GDAタグを利用した複数文書の要約

内山 将夫
通信総合研究所

橋田 浩一
電子技術総合研究所

1 はじめに

大規模な出来事は，通常，複数の出来事から構成される．それらの出来事は，互いに関連しているのは確かではあるが，一つのトピックになるほどではない．たとえば，戦争は，開戦，戦闘，交渉などからなるが，これらは，それ自体がトピックであるため，それ自体が一つの文書となりうる．

そのような大規模な出来事の要約，すなわち，複数のトピックに関する複数の文書の要約が本稿の主要な目的である．このようなことは従来は試みられていなかった．従来の研究 (McKeown and Radev 1995; Barzilay, McKeown, and Elhadad 1999; Mani and Bloedorn 1999) では，一つのトピックのみに関する複数文書を要約することを目標としていた．

複数のトピックを含む複数文書を要約するためには，形式的な情報よりも意味的な情報の方が重要である．なぜなら，複数文書に渡るような一貫した文書構造というものは存在しないため，形式的な情報を複数文書の要約に利用することはできないからである．このことは，単一の文書を要約する場合とは対照的である．単一文書の要約においては，たとえば，新聞記事のように極端な場合には，第1段落だけをとれば要約として充分なこともある．しかし，複数文書においては，そのようなことは不可能であるので，意味的情報を利用する必要がある．

そのような意味的情報を利用するために，本稿では，活性拡散を利用する．そして，活性拡散の結果に基づいた重要度を利用することにより，複数トピックに関する複数文書を要約することを試みる．要約のための主要な方法は，「重要文書の抽出」と「重要語句間の関係の表示」である．このとき，抽出された重要文書は個々に要約される．さらに，重要語句間の関係を含むような文集合は，全文書に渡る簡潔な要約となる．なお，重要語句間の関係の表示は，情報検索の分野で研究されているが (Niwa, Nishioka, Iwayama, Takano, and Nitta 1997; Sanderson and Croft 1999)，複数文書の要約という観点からの研究は我々の知るかぎりは存在しない．

本稿で要約の対象とする文書は，GDA (Global Document Annotation) タグ (Hasida 1997; Nagao and Hasida 1998) によりタグ付けされたものである．本稿で述べる方法は，基本的には，GDAタグの情報のみを利用しており，GDAタグは，個別言語に依存しないようにデザインされているため，提案手法は，本質的に汎言語的である．タグ付けされた文書を用いるもうひとつの利点は，要約研究における文書解析の段階を，その後の段階と切り分けることにより，研究の対象を要約研究における後半の段階 (たとえば生成) に絞ることができることである．更に，タグ付け文書は，要約研究だけでなく，翻訳などの他の自然言語処理のタスクにおける共通入力ともなりえる．

2 Global Document Annotation

GDAは，言語的なタグを電子化文書に付けることにより，文書の意味を計算機により理解可能にするとともに，それを利用した内容指向のプレゼンテーション，質問応答，要約，翻訳などを高精度で達成するような自然言語処理技術を実現することを目的とするプロジェクトである．そうすることにより，GDAは，電子化されたコンテンツのプレゼンテーションや再利用のための統合的なプラットフォームとなりうる．

現状の自然言語処理技術では，GDAタギン

グには人手の介入が必要であるが，このときのコストを上回るだけの利益がある．なぜなら，GDA タグによりタグ付けされた文書は，要約だけでなく，多用途に再利用可能であるからである．

GDA タグセット[1] は，XML (eXtensible Markup Language) の一種である．GDA タグセットが表現しうる言語的な情報は，構文関係，照応，語義，スコープ，発話行為など様々である．これらのなかで，本稿の目的にとって必須なのは，構文情報と照応の情報である．構文情報は，文書の構成要素 (エレメント) 間の統語関係を表現するものである．また，照応の情報には，共参照，代名詞照応，零代名詞照応などがある．

3 複数文書の要約

3.1 活性拡散

GDA タグが付けられた複数文書は，図 1 に示すようなネットワークを構成する．ネットワークのノードは，GDA エレメントを示し，リンクは，構文あるいは意味情報を示す．このネットワークは，GDA エレメントからなる構文木に加えて，エレメント間の参照関係を表すリンクからなる．

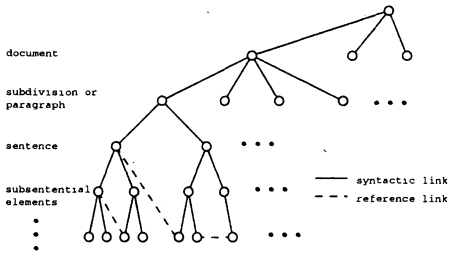

図 1: Multi-document network.

