5J-2

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 ¶

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．

本研究では 1 つの文をより短か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．

2 文節の対応

本研究では文中での意味の単位として文節を用いる．原文と要約で，文中での意味が同じ文節同士を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．

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 $R : \{O\} \rightarrow \{S, \phi\}$ の評価関数を $f(R) := -\mathrm{SCD}(R) + w\mathrm{SDD}(R)$ と定義する．ここで $\mathrm{SCD}(R)$ は後で定義する文節間の概念距離，$\mathrm{SDD}(R)$ は係り受け構造の保持度，w は重みである．

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

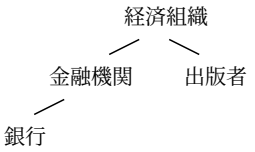

図 3: 概念の木構造

3.1 文節間の概念距離

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 CD と呼ぶ．図 3 の例では CD(銀行, 出版者) = 3 となる．本研究では概念距離にいくつかの拡張を行なった．

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．対応する文節がない場合にはパラメータ 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には経路がある場合の最大の距離に 1 を加えたものを概念距離とする．

原文 O の文節全てについて $\mathrm{CD}(o_i, R(o_i))$ を求め，その和を $\mathrm{SCD}(R)$ とする．

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

* Aligning Phrases in Original Text and its Summary Using Concept Distance and Inter-phrase Dependency
† Fukutomi Satoshi
‡ Takagi Kazuyuki
§ Ozeki Kazuhiko
¶ The University of Electro-Communications

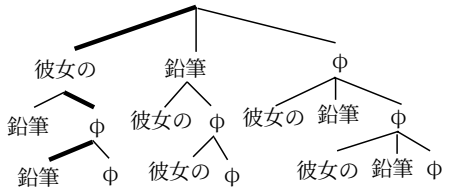

図 4: 原文 “彼女が所有する鉛筆” と要約 “彼女の鉛筆” との対応の探索木．太線が R_{max} を表わす．

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．文節 q_i, q_{i+n} に対して，文節列 $q_i, q_{i+1}, \ldots, q_{i+n}$ ですべての $i \leq j < i+n$ について q_j が q_{j+1} にかかるとき $\mathrm{DD}(q_i, q_{i+n}) := n$ と定義する．そうではない場合には $\mathrm{DD}(q_i, q_j) := \infty$ とする．

係り受け構造の保持度の評価関数を次のように定義する．

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

4 探索アルゴリズム

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 4 のように表わす．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．

ルートから各ノードまでの経路が対応 R を表わす．それぞれについて $f(R)$ を求め，親ノードの評価値との差分が閾値 θ よりも小さい場合にはノードを展開しない．

5 実験

毎日新聞の 2002 年度の記事 [2] を利用して実験した．記事の第 1 文とその記事の 54 文字の要約とを比較し，文節の対応付けを行った．

係り受け解析には茶筌 [3] と南瓜 [4]，概念距離の計算には EDR 概念体系辞書 [5] を用いた．

対象の記事数は 200 である．ϕ との概念距離 p，係り受け保持度の重み w を $0.0 \leq p \leq 3.0, 0.0 \leq w \leq 3.0$ 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．文節の主辞の文字列や品詞分類を用いた手法 [6] に基づくプログラムを作り，ベースラインとした．結果のうち F 尺度による上位 3 組を表 1 に示す．

表 1: 実験結果

p	w	再現率	適合率	F 尺度
1.0	2.0	90.3%	82.3%	0.861
1.5	2.0	89.2%	81.5%	0.852
1.0	3.0	88.9%	81.0%	0.848
ベースライン		92.0%	40.6%	0.563

6 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 90.3%，適合率 82.3%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識別子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] “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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.