複数記事要約のためのサマリパッセージの抽出

橋本 力† 奥村 学‡† 島津 明†

†北陸先端科学技術大学院大学 情報科学研究科 ‡東京工業大学 精密工学研究所

1 はじめに

複数の新聞記事を要約する際には，複数記事中の重要個所を同定する必要がある．従来の手法には，語の出現頻度などの統計的情報や，文，段落などの出現位置の情報を用いて重要個所を同定するものや，予め抽出すべき情報を設定してあるテンプレートを用いて重要個所を同定するものがある [1]．それに対し本研究で提案する手法では複数記事中の意見記事，解説記事，まとめ記事中の情報を利用する．これらの記事の中には過去の主要な出来事が新聞記者の視点で要約されている個所が存在する．我々はこの個所を**サマリパッセージ**と呼び，重要個所同定に有効であると考える．従来の統計的手法などは記事内容を理解せずに重要個所を同定する方法と言えるが，サマリパッセージを用いた重要個所同定法は，内容を熟知している記者の知識を利用した方法であり，より自然な重要個所同定が可能と考えられる．本稿では検索結果の記事集合からサマリパッセージを抽出する手法と，その手法を実装したシステムの評価結果について述べる．その後，サマリパッセージを用いた複数記事要約の手法を提案する．本研究では記事コーパスとして毎日新聞社のものを用いる．

2 サマリパッセージ

サマリパッセージは，意見，解説，まとめ記事中の，過去の主要な出来事が新聞記者の視点で要約されている個所である．例えば，意見，解説記事は，過去に起こった出来事について述べた後，その出来事についての意見や解説を述べる，という構造を持つと考えられる．以下に，例として意見記事の一部を挙げる．

…**被疑者不詳としながらも，河野さんの自宅を殺人の容疑で捜索した．**県警は「法律上，問題はない」としているが，情報がこれで独り歩きすることを考えなかったのか．…**やがて，捜査の目は河野さんから教団に向けられていく．**しかし…

太字部分が過去の出来事についての記述，サマリパッセージである．解説記事も同様の構造を持つと考えられる．まとめ記事は，それまでの経緯が分かるように，過去の重要な出来事を網羅的にまとめている．以下に，例として箇条書のまとめ記事の一部を挙げる．

3・20・東京で地下鉄サリン事件発生
21・麻原彰晃容疑者がロシア・ウラジオストクからのラジオ番組で「悔いのない…
⋮
5・ 2・日劇の元ダンサーとして知られる鹿島とも子容疑者を逮捕監禁容疑で逮捕
3・教団「法務省」大臣で顧問弁護士の…

箇条書の一行一行がサマリパッセージである．

記者が意見，解説，まとめ記事を書く際，過去のどの内容に触れるかは，その記者が過去のどの内容が重要かを考慮して決定していると考えられる．従ってサマリパッセージを参照すれば，対象となっている話題のそれまでの経緯の中で，どの内容が重要なのか判断でき，複数記事要約の際の重要個所同定に有効である．

3 サマリパッセージの抽出

サマリパッセージは意見，解説，まとめ記事にあるので，検索結果の記事集合からサマリパッセージを抽出してくるには 1) 記事集合から意見，解説，まとめ記事を検出し 2) 検出された各記事毎にサマリパッセージを抽出する．1) のステップを介さず 2) を行うと，意見，解説，まとめ記事以外 (報道記事など) から過去の記述が抽出されることがある．報道記事などにある過去の記述は，最新の出来事とそれを補う情報についての記述であり，複数記事要約に有効でないものが多いと考えられ，抽出対象から除外する．

3.1 本研究における記事のモデル

記事によっては異なるカテゴリ (意見，解説，まとめ) に属する内容が 1 記事中に存在する．例えば記事の前半に解説的な個所があり後半にまとめ的な個所が存在する場合がある．このような記事に対応するためカテゴリ毎のまとまりを単位として記事を処理する．本研

究ではカテゴリ毎のまとまりをセクションと呼ぶ．セクションは記事中で明示的に分割されている場合が多い．各セクションを以下のように定義する．

意見セクション　過去の出来事の記述 (サマリパッセージ) と，その出来事についての意見，主張が述べられているセクション

解説セクション　過去の出来事の記述 (サマリパッセージ) と，その出来事の背景にある事情や一般的見解などが述べられているセクション

まとめセクション　過去の出来事の記述 (サマリパッセージ) が，意見，解説的な内容を含まないでまとめられているセクション

意見，解説，まとめセクションは全てサマリパッセージを含むので，これらを一括して**サマリを含むセクション**を呼ぶ．これら以外のセクションとして，報道的なセクション，エピソードを述べているセクション，裁判記録などの公文書をそのまま載せてあるセクション，用語解説セクション，一覧表形式のセクションがある．これらを**その他のセクション**と呼ぶ．

