GDA文書を用いた複数文書要約

伊藤 誠悟† 橋田 浩一‡* 宮田 高志*

東京理科大学大学院†
産業技術総合研究所‡ 科学技術振興事業団 (CREST)*

1 はじめに

単一文書の要約を行う場合はしばしばにもとの文書に含まれる文を選択する方法が用いられるが, 複数文書の要約を行う場合には結束性を持たせるために文章生成に近い方法を用いる必要がある. また単一・複数文書要約どちらの場合においても要約を生成する際, 利用者により背景知識や興味が違うため利用者ごとに適合した要約, つまりパーソナライゼーション (個人化) された要約の生成が必要である. このような問題に対し本稿では複数のGDA文書から生成される意味ネットワークより, 元の文を参照しながら要約を生成する方法を論ずる. ネットワークは依存関係と意味的参照関係のリンクを持ち, 共参照しているノードは同一ノードとみなすネットワークである. このネットワークに活性拡散を行い全ノードの重要性を評価し, それに基づき要約の生成を行う. パーソナライゼーションに関しては活性拡散の際に活性値を変化させることにより利用者の興味のある部分の重要性を高めそれぞれに適応した要約を生成する. 本研究以外で現在までのネットワークを用いて要約を行う試みとして次のようなものがある.[4][5][9][10]

2 大域文書修飾 GDA

GDA (Global Document Annotation)[1][3] はそのうちの統語照応構造, 修辞構造, 対話構造, 語義などをタグによって明示的にアノテーションするためのXMLタグセットである. GDAを用いることにより文書の意味を計算機が理解可能とし, その情報を利用して応用を高精度で実現することができる. 現在までにGDAを利用した研究報告として回答抽出 [8], 要約 [6][7][11], プレゼンテーション [12], タグ体系変換 [2] などがある.

3 複数文書の要約

3.1 GDAによるネットワーク

GDAタグを含む文書は意味ネットワークに変換できる。

```
<su syn="f">
  <adp opr="obj">
    <placenamep id="jpn">日本</placenamep>
    <n id="tagid00016">大使</n>
    <n id="tagid00017">公邸</n>
    <ad>に</ad>
  </adp>
  <adp opr="agt">
    <np opr="obj" id="tagid00018">
      <n>武装</n>
      <n>ゲリラ</n>
    </np>
    <ad>が</ad>
  </adp>
  <v>乱入</v>
</su>
```

例えば上記のようなGDA文書では「日本大使公邸に」と「武装ゲリラが」が、それぞれ動作対象を示すobjリンクと、行為の主体を示すagtリンクにより「乱入」と結ばれている。 上記のようなGDA文書から3.2で述べるコンテンツネットワークを生成する前準備として図1に示すような意味関係によるネットワークの変換を行う. この変換には開発したGDA Parser for Javaを用いる.

3.2 コンテンツネットワーク

対象となる各GDA文書は図1で示したネットワークを構成している. 各文書のネットワークと文

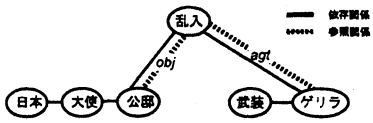

図 1: 意味関係のネットワーク

書間の共参照情報から 1 つの巨大なネットワークを生成する．文書間の共参照情報は各 GDA 文書と独立して次のように記述される．relation タグがひとつの共参照集合を表し，source 属性に構成要素となるノードのファイル名とファイル内での ID 名を記述する．

```
<relation source="9704/13/05.gda#tagid00035
                  9704/07/02.gda#tagid00017
                  9612/19/07.gda#tagid00003
                  9703/19/09.gda#guerilla"/>
```

このような情報から構成されるネットワークをコンテンツネットワークと呼ぶ．コンテンツネットワーク上では共参照リンクや eq (等価を表す関係子) リンクで結ばれるノードは同一ノードとみなされる．例えば図 1 のネットワークの他に「トゥパクアマルがグァテマラ大使を解放」という文がありこの文中の「トゥパクアマル」と図 1 中の「ゲリラ」が共参照として記述されていた場合 2 つのノードが融合され図 2 のようなネットワークを構築する．融合された後の共参照ノードは複数のラベルを保持している．

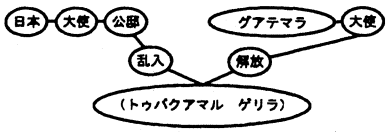

図 2: コンテンツネットワーク

3.3 活性拡散による重要性の評価

3.2 にて生成したコンテンツネットワークに対し活性拡散を行って各ノードの重要性を評価する．活性拡散とは，各ノードの活性値（重要度）を係り受け関係や参照関係およびその他さまざまなリンクを通じ伝播させていき重要性を評価する手法である．この手法の有用性は単一文書要約においては [5] により証明されている．本稿のコンテンツネットワークに対して活性拡散は次のように行う．

コンテンツネットワークの全ノード数が n 個である場合，そのネットワークの接続行列を A，各ノードの活性値を X，外部入力値を C とする．このときの漸化式は次のようになる．

