アノテーション付き多文書データからの要約生成

綾 聡平[1]　宮田高志[3]　橋田浩一[2 3]　石塚 満[1]

東京大学大学院情報理工学系研究科電子情報学専攻石塚研究室[1]
産業技術総合研究所サイバーアシスト研究センター[2]
科学技術振興事業団 CREST[3]

〒 113-8656 東京都文京区本郷 7-3-1
Tel: 03-5841-6755
Fax: 03-5841-8570
e-mail: s-aya@miv.t.u-tokyo.ac.jp
URL: http://www.miv.t.u-tokyo.ac.jp

Abstract

新聞や，WWW（World Wide Web）上の電子文書等を筆頭に，今日我々の入手出来る文書は増加の一途を辿っている．しかし，その量は膨大であり，その全て（若しくは殆ど）を読むことは出来ない．文書要約は，このような状況を改善すべく研究されている分野である．文書／文書集合から重要な情報を抽出し，その要諦のみをユーザに提示する．

自動要約の古典的手法としては重要文抽出法が挙げられるが，この手法は冗長性，重複性，結束性等の問題を孕んでいた．このような問題を解決すべく本研究では，文生成に近い形で要約を行うシステムを提案する．また，今日いくつかの自然言語解析ツールが開発されてきてはいるものの，それでも尚高精度での自然言語処理は難しい．そこで今回，GDA と呼ばれるアノテーションを利用して予め明示的に様々な情報を与えてやることで精度の高い要約作成を目指している．

1　はじめに

WWW（World Wide Web）の発達を筆頭に，巷に溢れる情報は日々増加している．しかしその一方で，残念ながら我々の情報処理能力はその速度に追いついていない．そこで，ある文書集合があったときに，その内容をまとめて簡潔に提示するシステムがあれば，短時間で情報を把握でき，非常に有益だと考えられる．自動要約は，このようなニーズに応えるべく研究されている分野である．

要約の手段としてはこれまで，原文から重要と判断される文を抜き出す手法（重要文抽出）が主に用いられてきた．しかし，文単位の抽出は不必要な情報が多く含まれ，要約率があまり上がらない，文間の結束性や一貫性等が維持しづらい等の問題を孕んでいる．また，このような手法は重要語を多く含む文のスコアが高くなることが多い為に長く複雑な文が選択されがちであり，読み手の負担が大きいとの指摘もある [1]．

そこで本研究では，原文から，認知科学で登場した活性拡散という手法を用いて重要語を推定し，そこから文生成に近い形で要約を生成する手法を提案する．これにより，冗長性，重複性を避けた文章を出力することを目標に掲げている．

また，近年 JUMAN，茶筅等の自然言語解析ツールが開発されてきているとはいえ，高精度の自然言語処理を完全自動化することは難しい．そこで，本研究では GDA と呼ばれるアノテーションのついた文書を

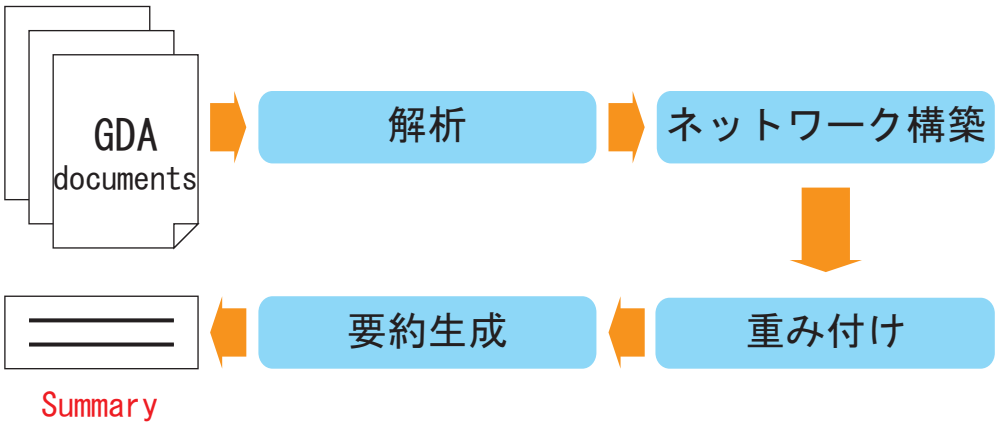

図 1: 提案システムの概要

要約対象とする．GDA の詳細については第 2.1 節で扱うが，係り受け関係，照応・共参照関係等を明示的に与えるために開発された XML の形式である．既存のツールに加え，人手を利用してアノテーションを加えることで，より精度の高い自然言語処理を目指している．

