 概念距離の拡張

本研究では概念距離に次のような拡張を行なった．

o_i と s_j とが文字列として一致するか，または主辞の原型が一致する場合には $\mathrm{CD}(o_i, s_j) := -1$ とする．

ϕ と文節の間にも概念距離があるとみなす．パラメータ p を用いて $\mathrm{CD}(o_i, \phi) := p$ とする．

o_i と s_j とに経路がない場合，または文節に対応する概念が辞書になく，上記の条件に当てはまらない場合には，最初は概念距離の計算をしない．経路のある場合の概念距離を全て求めたのち，その最大値に 1 を加えたものを概念距離とする．

3.1.4 概念距離を用いた対応の評価尺度

原文 O の全ての文節について $\mathrm{CD}(o_i, R(o_i))$ を求め，その和を $\mathrm{SCD}(R)$ とする．この $\mathrm{SCD}(R)$ を概念距離を用いた対応の評価尺度として用いる．

$$\mathrm{SCD}(R) := \sum_{o_i \in O} \mathrm{CD}(o_i, R(o_i))$$

3.2 係り受け構造の保持度

原文の係り受け構造が要約でどの程度保存されているかを評価する．

図 2 の例では意味の上から要約中の文節 “東京地裁から” に対応する原文中の文節を “同地裁から” とした．この判断をシステムに行わせるために係り受け構造を利用する．例では要約中の文節 “東京地裁から” が “受ける” に係り，原文中の文節 “同地裁から” が “受けたと” に係っている．“受ける” と “受けたと” が対応していると仮定したときに，“東京地裁から” と “東京地裁に” を対応させるよりも高い評価を “東京地裁から” と “同地裁から” との対応に与えるようにする．

3.2.1 係り受け経路

文節 q_i, q_{i+n} に対して文節列 $Q = q_i, q_{i+1}, \ldots, q_{i+n}$ ですべての $i \leq j \leq i+n-1$ について q_j が q_{j+1} にかかるとき，Q を q_i, q_{i+n} の係り受け経路と呼ぶ．係り受け経路の長さを $\mathrm{DD}(q_i, q_{i+n}) := n$ と定義する．係り受け経路がない場合には $\mathrm{DD}(q_i, q_j) := \infty$ とする．

図 4 に係り受けの例を示す．この例文は 2 つの係り受け経路 (“彼女が”, “所有する”, “鉛筆は”, “短くなった。”) と (“ずいぶんと”, “短くなった。”) によってカバーできる．

係り受け経路の長さは例えば次のようになる．

$\mathrm{DD}(\text{“彼女が”}, \text{“短くなった。”}) = 3$

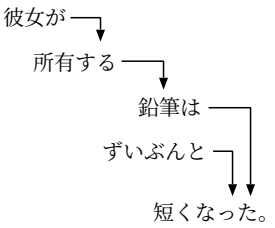

図 4: 例文 “彼女が所有する鉛筆はずいぶんと短くなった。” の係り受け解析木．矢印が係り受けを表わす．

3.2.2 係り受け構造を用いた対応の評価尺度

要約中の係り受けの関係にある文節 s_i, s_j $(i < j)$ について，対応する原文の文節 $R^{-1}(s_i), R^{-1}(s_j)$ を求める．それらの間に係り受け経路が存在すれば，経路の長さの 2 乗に反比例する値を (R, s_i, s_j) の評価値とする．ただし経路の長さが 1 のときに，評価値が 1 となるようにする．

全ての s_i, s_j についての評価値の合計を，係り受け構造を用いた対応の評価尺度と定義する．

$$\mathrm{SDD}(R) := \sum_{\mathrm{DD}(s_i,s_j)=1} 2 \cdot \left(\frac{1}{2}\right)^{\mathrm{DD}(R^{-1}(s_i),R^{-1}(s_j))}$$

