

Title	複数記事要約のためのサマリパッセージの抽出
Author(s)	橋本, 力
Citation	
Issue Date	2001-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/1426
Rights	
Description	Supervisor: 島津 明, 情報科学研究科, 修士

修 士 論 文

複数記事要約のためのサマリパッセージの抽出

指導教官 島津明 教授

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

橋本 力

2001 年 2 月 15 日

要 旨

本研究では、新聞記事コーパスを検索した結果の記事集合を対象に、それらの記事集合の中からサマリパッセージを抽出する方法と、抽出されたサマリパッセージを用いて記事集合を要約する方法について述べる。新聞記事の中には、意見記事、解説記事、まとめ記事があり、それらの記事の中には、対象となっている話題の過去の経緯における重要な出来事が、新聞記者の視点でまとめられている個所が存在する。本研究では、この個所をサマリパッセージと呼ぶ。ある話題に関する複数記事を要約する際は、その複数記事中の重要個所を同定する必要があるが、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。

検索結果の記事集合からサマリパッセージを抽出するには、まず、記事集合中から意見記事、解説記事、まとめ記事を検出するのだが、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进行处理する。このカテゴリ毎のまとまりをセクションと呼ぶ。従って、記事集合からサマリパッセージを抽出するには(1)記事集合から意見セクション、解説セクション、まとめセクション（この3つを一括してサマリを含むセクションと呼ぶ）を検出し(2)検出された各セクションから、そのセクションの形式に応じた方法でサマリパッセージを抽出する。本研究では、このような枠組で検索結果の記事集合からサマリパッセージを抽出するシステムを実装した。全体のシステムは(1)サマリを含むセクション検出システムと(2)サマリパッセージ抽出システムから構成される。サマリを含むセクションは、記事をセクションに分割するセクション分割モジュールと、分割された各セクションに、そのセクションが属するカテゴリを表すラベルを付与するラベル付与モジュールから構成される。

本研究では、サマリを含むセクション検出システムとサマリパッセージ抽出システムをそれぞれ評価した。サマリを含むセクション検出システムの評価では、セクション分割処理は人手で行ない、ラベル付与モジュールのみを定量的に評価した。また、各セクションからのサマリパッセージ抽出もおおむねうまくいくことを示した。

最後に、抽出されたサマリパッセージを用いた複数記事要約手法について論じる。また、本研究の提案手法と従来の手法の違いを述べる。

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。

目次

1	はじめに	1
1.1	研究の背景	1
1.2	複数記事要約におけるサマリパッセージの有効性	2
1.3	研究の目的	4
1.4	本論文の構成	7
2	本研究における新聞記事のモデル	8
2.1	セクション単位の扱い	8
2.2	意見セクション	9
2.3	解説セクション	10
2.4	まとめセクション	11
2.5	その他セクション	13
2.6	セクション区切り	15
3	記事集合からのサマリパッセージの抽出	18
3.1	サマリを含むセクション検出システム	18
3.1.1	セクション分割モジュール	19
3.1.2	ラベル付与モジュール	21
3.2	サマリパッセージ抽出システム	27
3.2.1	意見, 解説, 文章によるまとめ	27
3.2.2	箇条書によるまとめ	29
3.2.3	週間日誌	30
3.3	関連研究	31
3.3.1	要約的な個所の性質と検出方法	32
3.3.2	記事の分割方法	33

4 システムの評価	34
4.1 サマリを含むセクション検出システム	34
4.1.1 評価方法	34
4.1.2 評価結果	35
4.1.3 考察	37
4.2 サマリパッセージ抽出システム	42
4.2.1 文章によるまとめ, 意見, 解説セクション	42
4.2.2 箇条書によるまとめ	43
4.2.3 週間日誌	43
5 サマリパッセージを用いた複数記事要約	44
5.1 要約アルゴリズムの概要	44
5.2 関連研究	46
6 まとめ	48
7 今後の課題	49
7.1 記事集合からのサマリパッセージ抽出の精度向上	49
7.2 セクション分割アルゴリズムの精緻化と評価	50
7.3 サマリパッセージを用いた要約処理の詳細の検討	51
7.4 より客観性の高いデータ, ルールの作成	52

図 目 次

1.1	全体の処理の流れ	6
1.2	サマリを含むセクション検出システム	6
3.1	ラベル付与処理の大まかな流れ	22
5.1	サマリパッセージと過去の記事の対応付け	45
5.2	要約生成	45