尚，今回の提案手法は，ある事件とその続報記事を扱う一連の記事集合を要約対象としている．

2　提案手法の概要

文書を読み解く際，我々の脳が，どのように内容を把握していくのかは未だ不明であるが，記憶のメカニズムは徐々に解明されつつある．記憶には，ある語が想起（活性化）されるとその語に関連する語も同時に想起されるというプライミング効果（Priming Effect）[2, 3] があり，また，その速度は想起頻度に依存することが様々な認知実験により確認されている [4]．

さらにこのプライミング効果は，文書の内容を理解する際にも深く関与していると考えられている．文書を読み解くにつれて，話題が読者の頭の中に展開され，それに伴って記憶が活性化されていくことで文脈を理解し，内容を把握していると考えられるからである．

このような記憶のメカニズムを近似したモデルとして，活性拡散モデル（Spreading Activation Model）[5, 6, 7] がある．この理論は，人間の認知的側面から構築されたモデルであり，記憶はノードとして表現される語が，ノードの活性を伝播させるリンクで結びついたネットワーク構造で構築される．活性拡散が認知的側面から発展したものであることから，キーワード抽出 [8] 等に応用されており，要約にも Mani ら [9] 等が用いている．

本稿では，重要な部分をピックアップして提示する要約に於いても，このような活性拡散は有効であるとの立場を取り，提案手法は大凡，以下のような流れ（図 1 参照）に基づいて要約を生成する．

1. GDA 文書の解析
2. コンテンツネットワーク構築
3. 活性拡散に基づく重み付け
4. 要約文生成

以下に，これらの詳細について概説する．

```
<su syn="f">
  <adp opr="obj">
    <placenamep id="jpn">日本</placenamep>
    <n id="tagid01">大使</n>
    <n id="tagid02">公邸</n>
    <ad>に</ad>
  </adp>
  <adp opr="agt">
    <np opr="obj" id="tag03" eq="amaru">
      <n>武装</n>
      <n>ゲリラ</n>
    </np>
    <ad>が</ad>
  </adp>
  <v>乱入</v>
</su>
```

図 2: GDA 文書例

2.1 大域文書修飾 GDA

まず，要約対象記事にアノテーションとして用いている GDA について簡単に説明する．

意味や常識，概念等をコンピュータに理解させることは困難であり，それ故に自然言語処理の自動化に成功した例は少ない．そこで，このような状況を改善する為に考えられたのが GDA（Global Document Annotation）である．紙面の関係上，その子細については省略する [1]が，文書中の統語照応構造，修辞構造，対話構造，語義等の情報を予め明示的に示しておく為の，XML（Extensible Markup Language）タグセットである．既存のツールによる解析結果に加え，人手に基づく情報を与えてやることで，計算機上の自然言語処理精度を上げることを目的としている．

これまでに GDA を要約に利用した例としては，[10, 11] 等がある．GDA 文書の例を図 2 に示す．

2.2 コンテンツネットワークの生成

要約対象となる文書集合の GDA を解析し，その解析結果を元にコンテンツネットワークを生成する．このネットワークでは自立語をノードとして，その間にある係り受け，照応・共参照関係等をリンクとして表現する．尚この際，同一内容を示す語は同一ノードで表現し，また，機能語はノードとして表現せずに係り受けのリンク中にその情報を含める．これのような作業により，記事集合の全ての文を統合することが出来，最終的にひとつの巨大なネットワークとなる．