4 探索アルゴリズム

原文と要約中の文節の対応を木構造で表現する．ルートからノード N_k までの経路が部分的な対応 R_k を表わす．木のレベル l $(1 \le l \le n)$ は原文中の文節 o_{n-l+1} を表わし，ノード N_k は要約中の文節 $R_k(o_{n-l+1})$ を表わす．文節対応付けとは，この木構造から評価関数 $f(R)$ を最大にする対応 R_{max} を探索する問題である．

原文 “彼女が所有する鉛筆” と要約 “彼女の鉛筆” との文節の対応を図 5 に示す．要約中の文節 “鉛筆” に対応する原文中の文節は “彼女の”，“鉛筆”，ϕ の 3 通りである．このうち “鉛筆” を選び，同様にして “所有する” に対して ϕ，“彼女が” に対して “彼女の” を選ぶと 1 つの対応 R_k ができる．

ノードの探索は幅優先探索で行い，次のように枝刈りをする．N_k がレベル l のノードであるとし，その親を N_h とする．$f(R_k) - f(R_h) \ge \theta$ であるときにのみ N_k を展開する．

$f(R_k) - f(R_h) = -\Delta\mathrm{SCD} + w\Delta\mathrm{SDD}$ とすると以下の式が成り立つ．

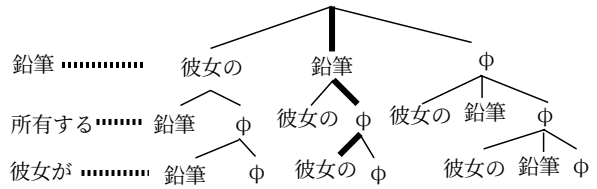

図 5: 原文 “彼女が所有する鉛筆” と要約 “彼女の鉛筆” との対応の探索木．太線が R_{max} を表わす．

$$\Delta\mathrm{SCD} = \mathrm{CD}(o_{n-l+1}, R_k(o_{n-l+1}))$$

$$\Delta\mathrm{SDD} = \begin{cases} 2 \cdot \left(\frac{1}{2}\right)^{\mathrm{DD}(o_{n-l+1},o_i)} & i \text{ が存在する} \\ 0 & \text{それ以外} \end{cases}$$

ただし i $(n-l+1 \le i \le n)$ は次の条件を満たす．

$$\mathrm{DD}(R_k(o_{n-l+1}), R_k(o_i)) = 1$$

5 実験

毎日新聞の 2002 年度の記事 [3] を利用して実験した．このコーパスは新聞記事と人手による 54 文字の要約が記事単位で対になっているものである．

記事の第 1 文と要約とを比較し，文節の対応付けを行った．このコーパスでは第 1 文は要約に採用されやすい傾向が報告されている [4]．

係り受け解析には茶筌 [5] と南瓜 [6]，概念距離の計算には EDR 概念体系辞書 [7] を用いた．

対象の記事数は 200 である．実験のための正解データは人手で作成した．正解データ中の文節の対応は 1127 あった．対応する文節間の概念距離の平均値は 1.47 であった．このうち概念間に経路のあるものは 1076 で，概念距離の平均値は 0.72 であった．

ϕ との概念距離 p，係り受け構造の保持度の重み w を $0.0 \le p \le 3.0$，$0.0 \le w \le 6.0$ の範囲で 0.5 刻みで変化させた．枝刈りの閾値 θ は -6.0 に固定した．

ベースラインとして次の 2 種類のプログラムを作り，利用した．1 つは竹内らの手法 [8] を基にしたものである．この手法では原記事中の文と要約記事中の文を対応付ける手掛りとして，次の条件のいずれかを満たす文節同士を対応させる．

1. 文節の主辞の品詞分類と原型が一致した場合．助詞の変化はある程度許容する．複合語は名詞や文字に区切って比較する．
2. 対応する複数の文節が係る場合．
3. 対応する文節が係り，その文節が別の対応する文節に係る場合．