以上のように記事をモデル化し，検索結果の記事集合からサマリパッセージを抽出する．全体の処理の流れを図 1に示す．

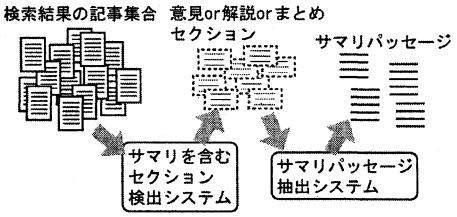

図 1: 全体の処理の流れ

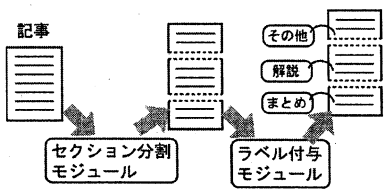

図 2: サマリを含むセクション検出処理

3.2　サマリを含むセクションの検出

検索結果の記事集合からサマリを含むセクション (意見，解説，まとめセクション) を検出する．サマリを含むセクション検出システムは図 2にあるように，**1)** 記事をセクションに分割するモジュールと **2)** 各セクションに意見，解説，まとめ，その他のラベルを付与するモジュールの 2 つから構成されている．

3.2.1　セクション分割モジュール

現在のセクション分割モジュールは以下の単純なルールから構成されている．以下のいずれかの条件を満たす行をセクション区切り行と見なす．

1. 行頭が◆，◇などの記号で，途中と末尾に「。」が無い行
2. ＜ … ＞，【 … 】などの括弧だけの行
3. ……の線だけの行

3.2.2　ラベル付与モジュール

ラベル付与モジュールは 9 つのフィルタから構成されており，各フィルタは 4 つのカテゴリのいずれかに対応している．入力されたセクションは以下で示す順にフィルタにかけられ，かかったフィルタに対応するカテゴリのラベルを付与され，その後のフィルタにはかけられず直ちに出力される．例えば 2 番目のフィルタにかかったら 3 番目以降のフィルタにはかけられない．フィルタの順番は，まず，見出しからカテゴリを判断できるセクション用のフィルタ (1〜4) を，次に，内容から判断するセクション用のフィルタ (5〜7) を，最後に，現在はまだ特徴をつかめていないセクション用のフィルタ (8〜9) を起動するようにしている．

1. まとめ (「週間日誌」)
2. 意見 (「社説」)
3. 解説 (「解説」)
4. その他 (公文書)
5. まとめ (箇条書)
6. 意見 (文末)
7. その他 (報道)
8. 解説 (残り)
9. その他

以下に各カテゴリ毎のラベル付与ルールを述べる．ルール作成には以下のデータを用いた．

- 95 年の記事を「地下鉄 サリン オウム」で検索した結果の上位 300 記事
- 96 年の記事を「核実験 CTBT」で検索した結果の上位 100 記事

意見セクション

記事全体の見出しに文字列「社説」が含まれている場合，フィルタ 2 がその記事全体に意見ラベルを付与する．また，セクション全体の中で意見を述べている文の占める割合が高い場合，フィルタ 6 がそのセクションに対して意見ラベルを付与する．ある文が意見文かどうかは文末表現を調べて判断する [2].

解説セクション

セクション全体の見出しに文字列「解説」がある場合，フィルタ 3 がその記事全体に対して解説ラベルを付与する．また，フィルタ 8 により，以下の 2 つの条件を満たすセクションに対しても解説ラベルが付与される．

1. 1〜7 のフィルタにかからずに残ったセクションで，
2. 閾値 (現在は 5 文) 以上の文があるセクション

2. の条件は，文がわずかしかないセクションや一覧表のセクションを候補から除外するために設けた．

まとめセクション

まとめには以下の 3 つのサブカテゴリがある．

- 「週間日誌」
- 箇条書によるまとめ

- 文章によるまとめ

多くの新聞には, ある一定期間内に起きた主要な出来事をまとめている, 定期的に連載される記事があると考えられる. 毎日新聞では「週間日誌」がそれに相当する. 「週間日誌」の見出しには文字列「週間日誌」が必ずあるので, 見出しにその文字列がある場合, フィルタ 1 によってまとめラベルが付与される.