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

これにより各ノードの活性値を得る．

上記のように，活性値を求めることによって各ノードの重要性を評価する．

3.4 要約生成アルゴリズム

はじめに以下で使用する語句の定義を行う．

重要ノード：活性値上位 N 個までのノード

核ノード：重要ノード中のサ変名詞，動詞ノード[1] 上位 M 個のもの

生成ノード：要約において発話されるノード

核ノードにおける M の数は，要約の出力文字数により数値を変化させる．例えば「200 文字の要約を生成する場合であるならば M は 10 が妥当である」のように決定する．3.3 における活性拡散の計算終了後，以下のようなアルゴリズムにより要約の生成を行う．

1. 活性拡散の結果より，活性値上位ノードから重要ノードを決定する
2. 重要ノードから核ノードを決定する
3. 核ノードより次の規則に従い生成ノードを形成する．この生成ノード形成は閉包演算である．
   - 核ノードと直接結ばれた重要ノードを生成ノードに追加する
   - 指定スコア以上のノードでなくても agt,sbj,obj などで指定される必須格のノードは生成ノードに加える．
4. 形成された生成ノードをイベントの発生した日付順に出力する

[1] 各ノードは Juman の解析結果を保持しているためそれにより判別

3.5 結束性

上記アルゴリズムにしたがって各イベントを発話し要約を生成する際に，要約としての結束性を保つ必要がある．そのため発話するイベント間で次のような処理を行う．以下の説明ではこれから発話するノード集合をB, Bの前に発話されたノード集合をAとする．

- あるイベントノード集合Bを発話する際には直前のイベントノード集合Aと連結しているかどうかをチェックする．
- BがAに連結していなかった場合は発話するイベントノード集合Bの各ノードのリンクを1つずつさかのぼっていき直前のイベントノード集合Aと連結する部分を探索する．
- 連結する部分が見つかったらその部分までののノードを発話するイベントノード集合Bに追加する
- Bを発話し，次のイベントも同様に処理する

3.6 共参照，日付

共参照グループのノードまたは日付ノードは要約として発話される場合はいくつかの規則により言い換えが行われる．共参照グループのノードに対しては一番最初に発話されるときはラベルの最も長い表現が選択され次からは最も短い表現を選択する．日付ノードに関しては年の変わり目月の変わり目などは詳しく発話しそうでないときは日付しか発話しない．

4 実験及び結果

4.1 実験データ

要約対象文書として1996年に発生した「ペルー日本大使公邸占拠事件」の関連記事群を用いる．この記事群は1996年12月から1997年4月まで4ヶ月にわたる新聞記事50記事から構成され「事件発生」「犯人グループとの交渉」「日本の対応」「人質家族関連」「事件の解決」などさまざまなトピックが含まれている．この文書は現在利用可能な自然言語ツールによる自動解析と人手を両方用いてGDAタグが付与されている．

4.2 要約文書の生成

4.1で示した対象文書に活性拡散を行い3.4のアルゴリズムを適用した結果を示す(イベント数は10，活性値に対して利用者の外部入力はなし)．ゴシック体の部分は核ノードである．

日本大使公邸が/1996 年 12 月 17 日に/「トゥパク・アマル革命運動」(MRTA) に/**襲撃**され/とられた。// リマからの報道によると/人質をとって立てこもったゲリラのメンバーは/18 日に/外国人記者の代表を通じてペルー政府に要求書を送り/服役中の一部ゲリラの釈放運動をしているランシェ神父ら2人を仲介者として派遣するよう/求めた。//ゲリラによる人質事件で/政府は/ペルー政府に対し人質の安全確保を要請するとともに/外務省の堀村隆彦中南米局審議官を/同夜/派遣した。// 19 日に/池田外相はフジモリ大統領を訪問し4 5 分間/ゲリラの要求と、大統領の対応に関して/会談した。//事件で/ペルー政府とゲリラの交渉舞台となる保証委員会のシプリアニ司教ら3人は/1997 年 2 月 6 日に/公邸内でゲリラ側と/会談/打診した。//日本大使公邸を占拠するゲリラとペルー政府双方のボイコットで延期されていた/予備的第1 0回対話/終了。// ペルー政府とゲリラの交渉仲介役を務める保証委員会は/4 月 5 日に/司教らが公邸に入り/ゲリラと非公式に/接触した。//フジモリ大統領は/司教ら関係者と/連日のように会談して/示している。// 外相が/大統領府でフジモリ大統領と会談、/日本大使公邸占拠事件の/解決について/謝意を伝えた。//事件で/人質となり救出された青木盛久・駐ペルー大使は/28 日に/外相と、公邸から救出された大使館員の木本博之公使、三村晴夫1等書記官兼医務官らとともに/羽田着の特別機で/帰国した。