表 1: 実験結果

p	w	再現率	適合率	F 尺度
0.0	1.5	91.4%	79.0%	0.847
0.5	1.5	90.7%	81.0%	0.856
1.0	1.5	90.4%	81.6%	0.858
1.5	1.5	90.4%	82.4%	0.862
2.0	1.5	87.8%	82.4%	0.850
2.5	1.5	86.5%	82.3%	0.843
3.0	1.5	85.5%	83.6%	0.845
0.5	0.0	88.2%	78.0%	0.828
0.5	1.0	90.7%	81.1%	0.856
0.5	2.0	90.6%	81.6%	0.859
0.5	3.0	90.4%	82.3%	0.862
0.5	4.0	89.3%	82.2%	0.856
0.5	5.0	89.0%	82.9%	0.858
0.5	6.0	87.6%	82.7%	0.851
ベースライン (1)		92.0%	40.6%	0.563
ベースライン (2)		83.8%	71.6%	0.772
ベースライン (3)		75.6%	76.7%	0.809

これをベースライン (1) と呼ぶ.

品詞分類の代りに概念距離を用いたものをベースライン (2),(3) と呼ぶ. それぞれ概念距離が 0,1 以下のときに文節を対応付ける.

結果のうち F 尺度を最大にするものを含む一部と, ベースラインの 3 組を表 1 に示す.

6 考察

ベースライン (1) とベースライン (2),(3) を比較すると, (2),(3) の F 尺度が大きい. 概念を利用した文節間の類似度の評価に効果があることを示している.

ベースライン (2),(3) と提案手法の比較から, 係り受け構造の評価には複数の対応を比較して, 最適なものを選ぶことに効果があると言える.

$w = 1.5$ のとき, p の増加に対して再現率が減少傾向, 適合率が増加傾向にある. $p = 1.5$ のときに F 尺度が最大になっている. これは $w \leq 1.5$ のときに見られる. $w \geq 2.0$ のときは F 尺度が最大になるのは $p = 0.5$ または $p = 1.0$ のときである. w が小さいときには, 評価関数への概念距離の影響が大きくなるが, 正解データの概念距離の平均値に近い p のときに F 尺度が大きくなっていると考えられる. $w \geq 2.0$ の範囲では概念間に経路のある場合の距離の平均値に近い p のときに F 尺度が大きくなっている.

$p = 0.5$ のとき, $w = 0$ の場合と $w > 0$ の場合を比較すると $w > 0$ の方が適合率, 再現率ともに高い. これは $p = 3.0, w \geq 4.5$ の範囲以外でみられる傾向である. 係り受け構造を対応の評価に利用することに効果があることを示している.

7 まとめ

原文と要約の文節間の対応付けを行った. 実験では再現率 90.4%, 適合率 82.4%という結果を得た. これは従来の手法による結果と比較して良好であり, 文節間の概念的な距離と係り受け構造の保持度を利用した効果があった.

今後の課題には文節の文字列から概念を求める手法の改良が挙げられる. また, 枝刈りをした部分木に R_{max} が含まれる可能性があるが, 結果にどの程度の影響を及ぼすかについても調べる. 将来的には原文と要約を比較し, その結果をもとに要約を生成するシステムを作りたい.

参考文献

[1] Regina Barzilay, Lillian Lee: "Learning to paraphrase: An unsupervised approach using multiple-sequence alignment," HLT-NAACL 2003, Main Proceedings, pp.16–23, 2003.

[2] Kevin Knight, Daniel Marcu: "Summarization beyond sentence extraction: A probabilistic approach to sentence compression," Artificial Intelligence 139, pp.91–107, 2002.

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

[4] 諸岡祐平他: "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, 2004.

[5] 松本裕治他: "日本語形態素解析システム『茶筌』version 2.2.1 使用説明書," 2000.

[6] 工藤拓, 松本裕治: "チャンキングの段階適用による日本語係り受け解析," 情報処理学会論文誌, Vol.43, No.6, pp.4834–1842, 2002.

[7] 日本電子化辞書研究所: "EDR 電子化辞書仕様説明書," 日本電子化辞書研究所, 1995.

[8] 竹内和広, 松本裕治: "自動文節対応付け手法を用いた要約生成操作の調査," 情報処理学会研究報告, 2002-NL-147, pp.21–28, 2002.