新聞記事には, 行頭に日付, 次にその日の出来事の説明が述べられている箇条書の形式で主要な出来事をまとめている記事がある. フィルタ 5 は, 先頭に数字が来る行が閾値 (現在は 3 行) 以上連続して現われるセクションにまとめラベルを付与する.

また, それまでの経緯を文章でまとめているセクションもある. しかし現状のシステムでは, 文章によるまとめのセクションをそれと認識することはできない.

その他のセクション

公文書をそのまま載せてあるセクションの見出しには, 文字列「要旨」か「全文」があることがほとんどである. 見出しにそのような文字列があれば, フィルタ 4 がその他ラベルを付与する.

報道セクションには以下の 3 つの特徴のいずれかが含まれている場合が多い. セクションにこの 3 つの特徴のいずれかがある場合, フィルタ 7 がその他ラベルを付与する. 1 つ目の特徴は [3] による.

1. セクションの先頭のある閾値 (現在は 1) 文以内に「～(事件 | 事故 | 問題) で」か「～について」を含む
2. セクションの先頭のある閾値 (現在は 1) 文以内に「～?日 ～ (た | る).」を含む (但し,?日は, 記事出版日から閾値 (現在は 3) 日以内の日付とする)
3. 「(～面に関連記事)」を含む

どのラベルも付与されず, 解説ラベルも付与されなかったセクションに対しては, 最終的に, フィルタ 9 がその他ラベルを付与する.

現在はエピソードと用語解説, 一覧表の特徴は掴めておらず, それと認識することはできない.

3.3 サマリパッセージの抽出

サマリを含むセクション検出後, サマリパッセージ抽出処理に移る. 抽出方法はセクションの形式によって異なる.

文章によるまとめ, 意見, 解説

これらのセクションは通常の文章で構成されており, サマリパッセージは以下の 2 つのタイプの文を抽出することで得られる. 文タイプは文末表現から判定する.

- 叙述文 過去
- 叙述文 状態

箇条書によるまとめ

行頭の数字やインデントの有無などを調べ, 箇条書全体を抽出するようにした.

「週間日誌」

週間日誌は以下のような構造になっている.

【△日】
その日の出来事 1
その日の出来事 2
...
【△+1 日】
その日の出来事 1
その日の出来事 2
...
...
【△+n 日】
その日の出来事 1
その日の出来事 2

週間日誌では, 検索意図から外れている話題についても述べているので, 記事を検索した時に使用したクエリを含む行だけを抽出するようにしている.

4 評価

4.1 ラベル付与の評価

前述の通りサマリを含むセクション検出システムは 2 つのモジュールからなるが, その内のラベル付与モジュールの評価を行った. 今回, ラベル付与の前段階であるセクション分割は人手で行った.

4.1.1 評価方法

意見, 解説, まとめのラベル付与結果と, サマリを含むセクション全体 (意見, 解説, まとめ) のラベル付与結果の評価を行った. サマリを含むセクション全体の評価では意見, 解説, まとめの区別はしないので, 例えば意見セクションに解説のラベルが付与されていても正解と見なす. システムのラベル付与の結果を以下の式で評価した.

$$\text{再現率} = \frac{\text{集めることができた正解セクション数}}{\text{全正解セクション数}}$$

$$\text{精度} = \frac{\text{集めることができた正解セクション数}}{\text{集めた全セクション数}}$$

評価には,97 年の記事を「山一 野村 証券」で検索した結果の上位 100 記事 (126 セクション) を用いた.

	意見	解説	まとめ	サマリ
セクション数	13	50	17	80

4.1.2 実験結果

※ 表中の ( ) 内の数値は再現率, 精度の分母, 分子に相当する.

	意見	解説	まとめ	サマリ
再現率 %	85($\frac{11}{13}$)	82($\frac{41}{50}$)	59($\frac{10}{17}$)	83($\frac{66}{80}$)
精度 %	85($\frac{11}{13}$)	82($\frac{41}{50}$)	77($\frac{10}{13}$)	87($\frac{66}{76}$)

意見セクション

間違って検出したセクションの多くは, 意見か解説か判断が人によって揺れると思われるセクションだった. 見落としたセクションの多くは, 文末表現リスト(意見文かどうかを判断するのに用いる) に無い意見表現 (「痛感する」「気がする」 など) が多く用いられているセクションだった. 見落としには文末表現リストを充実させることで対応可能と考えられる.

