複数の関連テキストに対する情報集約ツール

落谷 亮
富士通研究所
ochi@flab.fujitsu.co.jp

1 はじめに

商用データベースのインターネット経由のアクセスサービスなどテキストDBサービスの多様化に伴い、複数の大規模テキストDBに対し簡単な操作で検索を行なうということも身近に行なえるようになってきた。例えば、新聞記事データベースの横断検索サービスでは、一度の検索条件入力で複数の新聞社の記事データベースを一度に検索することが出来、あまり一般的でない希少な情報の検索や同じ事件の各種の関連情報を広く集めるのに非常に有効である。

しかし、この反面、横断検索により得られるデータは、関連分野のDB、例えば、複数の新聞社の新聞記事や複数の雑誌社の記事から検索された結果であり、それぞれ同じ内容の事実を表現したテキストを多く含んでいる。結果として、従来の見出し一覧による検索結果の表示では、同一内容の記事や類似の記事が表示され、画面上に同じような見出しが並ぶことになり、検索システムの利用者は、欲しいデータに辿り着くまでに同じ内容の見出しや記事を繰り返し読まされるというようなことが頻繁に起こる。このような状況は、検索結果のチェックに掛かる手間を増やすだけでなく、記事検索サービスの様な有料サービスでは見出しを表示させることさえも課金の対象になるため、利用者の検索意欲を妨げる障害になるとも考えられる。

この様な類似の見出しやテキストデータを手短に要約したり、整理して提示する問題は、新聞記事の検索やインターネットのデータ検索などの問題として最近注目され、類似記事の要約 [山本]、文書の融合 [柴田] などの研究結果が報告されるようになってきた。

山本ら [山本] による関連記事からの冗長部分削減による記事要約の研究では、共通部分の発見処理と修飾語句などの冗長部分の除去による要約を扱っており要約作成を主眼とした研究である。

柴田 [柴田] らの研究は文を単位とした文書の融合を目標に、文間で同一形態素の出現頻度により文書間の文の対応付けを行ない、情報の AND や OR 等の操作により複数文書をまとめる融合処理について報告している。

この様に、類似記事について、共通情報を用いて短く要約する、共通情報を融合して見せるなど幾つかの目標に向けての研究が行なわれているが、これらの目標を解決するには、以下の3つの点を解決する必要がある。

1. 共通情報の発見
2. 情報の順位付け
3. 情報の結合

このうち 2,3 は、検索者の意図などテキスト情報以外の状況に大きく左右される部分であり、情報の重要度計算や選択・表示手段を考える際には、これらの文脈情報も含めた処理が不可欠である。それに対し、1 のテキストの共通情報については、テキスト間の表現の共通性に着目することで処理が可能と考えられ、どの程度の共通性が実際見つかるかがテキストだけから判定出来ると考えられる。

この様な観点から、新聞記事テキストには、どの程度の共通表現が見られるのかを、2つの異なる新聞社の記事を対象に各種の照合方法に基づいて照合し原文に対する照合比率を調査したので、その結果について報告する。そして、最後に応用としての集約処理について簡単に紹介する。

2 類似テキストの照合

今回試みた共通情報の照合処理は、シソーラスなどの大規模な知識ベースを出来るだけ利用しない処理を想定し、処理は文字列、及び、部分文字列の照合のみを用いて実験を行なった。

類似照合として、元のテキストの段落、文、句を単位として、文字列、及び、形態素解析した結果に対して、完全一致、共通文字列照合を、それぞれ組み合わせた処理を想定した。以下に、今回の実験に用いた照合処理について順に述べる。

1. 単純な共有文字照合

   記事見出しの文字照合により関連記事判定がある程度可能であるという関連記事の検索に関する報告 [奥] もあり、共有される文字が原文でどの程度の部分を占めるかを測定した。共有率は、文字列 T の文字数を $len(T)$ とすると、文 S_i 、 S_i に対し、

   $$\text{共有率} = \frac{len(S_i\text{の共有文字全体}) + len(S_j\text{の共有文字全体})}{len(S_i)+len(S_j)}$$

   この照合条件は、もっとも緩い条件での最大照合率を示すと考えられる。字面の照合による共通化処理では、この値を越えることがないという目安になる。直観的には、この処理の照合率は最大になるが、文字の間のあまり意味のない照合も含んでおり、再現率は高いが適合率が低い処理の目安を与えていると考えられる。

2. 最長共通文字列照合

   最長共通文字列を照合し、その文字列長が原文に占める比率を測定した。

   $$\text{共有率} = \frac{len(S_i, S_j\text{の最長共通文字列}) \times 2}{len(S_i)+len(S_j)}$$

   実際のオンライン記事ＤＢでは、新聞記事の地方版等で、紙面の構成の都合上、若干の編集が入ったデータが存在し、殆んどの部分は同じで、一部分が編集されているような記事が混在する。そのようなケースでは最長共通文字列照合によりテキスト間の一致が取れる。最長共通文字列は再現率は低いが適合率の高い処理の目安と考えられる。