ネットワークの各ノードの重要度は，活性拡散を利用して求める．活性拡散は，単一文書の要約に利用され，その有効性が確かめられている (Hasida, Ishizaki, and Isahara 1987; Nagao and Hasida 1998) が，その効果は，単一文書よりも複数文書において高いと予想される．なぜなら，1 章でも述べたように，複数文書においては，文書構造に頼った要約は実現できず，必然的に，意味的な重要性を考慮する必要があるためである．

活性拡散は，共参照関係にあるエレメント間における活性値が同一になるという制約のもとで行われる (Hasida et al. 1987)．活性拡散終了後のノード i の活性値を $a(i)$ とすると，そのスコア，すなわち，ノード i の重要度は，

$$score(i) = a(i) + \sum_{j \in ch(i)} score(j) \quad (1)$$

である．ここで，$ch(i)$ は，構文木におけるノード i の子供のノード集合である．ただし，葉ノード (単語) においては，子供のノード集合は存在しない．

3.2 重要文書と重要文の抽出

重要文書の抽出は，文書に対応するノードのうちで，活性拡散により求められたスコアが高いものを選ぶことにより行なわれる．このとき，文書の長さに対するスコアの正規化 (単語数でスコアを割るなど) は行なわない．なぜなら，一般に，本稿の対象である新聞記事では，記事の長さはほぼ同等であり，かつ，長い記事の方が，重要な場合が多いからである．また，文書の長さで正規化した場合には，短い記事が極端に優先される場合があるからである．それに対して，正規化しない場合には，長いものを優先しつつ，短いものでも重要なものはとる，というようになるからである．しかし，このことは対象とする文書集合により異なる可能性がある．

抽出された重要文書からは，重要文が同様な手法により抽出され，必要な部分を除いて枝刈りされる．枝刈りの手続きは概略以下の通りである．

1. スコアの低い従属節は除く．
2. 主節における必須要素以外は除く．
3. 必須要素における深さ 2 以上の修飾節を除く．

[1] http://www.etl.go.jp/etl/nl/GDA/tagset.html

4. 照応先が重要文中に選ばれていないような代名詞は除く．

ここで，枝刈りには構文タグの情報を主に利用するが，必須格の判定には，格助詞などの表層情報を利用している．また，照応についてはタグを利用する．なお，照応については，単に削除するのではなく，照応先と置き換えることもできる．

提案手法の有効性を示すために，「ペルー日本大使公邸占拠事件」についての新聞記事の集合について，提案手法を適用した．これらの記事群は，1996年12月から1997年4月までの，4ヶ月におよぶ50記事からなり，事件の開始から，ゲリラ側との交渉，事件の解決まで，様々なトピックを含んでいる．

これらの記事群における，最も重要な記事としては，事件の開始(トゥパク・アマルによる日本大使公邸への襲撃)記事および事件の解決(ペルー政府による制圧)記事がある．これら二つの記事は，提案手法が事件の時間関係を陽には取り扱っていないにもかかわらず，上位4記事以内にランクされていた．このことは提案手法の有効性を示す．なお，事件開始の記事は，50記事中の1番目の記事だが，事件解決の記事は，50記事中で最後から6番目の記事である．したがって，単に最初と最後の記事を抽出するという単純な方法では，事件解決の記事を抽出することはできない．

これら二つの記事の25%分に相当する要約を以下に示す．

> ○日本大使公邸に武装ゲリラ、パーティーに乱入　銃撃、200人が人質－－ペルー
>
> 日本、ペルーの両国関係者多数が人質にとられた。武装グループは約20人で、うち複数が公邸に押し入った。現在、散発的に銃撃戦が展開されているという。
>
> ○ペルー日本大使公邸占拠事件　人質全員解放　権力基盤回復狙い－－フジモリ大統領
>
> フジモリ大統領は、強気の政治家であることを、日本大使公邸占拠事件の解決でみせつけた。フジモリ大統領は公邸敷地に入った。公邸訪問は、作戦の陣頭指揮を執っていることを印象付けた。なぜ、フジモリ大統領は武力行使を選択したのか。政治危機の根源にある公邸事件を、武力で解決することで、政治主導権の回復を狙ったといえる。

3.3 実体関係図

複数文書中における実体iと実体jとの関係についてのスコア$score(i,j)$は以下のものである．ただし，実体とは，文書中における語句が指示するものをいう．

$$score(i,j) = |E(i)|a(i) + |E(j)|a(j) + \sum_{s \in S(E(i)) \cap S(E(j))} score(s) \quad (2)$$

ここで，$E(i)$は，実体iを指示するようなGDAエレメントの集合であり，$a(i)$は，実体iの活性値である．また，$S(E(i))$は，$E(i)$中のエレメントを支配するような文の集合である．式中において，$|E(i)|a(i)$としているのは，実体iの重要度に加えて，頻度も考慮するためである．これは，情報検索で良く知られた尺度である，tf・idfとのアナロジーからこうした．

$score(i,j)$は，実体iと実体jの関係の強さを示すものであるが，それだけでなく，この値が大きいことは，$S(E(i)) \cap S(E(j))$，つまり，実体iとjとを含むような文集合は，文書に渡る簡潔な要約となりうることを示している．なお，共参照の連鎖に着目した要約は，単一文書については，(Azzam, Humphreys, and Gaizauskas 1999)で試みられている．

実体関係図は(2)式の値が上位のものから作られる．図2は，前述の50記事から作られた，上位11個の関係についての，実体関係図である．この図は，事件の全体を良く表現していると言える．

図2において，最高スコアの関係は，「ペルー」と「日本大使公邸」の関係である．これらを含む文を複数記事から抽出し，その一部をパラフレーズして示すと以下の通りである．なお，下線部がパラフレーズされた部分である．

1. 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1996年12月17日、左翼ゲリラ(トゥパク・アマル)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。
2. その人質事件で政府は12月18日、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、現地に派遣した。
...

