係り受け関係の被覆に基づく重要文抽出法

岡崎　直観 †　　松尾　豊 ‡　　石塚　満 †

† 東京大学大学院情報理工学系研究科
‡ 産業技術総合研究所サイバーアシスト研究センター

1　はじめに

日常生活の様々な場面で要約が活用されている．新聞記事の先頭に付いている見出し，Web ページのタイトルやその概要を記述する RSS (RDF Site Summary)，論文のアブストラクトなど，情報の発信者が自ら要約を作成するのはその典型例である．しかし，このような構造化が施されていない文書も多く存在し，仮に構造化されている文書であっても，ユーザ（文書の読者）は新聞のヘッドラインよりも詳細にわたる要約が欲しいなど，要約の質や量がユーザの望むものと合致しないことがある．さらに，ユーザが意図的に収集した複数の文書を要約する場合は，横断的な情報の整理が必要であるから，情報を利用する側が要約を作成しなければならない．

文書自動要約研究 [1] は，大量の文書の中からユーザにとって有用な情報を厳選し，ユーザに提供すべき要約を計算機を用いて作成することを目標としている．中でも，元文書の中から重要と思われる箇所を計算機で推定する重要文抽出法は，ほとんどの文書自動要約システムで中心的役割を果たしている．我々は，文が幾つかの情報断片から構成されていると考え，どの情報断片がユーザーにとって重要で，どの情報断片がユーザにとって不要なのかを考慮しながら，重要文集合を得る手法を提案する．文が保有する情報断片の推定には，文の係り受け構造を利用する．提案手法はユーザが必要としている情報断片を多く含む文を出発点とし，すでに要約に取り込まれた情報断片を考慮しながら，次に抽出すべき文を選ぶ．提案手法の特色として，ユーザが欲しい情報を中心に要約を作成すること，抽出した文の内容の纏まりを考慮すること，冗長な情報をできるだけ省略することなどが挙げられ，複数文書向けの文抽出法としても応用可能である．

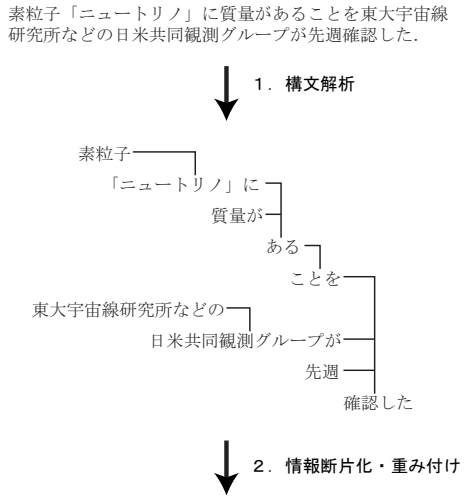
2　文の情報断片

人間は文の意味を解釈し，その解釈に基づいて文章中で重要な箇所を言い当てることができる．計算機も，文章の意味を解釈して何らかの内部表現に変換し，その内部表現に基づいて重要な箇所を決定できれば理想的である．計算機で文の意味を扱う内部表現としては，格文法 [2] や GDA (Global Document Annotation) [3] などがあるが，本研究では構文解析による係り受け関係 [4] から文章の内部表現を作成する．

文を単語ペアによる内部表現に変換する手順を，図 1 を用いて説明する．まず，要約対象の文書を文単位に区切り，構文解析器を用いて文の構文木表現を得る．次に，構文木から係り受け単語ペア（係りの向きは無視する）を取り出し，文を単語ペアの集合表現に変換する．このとき，格助詞や「ある」「こと」などの単語はストップワードとして除去する．図 1 の例では，入力文を 7 個の係り受け単語ペアで表現している．

このように抽出した係り受けペアは，それぞれ「ニュートリノは素粒子だ」「ニュートリノは質量を持つ」「ニュートリノを確認する」「質量を確認する」「東大宇宙線研究所は日米共同観測グループだ」「日米共同観測グループは確認する」「先週確認する」と書き下せる．これらは，元の文が伝えようとしている意味を断片的に表現しているが，図 1 の係り受けペアをこのように書き下せるのは，単語間の関係を人間が補完しているからであり，係り受けペアだけで単語間の意味的な関係を明らかにするわけではない．しかし，計算機が単語間の関係の意味を解釈できなくても，文の書き手は係り受けペアの単語と単語を結び付けて読者に伝えようとしているので，このように抽出した係り受けペアを文の情報断片と呼ぶことにする．