3. 共有形態素照合

   1 の文字照合と同じやり方で、形態素単位での共有が原文に占める率を測定した。

   $$\text{共有率} = \frac{len(S_i\text{の共有形態素全体}) + len(S_j\text{の共有形態素全体})}{len(S_i)+len(S_j)}$$

   文字単位の照合よりは単位が長いので、再現率が 1 の共有文字照合よりは若干下がり適合率は上がると考えられる。

4. 最長共通形態素照合

   $$\text{共有率} = \frac{len(S_i、\ S_j\text{の最長共通形態素}) \times 2}{len(S_i)+len(S_j)}$$

   2 の文字の最長共通文字列照合と同じやり方で、形態素を単位として単位で最長共通列を探し原文に占める率を測定した。形態素解析の誤りなどにより、文字単位の最長共通部分よりは若干短くなる部分が増えると考えられる。

5. 句と句の共有形態素照合

   日本語の句の順序は自由であるため、原文の形態素順序を保存した最長共通形態素照合では句の表現順所が変わると照合できない。厳密な処理であれば、構文解析結果に対し、部分照合するのが望ましいが、ここでは、句読点、格助詞により句の切れ目を決定し、それらのを切れ目で区切った単位に局所的に句の間の最長共通文字列照合を行ない、大局的には句の間の順序を任意に変えた組合せで形態素の文字共有度の和が最大になるものを選び、最大候補に対して、原文に対する文字列の比率を計算した。

   これを式で表すと以下の様に表される。

   S_i の句 K_{il}、 S_j の句 K_{jm} に対し

   $$\text{共有率} = \frac{max(\sum_{K_{il}=K_{jm}} len(K_{il})) \times 2}{len(S_i)+len(S_j)}$$

   但し、K_{il} は K_{jm} は、複数要素に照合しないとする。

6. 句と句の最長共通文字列

   先の場合と同様に、句を区切った後、句の文字列に対し最長共通文字列を計算した。句の組合せ全てに対しそれらの最長共通文字列を求め、その共有文字数の総和が最大になるものを選び、原文に対する文字列の比率を計算した。

   式で表すと以下の様に表される。

   $$\text{共有率} = \frac{max(\sum len(K_{il}, K_{jm}\text{の共通文字列})) \times 2}{len(S_i)+len(S_j)}$$

   但し、K_{il} は K_{jm} は、複数要素に照合しないとする。

	処理方法	照合比率 (%)		
		第 1 段落	見出し	第 1 文
1	共有文字	71.2	51.2	68.6
2	最長共通文字列	40.0	35.5	46.5
3	共有形態素	59.0	45.0	59.2
4	最長共通形態素	33.9	31.9	42.7
5	句の共有形態素	33.2	32.7	41.0
6	句の最長共通列	32.4	31.4	40.8

表 1: 共通部分の照合比率

3 実験

3.1 実験対象データ

日経新聞 [日経] 及び毎日新聞 [毎日] の記事データについて同一時期（1995 年 6 月）1カ月分の記事（日経 16323 記事、毎日 9268 記事）を用いた。実験に当たっては、各記事の間で類似記事を絞り込んだ後に、類似記事の間で、先に述べた照合を処理を行ない照合率を計算した。

3.2 類似テキストの絞り込み

各記事の見出し及び第 1 段落のテキストを類似度判定することにより記事を絞る。ここでは、見出し及び第 1 段落の文全部の形態素解析を行ない、その結果から、形態素を単位とした頻度ベクトルを作成しベクトル間のコサイン値を計算し、コサイン値が一定の閾値を越えるものを類似記事とした。元データ約 16,000 記事と 9,000 記事の組合せに対し、この処理で選ばれた類似記事の組合せ総数は 1082 記事となった。

記事全体のテキストを利用して、第 2 段落以降も段落や文毎の単位に分け、それらを過去の記事と照合することにより、事件の経過など一連の関連記事の参照表現なども処理できると考えられるが、今回の実験では第 1 段落だけに処理対象を限定して行なった。

3.3 照合率

各処理方法により実験を行ない、照合部分の文字列長を元のテキストの文字列長に対する比率として、先の式に従い計算した値を表 1 に示す。

1,2 は要素間に複数回の照合を許しているため照合率は高く、5 ～ 6 割程度の部分が照合しており、2,4,5,6 は、一度の照合しか許さないため照合率は 3 ～ 4 割程度と低くなる傾向がある。

直観的には、見出しは類似度が高そうであるが、第一文より 1 割程度は照合率が低いという結果が得られている。

4 類似照合パターンの分析

先の実験では機械的な近似処理で類似照合を行なったが、実際、どのような類似・相違のパターンがあるのかは、これらの照合率を向上させたり、他の情報抽出等の処理の際のパターン規則の参考にもなるので、最長共通文字列を照合した際に、照合しない部分文字列について調べた。以下に、顕著な相違を列挙する。

