共通意味断片の抽出による複数文書要約

上田良寛　小山剛弘
富士ゼロックス株式会社
{Ueda.Yoshihiro, koyama.takahiro}@fujixerox.co.jp

1 はじめに

文書分類、特にクラスタリング[1]に関する研究は広く行われているが、分類された結果それぞれのクラスターが何を意味しているのかを提示する概要提示技術の研究は遅れており、キーワード列挙以上の有効な方法は示されていない。全体の傾向を判断するには、結局分類された個々の文書を見ていく必要が生じる。

複数文書の概要を示す技術があれば、傾向の把握が容易になる。また、検索結果が大量になる場合、クラスタリングを行って、個々のクラスターの概要が提示されれば、クラスターごとにふるい分けを行うことができ、効率的な情報入手プロセスを支援することができる。

このように複数文書要約、特に「ある性質で集められた文書集合の要約」は、大量の情報が氾濫する時代に求められている技術であろう。我々は、共通意味断片を抽出することによって複数文書の要約を与える方法を開発し、実装を進めているので、ここで報告する。

2 複数文書要約の既存手法

キーワード列挙以外の複数文書で用いられる要約技術としては以下のようなものがあげられる(キーワード列挙とその拡張に関しては次章で述べる)。

代表文: 文書分類技術の Scatter/Gather (Cutting et al. 1992)では、分類されたクラスタに対し、中心に近い数文書から見出しを 20 文字づつ切り出ししたものと、クラスタ中で現れる頻出単語を列挙したものを、ラベルとして与える。後者はキーワード列挙であるが、前者は代表文ということができる。代表文は文書群の代表である文書から、その文書の代表である文を切り出したものであり、文書群を代表して文書群の性質を表現するものとは言い難い。

抽出された意味から文を生成するもの: SUMMONS (McKeown and Radev 1995)は、あらかじめ用意したテンプレートのスロットを複数文書から抽出した情報で埋めたものを意味構造とし、その意味構造からあるパターンにあった文を要約として生成するものである。対象となる文書が強く限定される。たとえばここでは、テロリストが起こした事件に対する記事(「誰が、どこで、いつ、どのような攻撃を起こして、人間の犠牲者、および破壊された建造物は...」)しか対象にしていない。事件のタイプごとにこのような意味テンプレートをあらかじめ作成しておく必要がある。また、同じ事件に対する記事のみを対象にしている。

続報記事の合成: 船坂(1996)では、続報を伝える複数記事から冗長部分を削除して合成する。一般に続報記事は、背景としてこれまでの経過を伝える部分が存在する。記事を合成する際に冗長なこの部分を削除する。一連の同じ事件であることが要約の条件になっている。

複数文の合成: 同じ内容の文書(同じ事件を伝える複数新聞社の新聞記事)から同じ意味を共有する文を特定し合成する(柴田他 1997)。合成方法として、AND 型(要素の共通集合)、OR 型(合併集合)などがある。

ネットワークを用いた複数テキストの要約(豊浦他 1999): 数テキストからひとつの係り受けのネットワークを構築し、入る方向のリンクが多いノードを話題の中心となる重要ノードとみなし、そうしたノードを中心に要約を構成する。違った意味をもつ文もマージしてしまう可能性があり、適切な要約は得られない。

キーワード列挙以外の方法は、同じテーマをもつ文書(続報、同じ事件を伝える複数ニュースソースの記事)が対象となっている。分類で得られた文書群や、検索結果で得られた文書群に対して、概要把握を行うという用途には適していない。

3 共通意味概念の抽出

我々は、自由な検索・クラスタリングの結果に適用できる複数文書要約技術を目的としている。こうして得られた文書集合は、文書のテーマはそれぞれ違っていても、なんらかの共通する意味をもつはずである。我々はこの点に着目して、「共通意味概念を抽出すること」をねらい

[1] 分類は、予め分類枠が規定されているカテゴライゼーションと、分類枠を予め規定せず似たものを集めることによって分類するクラスタリングの 2 種類に大別される。カテゴライゼーションの場合、分類枠の意味は先に分かっているので、この意味では概要提示技術の必要はないといえる。しかしこの場合でも、集まってきたものの集合に、集めた基準以外の共通性質を見いだすことは意味があると考える。