要約対象の文書に含まれるすべての文を情報断片で




```
(
  (素粒子　ニュートリノ　4.35)
  (ニュートリノ　質量　5.20)
  (ニュートリノ　確認　3.31)
  (質量　確認　2.75)
  (東大宇宙線研究所　日米共同観測グループ　2.42)
  (日米共同観測グループ　確認　1.94)
  (先週　確認　1.44)
)
```

図 1: 係り受け構造から情報断片（単語ペア）への変換

表現すると，ある文を抽出するという操作によって，どの情報断片を読者に伝えたことになるのか把握できる．さらに，それぞれの情報断片に重要度（重み）付与すれば，どの文が重要な情報断片を保有しているのか調べ，そのような文を優先的に抽出することができる．今回は係り受けペアの単語の tf*idf 値の相乗平均を情報断片の重みとし，文を単語ペアとその重みの集合で表現した．

3　情報断片の被覆に基づく文抽出

文が情報断片とその重みで表現されているとき，重要文抽出問題は，重要な情報断片を出来るだけ多く含む文の集合を決定する問題として以下のように定式化できる．要約したい文書 D（複数の文書でもよい）が n 個の文 $\{S_1, S_2, ..., S_n\}$ で構成されており，文 S_i の文字数を $\mathrm{length}(S_i)$ で表す．その文書 D に含まれるすべての文を分析した結果，全部で m 個の情報断片 $\{c_1, c_2, ..., c_m\}$ が見つかったとしよう．このとき，文 S_i が保有する情報断片 c_j の重み w_{ij} を何らかの方法で算出し，その重み w_{ij} を要素とする文–情報断片行列 W（n 行 m 列）を作成する：

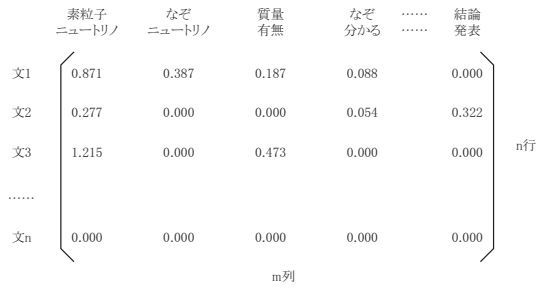


図 2: 文–情報断片行列の例

$$w_{ij} = \begin{cases} S_i \text{ における } c_i \text{ の重み} & (c_i \in S_i) \\ 0 & (c_i \notin S_i) \end{cases} . \quad (1)$$

文–情報断片行列の例を図 2 に示した．

このように表現された文–情報断片行列 W を使って，制限文字数 L 以内で重要文を抽出する方法を考える．すなわち，文書 D から最も重要と思われる（以下の f を最大化する）文 S_i を一つ選び：

$$f = \underset{S_i \in D \setminus E}{\mathrm{argmax}}[\mathrm{weight}(S_i, W, E)], \quad (2)$$

選んだ文を要約文集合 E に加えるという操作を，要約文字数の制約条件：

$$\sum_{S_i \in E} \mathrm{length}(S_i) \leq L \quad (3)$$

が成立している間繰り返す．ただし，$\mathrm{weight}(S_i, W, E)$ は文 S_i の重要度を決定する関数である．

重要な情報断片を多く含む文は，やはり重要な情報を伝達する文であるから，文が保有する情報断片の重みの和を文 S_i の重要度とするのは自然なことであろう：

$$\mathrm{weight}_0(S_i, W, E) = \sum_{j=1}^{m} w_{ij}. \quad (4)$$

この重要度関数 weight_0 を用いて式 2, 3 を適用すると，重要な情報断片を多く含む文から順番に抜き出す文抽出になる．しかし，この抽出法は要約文集合 E 全体を眺めたときの情報断片の重複を考慮していないため，要約したい文書の性質によっては，抽出した文に類似する情報が含まれたり，特定の話題に偏って文を抽出してしまう恐れがある．あるクエリーで検索された文書集合のように，要約したい文書が関連する複数の文

文 1: ((A 1) (D 1))
文 2: ((B 0.8) (D 1))
文 3: ((C 0.5) (D 1))
文 4: ((A 1) (B 0.8))