1. 数字表現

   漢数字と算用数字の用法は、例えば、日付、金額の表記など、新聞別だけでなく、記事の間でも微妙に異なる。

   また、「１９９５」と「９５」などの年の表記の違いも新聞毎に傾向が異なる。
2. 括弧「」の利用

   特別な「ＸＸ会議」、「ＹＹ基金」などの固有名詞を括弧「」で囲むかどうかの基準が新聞毎に異なる。
3. 格の違い、主題化の違いによる助詞の相違

   「は」と「が」、「で」と「に」などの違いが高い頻度で起こる。
4. 情報の詳細度

   「Ｘ日午前」の午前や午後など詳細時間の表現や、「村山富市首相」と「村山首相」、「亀井静香運輸相」と「亀井運輸相」など、人名の表記法の詳細度の相違が見られる。
5. 言い回し

   「明らかになる」と「分かる」、「発表する」と「明らかにする」など述語や「全面」と「すべて」の言い回しの違いにも傾向がある。
6. 記号

   ’－’と’—’、及び、’＝’の使用等、記号の用法は規則が異なる。

今回の実験では、これらの違いは照合率に反映していないが、共通事実情報の抽出率を考える場合には、これらの表現上の違いを吸収して計算する必要があると考えられる。

	処理方法	照合比率 (%)		
		第 1 段落	見出し	第 1 文
3	共有形態素	39.6	40.8	45.0
4	最長共通形態素	28.1	30.1	38.2
5	句の共有形態素	31.4	31.8	39.4
6	句の最長共通列	30.4	30.9	39.0

表 2: 機能語補正後の照合比率

5 応用

共通情報の抽出結果の応用として、以下に挙げるような幾つかの処理を想定している。

1. 共通事実文の生成と表示。

   共通部分の句を組合せることにより文を生成する。手法としては、山本らの要約文生成手法と類似の方法で、機能語だけが生成されるのを防ぐため、前に現れる自立語が共通でなく生成されない場合には、機能語だけが生成文に残らないように、その機能語を削除する手法を取る。

   このように孤立した機能語除去を行ない生成された文は、先の実験で計算した照合部分より短くなる。どの程度短くなるかを実際に計算したものを表 2 に示す。この結果では句の照合の方が最長形態素照合よりは若干ではあるが共有率が高くなっている。

   見出し間で照合を行ない文生成をさせた結果と、第 1 段落に対し処理した結果の生成文を原文と共に図 1 に示す。

2. 情報抽出の前処理

   会社名、日時、金額など記事からの情報抽出の前に、類似記事との照合を行なうことにより、表現のバリエーションや冗長な説明語句を除去することが出来、抽出処理の処理対象を狭めることが出来るとも考えられる。

   また、先に述べた文字列照合の相違情報を広範に集めることにより、表現のバリエーションや相違、冗長な表現の一般的傾向なども収集できるので、情報抽出システムの規則開発の支援ツールも応用として考えられる。

6 まとめ

新聞記事を対象として、類似の事実を表現した文で、どの程度の率でテキスト間に照合ができるかについての実験結果について述べた。

見出し 1	生産者麦価、据え置きへーーきょう、米審に諮問
見出し 2	政府・与党、生産者麦価据え置き——農相、きょう諮問。
生成文	生産者麦価、諮問
原文 1	大河原太一郎農相は 1 日、１９９５年産麦の政府買い入れ価格（生産者麦価）を据え置くよう米価審議会（渡辺五郎会長）に諮問する。
原文 2	大河原農相は 1 日午前から開かれる米価審議会（渡辺五郎会長）に９５年産の麦の政府買い入れ価格（生産者麦価）を諮問する。諮問案をめぐる３１日から…（略）…
生成文	大河原農相は 1 日年９５産麦の政府買い入れ価格（生産者麦価）を米価審議会（渡辺五郎会長）に諮問する。

図 1: 生成文の例

今回の実験では表記上の照合だけを集計したが、実際に表現内容の事実情報の対応が取れているかどうかについて調べることも実験としては興味深い。特に、単純に形態素の共有最長列を選ぶのと、句の順序を入れ替えて照合させた結果の違いなどの比較も課題であると考えている。

新聞記事などで、一般的な事実を記述した部分と、記者の判断や予測等の独自の創作部分を分けることは、著作としての記事独自の情報を守りながら、検索者が探している情報への手がかりを提示したり、手がかり情報のみを著作と切り離して流通させていくなど、情報流通を実現する技術の一部として重要と考えており、今後の大きな目標と考えている。

参考文献

[奥] 奥 雅博, 鷲崎 誠司, 田中智博, 関連記事の判定に関する検討, 言語処理学会第 2 回年次大会, pp. 89–92. 1996.

[柴田] 柴田 昇吾, 上田 隆也, 池田 裕治, 複合文書の融合, 電子情報通信学会技術報告, NLC97-19, pp. 7–12. 1997.

[山本] 山本 和英, 増山 繁, 内藤 昭三, 関連テキストを利用した重複表現削減による要約, 電子情報通信学会論文誌 (D-II), pp. 1968-1972. 1997.

[日経] 日経全文記事データベース 日本経済新聞 CD-ROM 95 年版, 日本経済新聞社

[毎日] CD 毎日新聞 95 年版, 毎日新聞社