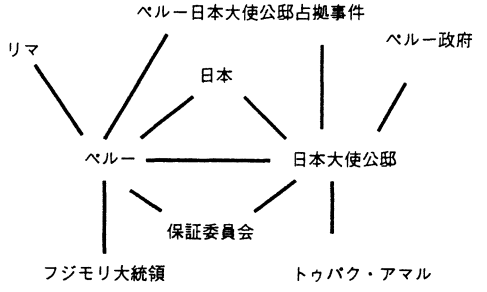

図 2: E-R graph of Peru hostage incident.

3. 1997年4月22日、ペルーの日本大使公邸への突入作戦を成功させたことで、フジモリ大統領の政治的威信は再び回復に向かうことになるだろう。

上記において，日付表現は，もともとは「17日」などであるが，それらは「1996年12月17日」などに書換えられる．また,「左翼ゲリラ(トゥパク・アマル)」は，もともとは,「左翼ゲリラとみられる武装グループ」という表現であったが，文書間に渡る共参照タグを利用して，より具体的な表現に書換えられた．このような書換えは，新聞記事の集合のように，徐々に情報が明らかになっていくような場合には必須である．更に,「その人質事件」は，もともとは「ペルーの日本大使公邸で発生した武装ゲリラによる人質事件」という表現であったが，この名詞句と第1文とが同一イベントであることが名詞と動詞の項構造から検出可能であるので,「その人質事件」というように書換えることができる．

4 おわりに

複数文書の要約を評価することは，単一文書の要約を評価することに比べて非常にコストが掛かり，そのためのテストベッドはまだ存在しない．そのため，現在のところは，ただ一つの記事群のみでしか提案手法を試していない．しかし，その結果は，提案手法が，複数文書の要約に対して，一般的に適用可能であることを示唆していると考える．

提案手法により，一連の記事の中から，事件の開始と解決を報じる最も重要な記事を抽出できたし，また，事件の鍵となる重要な実体とその関係も抽出できた．更に，文書間の共参照を利用すれば，曖昧な表現を具体的な表現に置き換えることができた．

これらは，GDAタグの情報のみを利用して実現可能である．そして，GDAタグは，ドメインやスタイルや言語に独立なものであるので，提案手法も，ドメインやスタイルや言語に独立に，複数文書から重要文書を抽出し，実体関係図を作成することができると考える．更に，GDAタグでタグ付けされた文書は，要約だけでなく，翻訳などの他の自然言語処理のタスクにおける共通入力ともなりうるであろう．

参考文献

Azzam, S., Humphreys, K., and Gaizauskas, R. (1999). "Using Coreference Chains for Text Summarization." In *ACL'99 Workshop on Coreference and Its Applications*, pp. 77–84.

Barzilay, R., McKeown, K. R., and Elhadad, M. (1999). "Information Fusion in the Context of Multi-Document Summarization." In *ACL'99*, pp. 550–557.

Hasida, K. (1997). "Global Document Annotation." In *NLPRS'97*, pp. 505–508.

Hasida, K., Ishizaki, S., and Isahara, H. (1987). "A connectionist approach to the generation of abstracts." In Kempen, G. (Ed.), *Natural Langauge Generation: New Results in Artificial Intelligence, Psychology, and Linguistics*, pp. 149–156. Martinus Nijhoff.

Mani, I. and Bloedorn, E. (1999). "Summarizing Similarities and Differences Among Related Documents." In Mani, I. and Maybury, M. T. (Eds.), *ADVANCES IN AUTOMATIC TEXT SUMMARIZATION*, chap. 23, pp. 357–379. The MIT Press.

McKeown, K. and Radev, D. R. (1995). "Generating Summaries of Multiple News Articles." In *SIGIR'95*, pp. 74–82.

Nagao, K. and Hasida, K. (1998). "Automatic Text Summarization Based on the Global Documentation." In *COLING-ACL'98*, pp. 917–921.

Niwa, Y., Nishioka, S., Iwayama, M., Takano, A., and Nitta, Y. (1997). "Topic Graph Generation for Query Naviagation: Use of Frequency Classes for Topic Extraction." In *NLPRS'97*, pp. 95–100.

Sanderson, M. and Croft, B. (1999). "Deriving concept hierarchies from text." In *SIGIR'99*, pp. 206–213.