表 目 次

3.1	「叙述文 過去」の文末表現の例	28
3.2	「叙述文 状態」の文末表現の例	28
4.1	地下鉄 サリン オウム '95	35
4.2	核実験 CTBT '96	35
4.3	山一 野村 証券 '97	35
4.4	地下鉄 サリン オウム '95 評価結果	36
4.5	核実験 CTBT '96 評価結果	36
4.6	山一 野村 証券 '97 評価結果	36
4.7	リストにない意見文文末表現の例	38

第 1 章

はじめに

1.1 研究の背景

近年, 電子化された大量の文書がオンラインで入手できるようになった. 新聞においても同様に, 新聞社各社で電子化された新聞記事を出版しており, それらは CD-ROM として出版されていたり, WWW 上で閲覧できるようになっていたりする. このように, 大量の記事に瞬時にアクセスできるようになったが, 人間の情報処理能力は限られているので, それらの膨大な量の情報全てを把握するのは困難である. そのため, 大量の記事の中から情報を得ようと思った時に, 膨大な情報の中から必要な情報だけを取捨選択するという, 人間にとって負荷の大きい作業をしなければならない. このような問題は情報洪水などと呼ばれており, 新聞記事以外の情報においても深刻な問題となっている.

大量の電子化記事の中から必要な情報を探し出す技術として広く利用されているものに情報検索がある. 電子化された新聞記事コーパスからある話題に関する記事を探す場合, この情報検索技術を用いれば, 種々雑多な話題の記事が含まれている膨大な量の新聞記事コーパスから, 自分が必要としている記事を絞り込むことができる. しかし, 検索結果が膨大になると, それらの記事全てに目を通すことは困難になる. また, 膨大な記事集合の中には, 内容が重複しているものや,それほど重要でない内容が含まれている可能性がある. 従って, 検索結果の記事集合から直接情報を得ようとするのは効率が悪く, これだけでは情報洪水の状態が解消されたとは言にくい.

そのような場合の解決策として, 検索結果の記事集合を要約する方法が挙げられる. 記事集合を要約することにより, 重要で, 冗長性の少ない情報を効率良く得ることができる.

検索結果が膨大である場合の対処法として, 要約すること以外に, 以下のものがある.

- 関連フィードバック法による再検索

- 検索結果のクラスタリング

関連フィードバック法による再検索では、最初の検索結果をユーザが評価し、その評価結果をもとに計算機がユーザの検索意図をさらに読み取り、再検索する。再検索の結果は再びユーザに評価され、検索が繰り返される。この方法により、最初の膨大な検索結果が、ユーザが必要としている情報へと絞り込まれていく。検索結果のクラスタリングでは、膨大な検索結果の記事集合を、内容が類似している記事同士を集めたクラスタの集合に変換する。この結果、ユーザはクラスタ単位で必要な情報を探することができる。

しかし、これら2つの方法によって、ある程度検索結果を絞り込むことができて、その結果は依然として複数の記事である。よって、それらの中には、それほど重要でない内容や、冗長な内容が含まれている可能性もあり、また、全てに目を通すのも面倒である。この場合も、再検索の結果やクラスタリングの結果の記事集合を要約して1つにまとめることは有効であると考えられる。

このように、情報洪水と呼ばれる現象が深刻になっている現在、複数の新聞記事を要約する技術の確立は重要といえる。

1.2 複数記事要約におけるサマリパッセージの有効性

計算機による要約は、一般に以下の2つのステップから構成される。

1. 文書中からの重要個所の抽出
2. 抽出された重要個所を元にした要約文の生成

複数の新聞記事を要約する場合も同様に、まず、複数記事中の重要個所を同定する必要がある。従来手法には、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。これらの手法は、要約対象の記事集合の内容を理解することなしに、「重要な情報は頻出する」などのヒューリスティックスを用いることにより、重要個所を近似する手法と言える。しかし、これらのヒューリスティックスにより常に正しく重要個所を同定できるわけではない。

一方、本研究では、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して重要個所を同定する手法を提案する。

以下では、まず、意見記事、解説記事、まとめ記事の性質について述べる。

意見記事、解説記事では、過去に起こったある出来事について述べ、その後、その出来事についての意見や解説を述べる、という大まかな構造を持つと考えられる。以下に意見記事の例を挙げる。

：

しかし、長野県警は毒物が何であるかまだ判明していない事件発生の翌日、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。慎重さとは別の力、つまり思い込みが働いたと、思わざるを得ない。振り返ってみれば、これが教団につながる重要な情報を見落とすことになったのではないか。

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、長野県警が河野さんに「謝罪」するまでに一年がかかった。捜査の軌道修正にこだわりを感じる。