このようなネットワークを構築する目的としては，先述した人間の認知的モデルに近似させることの他にも，以下の 3 つが挙げられる．

1. 例えば「日本大使公邸に乱入した武装ゲリラ」と「トゥパク・アマルが日本大使公邸に乱入」という 2 つの文があった場合,「武装ゲリラ」と「トゥパク・アマル」が同一であることをアノテーションで明示的に与えてあれば，この 2 文はネットワーク的に縮退する為，出力する際にも重複を回避することが出来る．
2. 文書集合をひとつのネットワークで表現している為，その中での距離や隣接性が内容的な近さをある程度反映していると考えることができる．そこで，これを利用することにより，結束性の維持を図ることが期待される．

[1] 詳細は　http://www.i-content.org/gda/

3. 要約を生成する際，文を一から生成し直すことが理想ではあるが，現在の技術では可読性の劣化は避けられない．そこで，原文の係り受け関係や機能語をネットワークで保持する，即ち原文の形もネットワークで表現することにより，原文の情報を利用しながら可読性に配慮した文を作成することが可能となる．

尚，リンク強度は Mani らの手法を参考に

$$\text{照応・共参照関係} > \text{係り受け関係} > \text{その他}$$

と定めており，さらに，関係回数が多い程強度が大きくなるように設計している．また，ネットワークを表現する接続行列 R を作成するにあたり，これを確率行列とする為に，

$$R_{ij} = \frac{l_{ij}}{\sum_k l_{ik}}$$

と正規化している．ここで，語 w_i から w_j に接続しているリンク強度が l_{ij} である．

ネットワークの一例を図 3 [2]に示す．この図中で，リンクの色の濃さは，濃い順に照応・共参照関係，係り受け関係，その他を示している．さらに，リンクの添え字は関係名であり，大凡 “sbj”，“agt” が主格，“obj” が目的格に相当する．“dep” は単純な係り受け関係である．

2.3 活性拡散に基づく重み付け

活性拡散には，いくつかのモデルが提案されており，その多くが外部入力値を考慮に入れている．しかし，読者が文書の内容を理解していく上では，前の記憶状態に無関係に作用する要素があるとは考えにくい．そこで，今回利用している活性拡散のモデルでは，外部入力値を考慮しない，以下の式を用いる．

$$\mathbf{A}\,(t) = \{(1-\rho)\,\mathbf{I} + \rho\mathbf{R}\}\,\mathbf{A}\,(t-1)$$

ここで ρ は減衰定数，$\mathbf{A}\,(t)$ は活性値，$\mathbf{R}$ はネットワークの接続行列である．この式を用いて，活性値が収束するまで計算を繰り返す．

さらに，ここで計算された活性値を用いて，イベント毎の重要度を計算する．具体的には，動詞はあるイベントを表現しているという仮定に基づき，以下のように算出する．

$$(\text{イベントの重要度}) = (\text{動詞の活性値}) + (\text{必須格の活性値})$$

また，これとは別に，活性値上位の語は重要語と捉え，要約生成に利用する．

2.4 要約の作成

活性拡散の結果に基づき，以下のような流れでコンテンツネットワークから要約を作成する．

1. 第 2.3 節で算出したイベントの重要度から，その上位の動詞（個数は要約率に依存する）を抽出し，整列する．今回は対象が新聞記事なので，イベントの起こった順に，時系列に並べている．
2. 先述の動詞をひとつ取り出し，これをある出力文の中心イベントと捉えて出力ノードに加える．

[2] この図は,「ペルーからの報道によると、首都リマ市にある日本大使公邸が１７日午後８時（日本時間１８日午前１０時）過ぎ、左翼ゲリラとみられる武装グループに襲撃され、日本、ペルーの両国関係者多数が人質にとられた。青木盛久大使ほか日本人関係者も公邸内に閉じ込められている。現在も警備の警官隊との銃撃戦が続いている模様。現在、地元警察および特殊治安部隊など約５００人が出動、公邸を包囲し、犯人グループの説得にあたるとともに、散発的に銃撃戦が展開されているという。」という 4 文をネットワークで表現したものである．

