アノテーション付き多文書データからの要約生成

Summarization of Annotated Multiple Documents

綾 聡平† 宮田 高志‡ 橋田 浩一‡ 石塚 満†
Sohei AYA† Takashi MIYATA‡ Hasida KÔITI‡ Mitsuru ISHIZUKA†

東京大学大学院情報理工学系研究科†
School of Information Science and Technology, University of Tokyo†
産業技術総合研究所サイバーアシスト研究センター‡
Cyber Assist Research Center, National Institude of Advanced Industrial Science and Technology‡

1 はじめに

1996年から1997年にかけて，ペルーの日本大使公邸で武装ゲリラが人質を取り，立てこもった事件があったことを覚えているだろうか．このペルー日本大使公邸占拠事件は解決までに数ヶ月を要しており，事件の顛末を調べるには，恐らく数十から数百程の記事を読む必要があるだろう．しかし，それら全てに目を通すのは多くの人にとって，非常に大きな負担である．自動要約は，このような手間を軽減する技術として期待されている．

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）が有力であったが，文単位の抽出で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問題があった．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く，これにより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．そこで，重要語句を抽出し，それを元に文章生成に近い形で要約するシステムの必要性が高まっている．

本稿では，複数のGDA（Global Document Annotation）文書から生成されるコンテンツネットワークに対し，活性拡散を用いて各々のノードに重み付けを行い，その重要度を元に文章生成に近い形で要約を行う手法を提案する．

2 大域文書修飾 GDA

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，詳細については省略する[1]が，XML（Extensible Markup Language）タグセットで与えられ，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ことにより，計算機が高精度で自然言語を処理す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，[1, 2]等がある．GDA文書の例を図1に示す．

[1]詳細は http://www.i-content.org/gda

```
<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1 GDAに基づくアノテーションの例

3 要約の流れ

ここで提案する手法は，大凡このような流れを踏んで要約を生成する．

1. ネットワークの生成
2. 活性拡散に基づく重み付け
3. 要約生成

以下に，これらの詳細について概説する．

3.1 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，それを元にネットワークを生成する．例えば「日本大使公邸に武装ゲリラが乱入」という文があった場合，係り受け関係により「武装ゲリラ」が「乱入」の主語，「日本大使公邸」が目的語であることがわかる．さらに，「トゥパク・アマル」と「武装ゲリラ」が共参照関係にあることがアノテーションによる情報で与えられていれば，これらは同一とみなす事ができる．

このようにして記事集合の全ての文を統合し，巨大なネットワークを生成することができる．また，このようにしてネットワークを生成すれば，同じような内容の文は縮退し，冗長性を省くことができると考えられる．

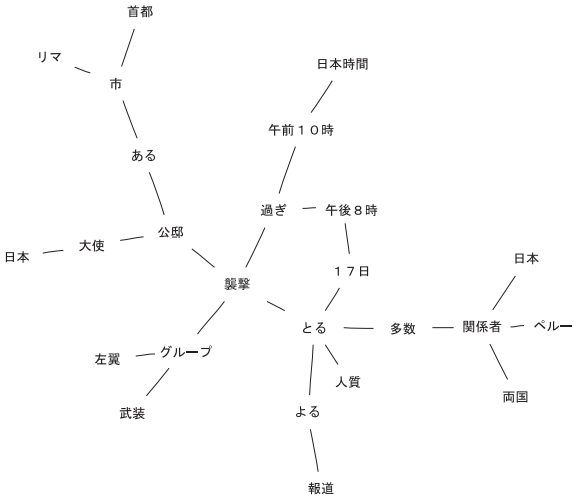

図 2　ネットワーク例

尚，リンク強度は

照応・共参照関係 > 係り受け関係 > その他

と定めており，さらに関係数が多い程強度が強くなるように設計している．また，機能語はノードとしては扱わない．ネットワークの一例を 図 2 [2] に示す．

3.2　活性拡散に基づく重み付け

第 3.1 節で作成したネットワークに対し，活性拡散を，活性値が収束するまで繰り返す．活性拡散は以下の式で表される．

$$X_i\,(t+1) = \sum_{j=1}^{n} A_{ji} X_j\,(t) + C$$

ここで n はネットワークに於けるノード数，A はネットワークの接続行列，$X(t)$ は各ノードの活性値，C が定数である．定数 C には，$TF * IDF$ に基づく値を用いている．

3.3　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. 活性拡散の結果に基づき，活性値上位（個数は要約率に依存する）のものを重要語と考える．
2. 重要語の中から動詞を取り出し，これらをそれぞれの文の中心と考え，出力ノードに加える．
3. 出力ノードと接続している主語，目的語等の必須格，もしくは重要語を出力対象に追加する．
4. 順番に遡りながら 3. の処理を繰り返す．尚このとき，既に出力された文中に同じ修飾が存在する場合には，枝刈りを行う．

さらに，このようにして生成された文たちをイベントの発生した日付順に出力する．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，日時も出力に加えてやる．

また，これだけでは可読性に欠ける場合があるので，結束性を保つ為に次のような処理も行う．

1. ある要約文 B を出力する際には，直前に出力された文 A 中に含まれるノード集合と連結しているかどうかを調べる．
2. B が A に連結していなければノードを順番に探索し，A と隣接するノードまでを B の出力ノードに追加する．

4　実験と考察

1996 年に発生したペルー日本大使公邸占拠事件に関する 50 記事に GDA タグをつけたものを要約対象に，実験を行っている．紙面の都合上，本稿への結果の掲載は省略するが，要約結果を見ると，提案手法は文を生成するのに近い手法であるので，要約率が比較的高くても，事件の大まかな流れが把握できるようになっている．しかしその反面，現状では可読性がやや低い．特に自立語間を結ぶ機能語は，原文を元に補完している為，ノードを切り貼りしてしまうと，繋がりの悪くなる箇所が多くなってしまう．

5　まとめと今後の課題

本稿では，複数の GDA 文書からコンテンツネットワークを作成し，活性拡散を用いて要約を生成する手法を提案した．

今後の課題としては，可読性を向上させる為の文生成アルゴリズムの改善が最も重要だと考えている．また現在，ペルーの記事のみでしか要約を行っていないので，他の新聞記事等にも要約対象を広げ，その汎用性についても調査する必要がある．その他，生成された要約の評価方法についても今後検討する予定である．

参考文献

[1] 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.

[2] 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.

[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，左翼ゲリラと見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.」をネットワークで表現したものである．尚，この図ではリンク強度は表現していない