複数文書要約の評価基準は確立していないため以下では実行例に対し定性的な評価を行う．この要約を見てみると上位10イベントの中に，事件の開始，武装ゲリラとの会談，事件の解決について，一連の事件の重要と思われる出来事が抽出され要約に含まれているのがわかる．3.5の結束性の処理が働いている部分の1つを詳説する．第4文の「19日に池田外相はフジモリ大統領を訪問し4 5分間ゲリラ側の要求と、大統領の対応に関して会談した」は結束性の処理をしなければ「19日に池田外相はフジモリ大統領を訪問し4 5分間会談した。」といなる，しかしこの文は前文の「発生したゲリラによる人質事件で政府はペルー政府に対し人質の安全確保を要請するとともに外務省の堀村隆彦中南米局審議官を同夜派遣した。」とのコンテンツネットワーク中でのノードの重なりがないため3.5で述べた結束性処理が働く．すると「会談」ノードから繋がりかつ全文とノードの重なりがある部分に「ゲリラ側の

要求と、大統領の対応に関して」を発見する（「ゲリラ」が重なり部分）．この発見したものを加えて発話することにより結束性を保っている．

4.3 パーソナライゼーション

前節での要約は利用者が何も指定しないで，活性拡散を行った結果の要約である．ここでは利用者の興味にあった要約，つまりパーソナライゼーションされた要約を示す．例えば利用者が「人質にとられている人の中にペルー最高判事がいて，その子供の名前がウゴ・シビナ君である。」という情報をもっていて「ウゴ・シビナ」について詳しく知りたいと仮定する。その場合コンテンツネットワーク上の「ウゴ・シビナ」と対応しているノードの外部入力を高くする。その結果「ウゴ・シビナ」と付近のノードの活性値が高まり，「ウゴ」と関連した出来事が重要ノードとなり要約で生成される。以下は「ウゴ・シビナ」の外部入力値を高めた場合の要約結果である．ゴシック体の部分は日付と共参照の言い換えの一部である．

> ウゴ・シビナが 1997 年 2 月 3 日に橋本龍太郎首相あてに助けを求める手紙を書いた。「父と一緒に誕生日を迎えたい」とつづる。2 月 5 日に首相が手紙を読んだことに「とってもうれしい。人質の父を持つほかのペルー人の子供たちの励みにもなると思う」と喜びの声を上げた。「父と誕生日を迎えたい」と手紙を書いたウゴ。首相が「返事を書きたい」と語ったことを知ると「うれしい。人質の家族の中にはまだ手紙を書けない小さい子供たちがたくさんいるので、役に立てた」と話し喜んだ。

5 おわりに

本稿では，意味関係ネットワークと活性拡散を用いて重要性の評価を行い，複数文書の要約を行う手法を提案し実験した．複数文書の要約を評価するための統一的なテストベッドは存在しないため，現在はこの 50 記事に対してしか提案手法を試みていない．しかし実験結果からわかるように，提案手法により複数文書の要約が一般的に適用でき一連の重要な流れを獲得できる．ノード単位での結束性の処理を行うことによりある程度要約の頑健性が保たれている．しかし要約例を見てわかるようにノード単位だけでの処理では不足であり要約の頑健性を保つ他の方法が必要である．また利用者の興味をインタラクションを行い活性値に反映させることで獲得し，利用者の好みに合わせた柔軟な要約文が生成されることを示した．

参考文献

[1] GDA - 大域文書修飾 http://www.i-content.org/gda/

[2] 伊藤 誠悟, 白松 俊, 横山 憲司, 山本 浩司, 橋田 浩一, 奥乃 博. "2 種類の意味構造化コンテンツの体系 GDA と UNL 間での自動変換" 第 11 回 AI チャレンジ研究会, SIG-Challenge-01-11, pp.47-54, 人工知能学会, 2001.

[3] Kôiti Hasida. "Global Document Annotation." In NLPRS'97, pp505-508. 1997

[4] Inderjeet Mani, Eric Bloedorn Multi-document Summarization by Graph Search and Matching. In Proc. of the 14th National Conference on Artificial Intelligence, 622-628 1997.12

[5] Kôiti Hasida, Syun Ishizaki, and Hitoshi Isahara. 1987. A connectionist approach to thegeneration of abstracts. In Gerard Kempen,editor, Natural Langauge Generation: New Results in Artificial Intelligence, Psychology, and Linguistics, 149-156. Martinus Nijhoff.

[6] 長尾 確, 橋田 浩一, 宮田 高志 GDA タグを用いた文書の要約に関する一考察. 自然言語処理シンポジウム「実用的な自然言語処理にむけて」電子情報通信学会「言語理解とコミュニケーション」研究会 1997

[7] 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, pp380-383

[8] 鈴木 潤, 橋田 浩一. "GDA タグを利用した回答抽出システムの提案" 言語処理学会第 7 回年次大会発表論文集, pp313-316. 2001

[9] 豊浦 潤, 津高 新一郎, 瀬尾 和男, ネットワークを用いた複数テキストの要約方式の提案. 言語処理学会第 5 回年次大会発表論文集 225-226 1999

[10] 上田 良寛, 小山 剛弘, 共通意味断片の抽出による複数文書要約. 言語処理学会 第 6 回年次大会 発表論文集 360-360 2000

[11] Masao Utiyama, Koiti Hasida, "Multi-Topic Multi-Document Summarization", Proceedings of the 18 th International Conference on Computational Linguistics , 2000.

[12] 内山 将夫, 橋田浩一. "GDA タグを利用したインタラクティブなスライド生成" 「言語資源の共有と再利用」シンポジウム 1999