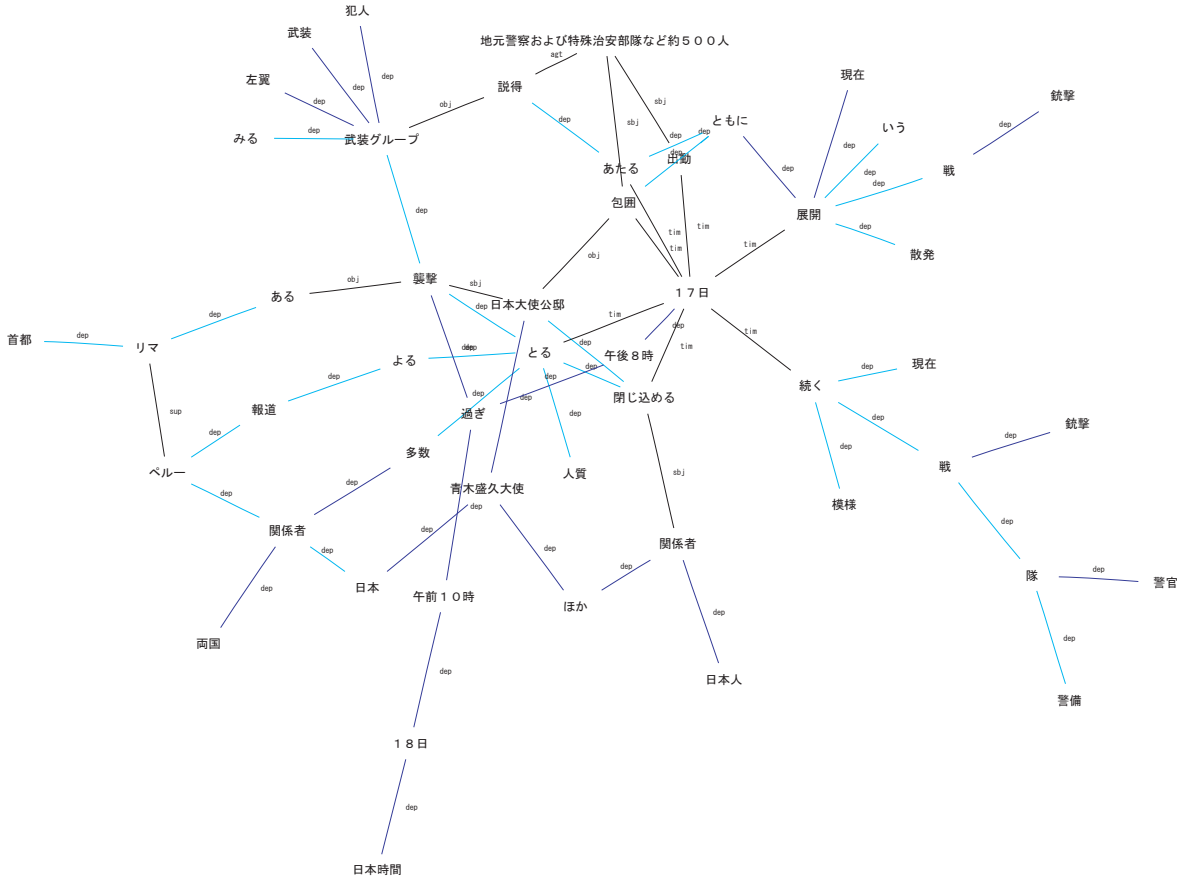

図 3: A sample of network

3. 出力ノードと接続している主語，目的語等の必須格，もしくは活性値上位の重要語を出力対象に加える．しかしこのとき，実際には必須格でもアノテーションが抜け落ちている箇所があり，その厳密な判断が難しい為，応急策として助詞を挟む係り受けも全てピックアップしている．
4. 同様のルールに基づいて順番に遡りながら 3. の処理を繰り返す．このとき，既に出力された文中に同じ修飾が存在する場合には枝刈を行う．
5. 同一の文中から取り出した必須格は，原文中の機能語を使用するが，他の文を参照して取り出した必須格は，助詞等が抜け落ちた形となってしまう．そこで，これらを主格，目的格等に応じて適当に補ってやる必要がある．
6. 出力ノードが固まった後，最後の動詞もしくは助動詞を整形する．新聞記事では，大半の文が過去の事実関係を述べていると考え，全て過去の終止形統一した．尚，この際には，JUMAN [3]による形態素解析の結果を用いている．