とした。例えば、「野党は、景気対策の一環として所得税の減税を提案している」、「衆議院で所得税減税が可決された」、「所得税は来年度から減税される」は、「所得税の減税(所得税を減税する)」という意味概念を共通に含む。我々は、自由な検索・クラスタリングの結果に対する要約として、このような共通意味概念を抽出したものが適していると考え、これらを抽出する技術の研究を行うことにした。この意味概念はそれぞれの元文書から見れば断片であるということができ、今後、**共通意味断片**と呼ぶことにする。

キーワード列挙も共通意味断片であるといえるが、その単語の文脈が示されないので、読者が自らの知識を動員してその文書群で言われていることを推測するしかない。また、キーワード列挙では、「所得税を増税し、消費税を減税する」を含んだ文書が多い場合でも、文書群全体として「所得税、減税」が得られる可能性もある。

これらの問題に解を与える試みとして、Grouper (Zamir and Etzioni 1999)や(小川他 1999)のように単語を単語連続に拡張しているものがある。動詞的概念(イベント、状態)を名詞句で表現することの多い英語ではある程度効果をもつが、日本語では表層表現が多様になるため効果が薄れる。これに対して、我々は、深層での「共通性」を見いだすことによってキーワード列挙の問題を解決することを目指す。

このため、第一に、文型の差異を吸収する。「所得税を減税する」の例では、「所得税の減税」(名詞句)、「所得税減税」(名詞連続)、「所得税が減税される」(受動態)も共通と見なす。

また、シソーラスを用いて上位概念で共通なものを探索することにより、単語レベルの差異も吸収する。例えば、「白菜からダイオキシンが検出された」、「ホウレンソウからダイオキシンが検出された」から「野菜からダイオキシンが検出された」を得ることができる。

シソーラスを単純に用いると、抽象度が高くなりすぎて文書群の共通意味を把握できなくなる。例えばここに「お茶からダイオキシンが検出された」が加わると、「農作物からダイオキシンが検出された」になるし、「サバから…」が加わると、「生物から…」になる。これを避けるために、抽象度を上げるときに意味断片に与えるスコアを一定の割合で落とす。「ホウレンソウからダイオキシンが検出された」を「野菜からダイオキシンが検出された」に変換する際にスコアを落とせば、断片「ホウレンソウから…」の頻度が高い場合には、他の野菜の場合から集積して得られた「野菜から…」よりも高いスコアを持つことになり、「ホウレンソウから…」が文書群の代表となる。

以上が、本方式の基本的な考え方である。

4 基本アルゴリズム

ここではアルゴリズムの基本形を、ステップを追って説明する。この基本形では詳細を未指定のままにしているところもあり、また、このまま実装するには効率が悪いため、実装時にはモディフィケーションが必要となるが、まず基本的な考えを示すために概略だけを説明する。

4.1 文解析

対象とする文書群に存在する全ての文書の、全ての文を構文解析する。この結果は一般的にグラフ構造または木構造で表現されるが、ここでは、自立語の係り受け関係を表現した木構造を採用する。これは自立語をノードに、その間の関係をアークとして表現している。関係としては表層格を採用し、ラベルには格助詞そのものを記載している。例を図 1 に示す。

4.2 サブツリーの登録と変換

この解析木の可能なサブツリーを複数出力する。さらに、それぞれのサブツリーに対し、以下のような変換を行ったものも同時に登録する(図 1)。

類義語・上位語変換: シソーラスを用いて各ノードを類義語・上位語に変換したツリーを全て登録する。

互換関係への変換: 意味的に等価な関係への変換を行い、全て登録する。その中には、受動態・能動態の変換(「軽量の携帯電話がフーバー社によって発売される」→「フーバー社が軽量の携帯電話を発売する」、助詞の変換(「は」と「が」など)、自動詞・他動詞変換(「全角スペースがシンタックスエラーを起こす」→「全角スペースでシンタックスエラーが起きる」)などがある。

4.3 解析木へのスコア付け

得られた解析木にスコアを与える。スコア付け方法は基本アルゴリズムでは規定しないが、係り受けの数、関係の種類により重み付けを変えて合計する方法、単語の重要度スコア(tf*IDF 積など)を加算する方法などが可能である。現在の実装では、ノード数(自立語数)をスコアとして採用しており、より小さなサブツリーになるに従ってスコアが減少する。

変換されたサブツリーには、その変換された度合いに応じて元の木よりも低いスコアを与える。上位語変換の際に、シソーラスの階層差に応じて元の単語のスコアを逓減させる。互換関係への変換でも同様なスコア逓減を行う。図 1 では、「フーバー社は」が「電機メーカーは」になった時点で0.5、「は」を「が」に変換して0.25と順次半減させている。なお、現在の実装では、変換にともなうスコア逓減を取り入れておらず、このため上位語変換のレベル数を制限することで対処している。