図 3: 探索的に文抽出を行ったほうが良い例

書の場合は，類似の情報を多く含む傾向があるため，この抽出法では不十分である．

情報断片はいったん伝達してしまうと，その情報が受け手にとっては既知となり，その情報断片の重要度は減少すると考えられる．すでに要約文集合 E の中で述べられた情報断片の重要度を減少させるために，文 S_i の新しい重要度関数 $\mathrm{weight}_1(S_i, W, E)$ を導入する：

$$\mathrm{weight}_1(S_i, W, E) = \sum_{j=1}^{m} \mathrm{novel}(c_i, E) \cdot w_{ij}. \quad (5)$$

ここで，$\mathrm{novel}(c_i, E)$ は情報断片 c_i の要約文集合 E に対する目新しさを表す関数で，簡単のため次のように定義する：

$$\mathrm{novel}(c_i, E) = \begin{cases} 0 & (c_i \text{ が } E \text{ に含まれている場合}) \\ 1 & (c_i \text{ が } E \text{ に含まれない場合}) \end{cases}. \quad (6)$$

つまり，式 5, 6 は要約文集合 E で述べられている情報断片の重要度を以降 0 として，新たに抽出する文の重要度を計算するという意味である．この重要度関数を利用して式 2, 3 で文抽出を行うと，すでに要約に含めた情報断片の重要度は 0 となるため，要約に含められた情報断片を含む文よりも，新しい情報断片を含む文が有利となる．ゆえに，冗長な情報を持つ文を抽出する代わりに，新しい情報を持つ文が優先的に選ばれるようになる．

この重要文抽出法を利用する際には，いくつか有用な変形が考えられる．例えば，図 3 に示すような情報断片 A, B, C, D から構成される 4 つの文があり，ここから 2 つの文を抽出することを考える．重要度関数 weight_1 を用い，式 2 を適用して文を 1 つずつ抜き出すと，まず文 1 が選ばれ，次に文 2 または文 4 が選ばれ，文の重み和は 2.8 となる．しかし文 3 と文 4 を選べば，すべての情報断片が抽出文に含まれ，文の重み和も 3.3 と改善される．このように，抽出する文は組み合わせ

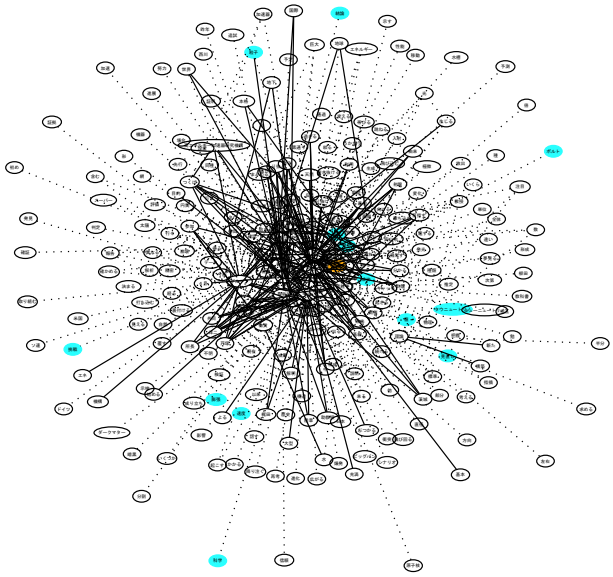


図 4: 情報断片行列のグラフ表示（一部）

的に決めるべきできであるから，式 2 による抽出を探索問題に変形する：

$$F = \underset{E \subseteq D}{\mathrm{argmax}} \sum_{S_i \in E} [\mathrm{weight}(S_i, W, E)]. \quad (7)$$

ただし，F を最大にする E を求めるのは困難であるので，実際には準最適解を求めることになる．

その他の変形として，あるノードを展開（ある文を選択）した後，その文が保有する情報断片に関連する情報断片 [1] の重みを一時的に高く評価し，次に抽出する文の重要度を計算することで，抽出した文集合の内容的な纏まりを改善することが考えられる．また，クエリー等からユーザの知りたい内容が推定できるときは，探索の出発点をクエリーを含む文に限定し，ユーザが欲しい情報を中心に要約を作成するように変形することも可能である．

図 4 はノードを単語，エッジを係り受け関係として情報断片行列を可視化したものである．提案手法は図 4 のエッジの中で重要なものを出来るだけ多く被覆するように文を抽出する手法と捉えてもよい．図 4 から重要文抽出を行い，要約文の中に取り込まれた情報断片は実線で示した．