また，今回の要約対象は新聞記事であり，その特性上，イベントの起こった日時は重要な情報だと考えられるので，日時も必須格として扱い，出力している．

[3]JUMAN Homepage
http://www.kc.t.u-tokyo.ac.jp/nl-resource/juman.html

フジモリ大統領
ペルー日本大使公邸占拠事件
トゥパク・アマル革命運動（ＭＲＴＡ）
ペルー
シプリアニ司教（バチカン代表）
保証委員会
平和的解決
日本大使公邸
リマ
人質
池田行彦外相
ウゴ・シビナ君（１１）
ペルー政府
日本
橋本龍太郎首相

図 4: 活性拡散の結果（上位 15 語）

3　実験と評価

3.1　実験

前節で説明したシステムを用い，要約作成を行った．今回要約対象としたのは，1996 年から 1997 年にかけてのペルー日本大使公邸占拠事件に関する毎日新聞の記事，全 50 記事に GDA に基づくアノテーションを付与したものである．

活性拡散の結果を図 4 に，要約結果は図 5 に示す．本手法では，まだ字数制限などは行うことができない為，重要度上位 10 イベントから要約を作成，出力している．

3.2　考察

(1) 活性拡散の結果

妥当性に関しては，今後詳細に検討していく必要はあるが，事件の大凡の顛末を考えると，ある程度納得のいく結果が出力されていることが分かる．

(2) 省略語の獲得

GDA 文書を利用している為，主語や日時等が省略されている場合にも，正確に獲得することができる．例えば，要約結果の「フジモリ大統領は２２日姿勢を貫いたことを強調した」は，原文中では「さらに『ペルーはテロリストを認めない．国際社会に我々は模範を示した』と述べ，トゥパク服役囚の釈放要求には一切応じず，テロに屈服しない姿勢を貫いたことを強調した.」となっており,「強調した」というイベントの主語である「フジモリ大統領」やその日付である「２２日」を，GDA で示された必須格に基づいて出力することに成功している．

(3) 内容重複の回避

例えば「日本大使公邸に乱入した武装ゲリラ」と「トゥパク・アマルが日本大使公邸に乱入」というような同一のイベントはネットワークを構築する際に縮退する為，出力文中でも，内容の重複を回避することができている．これは，高い要約率を達成する為に，大変有効だと考えられる．

(4) 冗長性の回避

必須格と重要語を中心に文を組み立てている為，冗長性を回避できている．この冗長性の回避も，先述の重複の回避と共に，高い圧縮率を目指す上で有効である．

しかし，その一方で必須格以外を取得しないことで文がやや細切れになる傾向にあり，可読性に欠ける．

```
<1996 年 12 月 18 日>
    １８日/フジモリ大統領は/パレルモ教育相を/現場に/派遣した//
<1997 年 01 月 10 日>
    フジモリ大統領は/１０日/努力すると/言明した//
<1997 年 01 月 28 日>
    ２８日/１０日/大統領府で/フジモリ大統領と/会談し/公邸/占拠を/続けた//
<1997 年 02 月 17 日>
    2月１７日/大統領を/乗せた//
<1997 年 02 月 17 日>
    2月１７日/大統領は/貧困層の/支持を/強調することで/政府の/態度を/印象づけた//
<1997 年 03 月 03 日>
    フジモリ大統領は/３日/キューバを/訪問//
<1997 年 04 月>
    フジモリ大統領は/４月/内閣の/信任を/公言した//
<1997 年 04 月 22 日>
    フジモリ大統領は/２２日/事前には/日本政府に/通告しなかったことを/認めた//
<1997 年 04 月 22 日>
    フジモリ大統領は/２２日/記者会見の/直前に/首相と/電話で/会談した//
<1997 年 04 月 22 日>
    フジモリ大統領は/２２日/姿勢を/貫いたことを/強調した//
```