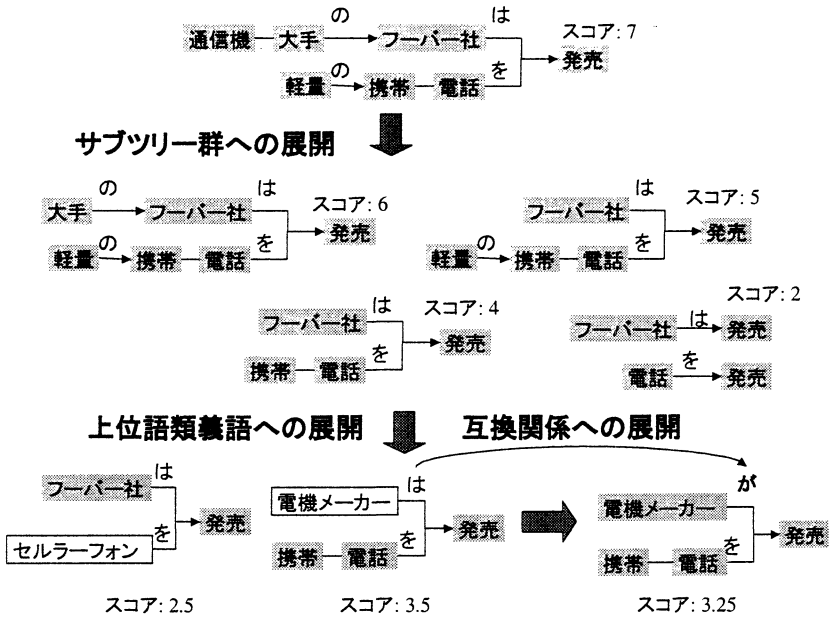

図 1 サブツリー群への展開

4.4 解析木スコアの累積

このようにして得られたサブツリーに対して、出現ごとのスコアを累計していく。「フーバー社は軽量の携帯電話を発売する」を含む文書と同じ文書集合に、フーバー社の他の製品、例えば、「簡易携帯電話」や、「W-CDMA方式の携帯電話」に言及があれば、「フーバー社は 携帯電話を 発売(する)」という共通ツリーのスコアが高くなる。一方、「電機メーカーが軽い携帯電話を発売する」に展開したものには、「富士通が軽い携帯電話を発売する」、「日本電気は軽量携帯電話を発売する」などを起源とするものが集まってくる。

4.5 文合成

蓄積されたサブツリーを累計されたスコアでソートする。上位の関係から必要数ピックアップし、サブツリーから文を組み立てる。生成される要約文がすでに生成した要約のいずれかのサブセットである場合は、その要約文をスキップする。

この方法により、文書集合の性質に応じて異なった共通意味断片が得られることになる。フーバー社の記事を集めた文書集合からは「フーバー社は携帯電話を 発売する」が生成され、フーバー社の主力商品が携帯電話であることが分かる。一方、最近の携帯電話に関する記事を集めた場合、「電機メーカーが軽い携帯電話を発売する」が生成され、「軽い携帯電話」がトレンドであることが分かることが期待される。

5 プロトタイプの実装

現在このアルゴリズムの実装を進めている。要約部分の開発言語は Java、関係解析部分は C++で、解析部分は(上田他 1999)で用いたものを共有している。

5.1 高速化

4章で示したアルゴリズムをこのまま実装すると、サブツリーの切り出しと上位語・類義語への展開で組み合わせ爆発をおこす。変換の際に適切な枝刈りが不可欠である。詳細は省略するが、以下の手法を取り入れている。

- ツリーの分割: 全文書から単語の2項関係(係り側の単語、間の関係、受け側の単語)を抽出し、低頻度の2項関係で解析木を分割し、分割された解析木ごとにサブツリーを切り出すことで、切り出されるサブツリーの数を絞り込む。
- 類義語展開の制限: 類義語展開した単語の2項関係が他に出現しない時、展開候補から除くことにより展開すべき候補数を減らすことができる。