[1] 情報断片同士の関連性の推定として，例えば同じ語を共有しているならば関連性を認めるなどの方法が考えられる．

要約長＼システム	提案手法	人手による要約	LEAD手法（ベース）	システム平均
short	0.230	0.385	0.160	0.210
long	0.248	0.402	0.159	0.243

表 1: 要約の内容の評価

要約長＼システム	提案手法	人手による要約	LEAD手法（ベース）	システム平均
short	0.067	0.033	1.500	0.458
long	0.167	0.033	2.833	0.725

表 2: 1 要約に含まれる内容重複文の数

4 評価

提案手法による重要文抽出法を組み込んだ自動要約システムを構築した．構文解析器として CaboCha [5] を用い，式 7 による抽出文の探索には，各ノードにおける分枝数を 3〜10 個 [2]に設定した最良優先探索を用いた．国立情報学研究所情報学資源研究センターの支援により開催されているワークショップ NTCIR-4 のテキスト自動要約タスク TSC-3 [6] に参加し，そのテストコレクション [3]を用いて提案手法による自動要約システムを評価した．我々のシステムの詳細，TSC-3 のテストコレクションや評価方法の詳細に関しては 2004 年 5 月に予定されている NTCIR の成果報告会，もしくは会議論文集を参照していただきたい．また，紙面の都合で作成された要約例の掲載を省略するが，TSC-3 では参加システムが作成した要約の公開を検討中とのことなので，併せてそちらも参照していただきたい．

TSC-3 はシステム要約を約 20 個の評価尺度で評価するが，ここでは，被験者が重要文抽出の性能を評価する「内容のスコア付け」と「重複内容の文数」の結果を示す．提案手法の結果のほかに，人手で作成した要約，LEAD 手法（ベースラインシステム）による要約，人手とベースラインシステム，提案手法を除く参加システム全体の平均の評価を比較のために載せた．表 1 は TSC-3 による内容の評価 [4]であり，値が大きいほど内容が優れていると言える．提案手法はいずれの要約長でも LEAD 手法よりも圧倒的に優れており，参加システムの平均よりも若干良いという結果が得られた．

表 2 は，１つの要約の中で内容の重複が認められる文の数の平均を示したものであり，値が小さいほど冗長な内容を含まないことを示す．要約文字数が大きいときは，重複内容を取り込んでしまう可能性が高まるが，それでも提案手法は１つの要約あたり 0.167 文しか冗長な内容を抽出しない．

[2]探索時間が長くなりすぎないように，展開回数は探索空間の広さ（要約の長さ）に応じて自動調節する．

[3]30 件の要約対象文書集合で構成され，参加者は指定された長短 2 種類の要約を作成する．

[4]評価者が自分の作成した要約とシステム要約との間で文対応付けを行い，そのスコア（対応度合い）に基づく要約の評価である．ただし，評価者が作成した要約に含まれる文には重要度のランクが付与されているのでそれも考慮して最終的な評価値を決定している．

5 結論

文の係り受け関係に基づいて文書を情報断片で表現し，重要な情報断片を多く取り込むような文集合を求める最適化問題によって重要文抽出を行う手法を提案した．TSC-3 での評価では，提案手法は LEAD 手法よりも優れており，内容の重複の少ない重要文抽出を実現しているが，重要な情報断片の推定に改善の余地が残されている．今後は，係り受け関係のラベルを活用する方法や，重みの算出方法の改良を通じて，精度向上に努めたいと考えている．

謝辞

本研究にあたっては，NTCIR-3, NTCIR-4 のテキスト自動要約タスク TSC-2, TSC-3 のテストコレクション（毎日新聞記事データ，読売新聞記事データ）と評価データを用いました．

参考文献

[1] 難波英嗣, 奥村学. ここまで来たテキスト自動要約. 情報処理, Vol. 43, No. 12, pp. 1287–1294, 12 2002.

[2] Charles J. Fillmore. The case for case. *Universals in Linguistic Theory*.

[3] Katashi Nagao and Koichi Hasida. Automatic text summarization based on the global document annotation. In *Proc. of COLING-ACL '98*, Montreal, Quebec, Canada, Aug. 1998.

[4] 立石健二, 大庭直行, 峯恒憲, 雨宮真人. 係り受け情報を利用した web 上の日本語テキスト検索システム. 研究報告「デジタル・ドキュメント」, No. 13, pp. 47–54, 1998.

[5] 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. Vol. 43, No. 6, pp. 1834–1842, 2002.

[6] Text Summarization Challenge Home Page. `http://www.lr.pi.titech.ac.jp/tsc/`.