解説セクション

現在はフィルタにかからずに残ったセクションから解説を特定しているが, このせいで,8 つめのフィルタ以前のフィルタの間違い, 見落としが解説セクション検出の結果に悪影響を与えた. よって解説独自の特徴を把握し, 解説独自の検出方法の開発が望まれる. 解説の特徴として, 背景などを表す「~からである」や「~わけである」, 推量を表す「~だろう」や「~ようだ」などの表現が有効な手がかりと考えられる.

まとめセクション

まとめサブカテゴリの一つ「週間日誌」は全て検出できた. 箇条書のまとめでは, いくかの不規則的な箇条書の形式には対応できず, 再現率を下げた. 不規則的な形式とは, 以下のような, 出来事の日付を表す数字が箇条書の行頭以外に現われるようなものなどである.

> 【山一証券】１９９５年１月, 一任勘定取引の損失の穴埋めなどとして約７９００万円を付け替え
> 【日興証券】９３年１０月に購入した東京電力株の値下がりで約８０００万円の損失を出し, その後１０００万円を付け替え
> ...

これらに対応できるように箇条書の認識方法をより柔軟にする必要がある. また, 現在文章によるまとめには対応できていない. 文章によるまとめには, どのラベルも付与されなかったセクションで, 過去の出来事の記述 (サマリパッセージ) がほとんどを占めるセクションを検出することで対応できる可能性がある.

4.2 サマリパッセージ抽出の結果について

文章によるまとめ, 意見, 解説からのサマリパッセージ抽出において抽出できなかったものとしては, 体言止めで過去を記述している文がほとんどだった. また, 過去の出来事の記述と, それに対する意見, 解説をまとめて一文で述べているような場合, その中の過去の記述は抽出できない. 例えば,

> 五日夜, 東京・新宿の地下鉄丸ノ内線新宿駅のトイレに青酸ガス発生装置が仕掛けられた事件から教訓を学びたい.

など. 箇条書のまとめからの抽出では, 前述した不規則的な形式の箇条書に対しては失敗する. 週間日誌は構造が決まっているため全てうまくいっている.

5 サマリパッセージを用いた要約

抽出したサマリパッセージから複数記事要約を生成する方法として, 図 3のように考えている.

1. サマリパッセージをクエリとして過去の記事集合を検索し, 最も関連の強い個所と対応づける.
2. サマリパッセージと関連の強い個所を重要個所と見なし, それらを元に要約を生成する.

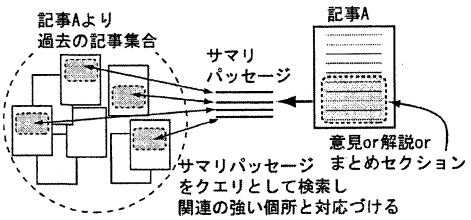

図 3: サマリパッセージを用いた複数記事要約

6 おわりに

本稿では, サマリパッセージが複数記事要約に有効であることを論じ, 記事集合からのサマリパッセージの抽出方法と評価結果, サマリパッセージから要約を生成するまでの処理について示した. 今後は **1)** サマリパッセージ抽出の精度向上を目指すとともに **2)** 今回は行わなかったセクション分割モジュールの評価と **3)** サマリパッセージから要約を生成するまでの処理の詳細の検討が必要である. また今回は一人によって作成されたデータを用いたが,**4)** 複数の被験者を募り意見, 解説, まとめの判定実験を行い, その結果を元により客観性の高いデータ, ルール作成を行う必要がある.

謝辞

新聞記事データの利用を許諾してくださった毎日新聞社に感謝申し上げます.

参考文献

[1] 奥村学, 難波英嗣. テキスト自動要約に関する研究動向. 自然言語処理,Vol.6,No.6,pp.1-26. 1999.

[2] 福本淳一, 安原宏. 日本語文章の構造化解析. 情報処理学会自然言語処理研究会 85-11. 1991.

[3] 阪元慶隆, 渡辺靖彦, 岡田至弘. 表層表現を手がかりにした続報記事の抽出. 言語処理学会第 4 回年次大会発表論文集 pp368-371. 1998.