図 5: 要約結果（<>内は日付データ）

現時点ではほとんど入っていないが，複数の文を組み合わせて読みやすくする処理等も今後入れていく必要がある．

また，先程の例の「テロに屈服しない姿勢」等のように，出力した方が自然な文となる部分も削除されているケースがある．出力する／しないの判断についても，今後の検討課題である．

(5) 重要度判定

出力結果をご覧頂ければお分かりいただけると思うが，事件の開始／解決に関連する文が抜け落ちている．現在 (動詞の活性値) + (必須格の活性値) としているイベントの重要度算出方法を，もう少し精査する必要がある．

(6) 結束性

今回，全ての出力文がフジモリ大統領に関連する内容であるが，フジモリ大統領は図 4 を見てわかる通り，活性拡散の結果，最も活性値の高かった語である．おそらく，イベントの重要度を (動詞の活性値) + (名詞，目的語等の必須格の活性値) と定めた為，最も活性値の高かったフジモリ大統領を必須格に取る文ばかりが出力されたものと考えられる．この方法も，結束性の維持という意味では，ある一定の効果はあるかと思うが，実際にはネットワークの隣接性等を考慮に入れ，もう少し厳密に処理を行う必要がある．

(7) 網羅性

事件の一連の流れを把握する上で，できれば満遍なく様々な箇所を抽出する必要があると考えられるが，2 月 17 日から 2 つ，4 月 22 日から 3 つのイベントを選択しており，かなり偏在した内容を出力している．同じ日付のイベントを選択した場合にはペナルティを与えるなどの処理を行うべきかもしれない．

4 結論と今後の課題

本稿では，アノテーション付きの多文書データから，コンテンツネットワークを構築し，文生成に近い形で要約を出力するシステムを提案した．このシステムをペルー日本大使公邸占拠事件を扱った一連の記事集合に用いた結果，現時点で高い圧縮率が見込めるものの，文章の可読性や重要度算出法等にまだ多くの

課題が残されていることがわかった．

今後の課題としては，重要度評価方法の再考や文生成アルゴリズムの改善と共に，現在ほとんど考慮していない句読点の処理，文の連結処理等も必要である．さらに，評価方法として，アノテーションを利用しない手法との比較や，ペルー日本大使公邸占拠事件以外の記事集合に対する実験等も考慮していかねばならない．

参考文献

[1] 上田良寛, 岡満美子, 個山剛弘, 宮内忠信. 句表現要約手法に基づく要約システムの開発と評価. 自然言語処理, Vol. 9, No. 4, pp. 75–96, 2002.

[2] R.F.Lorch. Priming and searching processes in semantic memory: A test of three models of spreading activation. *Journal of Verbal Learning and Verbal Behavior*, pp. 468–492, 1982.

[3] D.A.balota and R.F.Lorch. Depth of automatic spreading activation: mediated priming effects in pronunciation but not in lexical decision. *Journal of Experimental Psychology: Learning, Memory, Cognition*, pp. 336–345, 1986.

[4] 阿部純一, 桃内佳雄, 金子康朗, 李光五. 人間の言語情報処理. サイエンス社, 1994.

[5] M.R.Quillian. Semantic memory. *Semantic information processing*, pp. 227–270, 1968.

[6] A.M.Cillins and E.F.Loftus. A spreading activation theory of semantic processing. *Psychological Review*, pp. 407–428, 1975.

[7] J.R.Anderson. A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*, pp. 261–295, 1983.

[8] 松村真宏, 大澤幸生, 石塚満. 語の活性度に基づくキーワード抽出法. 人工知能学会論文誌, Vol. 17, No. 4, pp. 398–406, 2002.

[9] E. Bloedorn Mani, I. Multi-document summarization by graph search and matching. In *Proc. of AAAI-97*, pp. 622–628, 1997.

[10] 長尾確, 白井良成, 橋田浩一. 言語的アノテーションに基づくマルチメディア要約. 言語処理学会第 6 回年次大会発表論文集, pp. 380–383, 2000.

[11] 伊藤誠悟, 橋田浩一, 宮田高志. Gda 文書を用いた複数文書要約. 言語処理学会第 8 回年次大会発表論文集, pp. 555–558, 2002.