これらを組み合わせることにより、生成されるサブツリーの数を 1/100 以下に絞り込むことができる。

5.2 基本アルゴリズムとの差異

基本アルゴリズムでは、実装方法に選択肢を残していた。また、現在の段階で実装に含まれていない部分も多い。ここでは基本アルゴリズムとの差異を簡単にまとめる。

- スコアリング: スコアには tf*IDF は用いておらず、自立語数×サブツリーの出現頻度を用いている。
- サブツリーの登録・変換: 登録は互換関係(助詞)のチェックを行っていない。また、態の変換など互換関係変換は行っていない。上位語変換にともなうスコア逓減を行っていないので、変換の段数を制限している。

6 実行例

複数文書要約における正しい要約の評価基準は存在しないため、ここでは実行例に対し定性的な評価を行う。

図2の例は毎日新聞の記事(99.10～12)を「日産」で検索した結果 25 件を要約したものである。ここでは、要約表現を元の文書中での表現(+)と対応づけて表示している。( )はスコアで、[ ]は頻度を表わす。これによると、「人員削減計画、リストラ計画、販売計画を発表」→「計画を発表」、「京都工場、久里浜工場、九州エンジン工場の閉鎖」→「工場の閉鎖」といった個々の表現が集まることにより抽象化された要約句が得られている。

図3の例は毎日新聞の記事(95.1～12)を「株式市場」で検索した結果を期間ごとに要約したものの一部である。これによると、「95 年の前半に低迷していた株価が後半から回復し市場が活発化していく」という株式市場の流れを大まかに把握することができる。

7 終わりに

共通意味断片を抽出することによる複数文書要約を考案し、そのプロトタイプを実装した。新聞記事の検索結果に対して適用したところ、文書群の性質を反映させた要約が得られた。

現在のプロトタイプは、もとのアルゴリズムの完全な実装ではない。今後フル実装をおこなうために必要な高速化アルゴリズムをさらに加えていく必要がある。また、現状ではシソーラスが 2.5 万語と不十分であるので、10～20 万語程度まで充実させる必要があろう。特に専門用語の拡充は、発見を行う意味で重要と考える。

謝辞

新聞データ(CD－毎日新聞 95 年版、99 年データ)の利用を許諾して頂いた毎日新聞社に感謝いたします。

```
1: 計画を発表 (16.0)        ( ):スコア, [ ]:頻度
    +計画を発表 [3]
    +人員削減計画を発表 [1]
    +リストラ計画を発表 [2]
    +販売計画を発表 [2]
2: 工場の閉鎖 (14.0)
    +工場の閉鎖 [2]
    +京都工場も閉鎖 [1]
    +日産久里浜工場と閉鎖 [1]
    +日産九州エンジン工場も閉鎖 [1]
    +エンジン工場も、閉鎖 [2]
3: 日産自動車がリバイバルプラン」を発表 (12.0)
    +日産自動車がリバイバルプラン」を発表 [2]
    +日産自動車は日産リバイバルプラン」を発表 [2]
4: 日産自動車は発表 (12.0)
  .....
```

図2「日産」で検索した結果の要約例

```
■95.1～95.3 (30 件)     ( ):対象記事件数, [ ]:頻度
  今年の最安値となる [2]
  理由で、中心に売る [2]
■95.4～95.6 (26 件)
  株価は下落 [6]
  安値を更新 [4]
  売りが出る [4]
■95.7～95.9 (30 件)
  一万八千円台を回復 [5]
  東証の平均株価は、回復 [3]
  東京株式市場は買いが活発化 [3]
■95.10～95.12 (22 件)
  ニューヨーク市場は、史上最高値を更新 [3]
  株式市場は活況を続ける [3]
```

図3「株式市場」で検索した結果の期間ごとの要約例

参考文献

Cutting, et al. (1992): "Scatter / Gather: A Cluster-based Approach to Browsing Large Document Collections" SIGIR-92.

Zamir and Etzioni (1999): "Grouper: A Dynamic Clustering Interface to Web Search Results" WWW8.

McKeown and Radev (1995): "Generating Summaries of Multiple News Articles" SIGIR-95.

上田、岡、小山、宮内 (1999):「句表現要約手法に基づく要約システム」言語処理学会第 5 回年次大会.

小川、落谷、西野 (1999):「文書クラスタの判別のための特徴表現付与」 言語処理学会第 5 回年次大会.

柴田、上田、池田 (1997):「複数文書の融合」自然言語処理研究会 120-12.

豊浦、津高、瀬尾 (1999):「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.

船坂、山本、増山 (1996):「冗長度削減による関連新聞記事の要約」自然言語処理研究会 114-7.