事件の処理が一段落した時点で、県警に捜査の検証と公表を求めたい。苦い経験から教訓を得るためには、不可欠の作業と考えるからだ。

：

太字の部分が過去の出来事の記述で、その後に、その出来事に対する記者の意見が述べられている。解説記事も同様の構造を持つと考えられる。意見記事、解説記事で新聞記者が取り上げる過去の出来事は、世間で議論を呼んだ出来事や、その後の話題の進展に影響を及ぼした出来事、その話題全体を理解する上で必要となる出来事などであると考えられる。

まとめ記事では、過去の主要な出来事を、記者の意見や解説を含まずに、網羅的に記述している。後で述べるように、まとめ記事として本研究では3つのサブカテゴリを考えているが、以下に、そのサブカテゴリの内の一つで、最も出現頻度の高い、箇条書によるまとめ記事の例を挙げる。

地下鉄サリン事件以後の主な動き - - オウム真理教

月	日	内容
---	---	----

3	20	東京で地下鉄サリン事件発生
---	----	---------------

2	1	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「悔いのない死を迎えよう」などと呼びかける
---	---	---

2	2	警視庁が目黒公証役場事務長、仮谷清志さん拉致事件の逮捕監禁容疑で山梨県上九一色村などの教団関連施設の家宅捜索を開始
---	---	---

4・2・警察庁の係官2人が、教団のロシアでの活動実態を調査するためモスクワ入り

4・警察当局は山梨県富沢町の「富士清流精舎」と呼ばれる教団の工場施設などを殺人予備容疑で捜索

6・東京・赤坂のビル地下駐車場で教団の「防衛庁」長官、岐部哲也容疑者ら3人を建造物侵入容疑で逮捕。銃器の部品とみられる銃身など多数を押収

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

15・警視庁は「諜報省」トップ、井上嘉浩容疑者を公務執行妨害容疑で逮捕

16・教祖の麻原彰晃容疑者らを殺人容疑で逮捕

このように、まとめ記事では、対象となっている話題のそれまでの経緯がわかるように、過去に起きた主要な出来事を網羅的に記述している。この際、どの出来事を述べるかは、新聞記者が過去のどの出来事がその話題全体を理解する上で重要なかを考慮して決定していると考えられる。

以上のように、意見記事、解説記事、まとめ記事には、対象となっている話題の中で重要な過去の出来事を記述している。我々はこの意見記事、解説記事、まとめ記事中にある、それまでの経緯の中での重要な出来事の記述をサマリパッセージと呼ぶ。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。

前述したように、従来の手法は、要約対象の記事集合の内容を理解することなしに重要個所を同定するが、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。

1.3 研究の目的

本研究の目的は、記事集合からサマリパッセージを抽出する方法を開発することである。記事集合は、新聞記事コーパスをあるクエリで検索して収集する。

以下に、記事集合からのサマリパッセージ抽出処理の概要について述べる。この処理は

大きく分けて以下の 2 つのステップから構成される。

1. 記事集合からの意見記事, 解説記事, まとめ記事の検出
2. 検出された記事からのサマリパッセージの抽出

記事集合から直接サマリパッセージを抽出せず 1 のステップを介するのは, 意見記事, 解説記事, まとめ記事以外 (報道記事など) から過去の出来事の記述が抽出されるのを防ぐためである。報道記事にある過去の出来事の記述は, 主に最新の出来事に関する記述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。

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して, 上記 1 のステップは, カテゴリ毎のまとまりを単位として意見, 解説, まとめを検出すべきである。本研究ではこのカテゴリ毎のまとまりをセクションと呼ぶ。従って, 上記 1 のステップは実際には,

意見セクション, 解説セクション, まとめセクションの検出

というように, セクション単位の処理になる。また, これら 3 つのカテゴリのセクションは全てサマリパッセージを含むので, 一括してサマリを含むセクションと呼ぶことにする。結局, 上記 1 のステップはさらに 2 つのステップから構成され, 全体の処理の流れは以下のようなになる (図 1.1)。

1. 記事集合からのサマリを含むセクション (意見セクション, 解説セクション, まとめセクション) の検出 (図 1.2)
 - (a) 各記事のセクションへの分割
 - (b) 分割された各セクションがどのカテゴリに属するかを表すラベルの付与 (サマリを含むセクションの検出)

2. 検出されたサマリを含むセクションからのサマリパッセージの抽出

以上のような枠組でシステムを実装し, その評価を行なったので, それについて報告する。また, サマリパッセージを用いた複数記事要約の方法を示す。
なお, 本研究では, 新聞記事コーパスとして毎日新聞社のものを用いる。

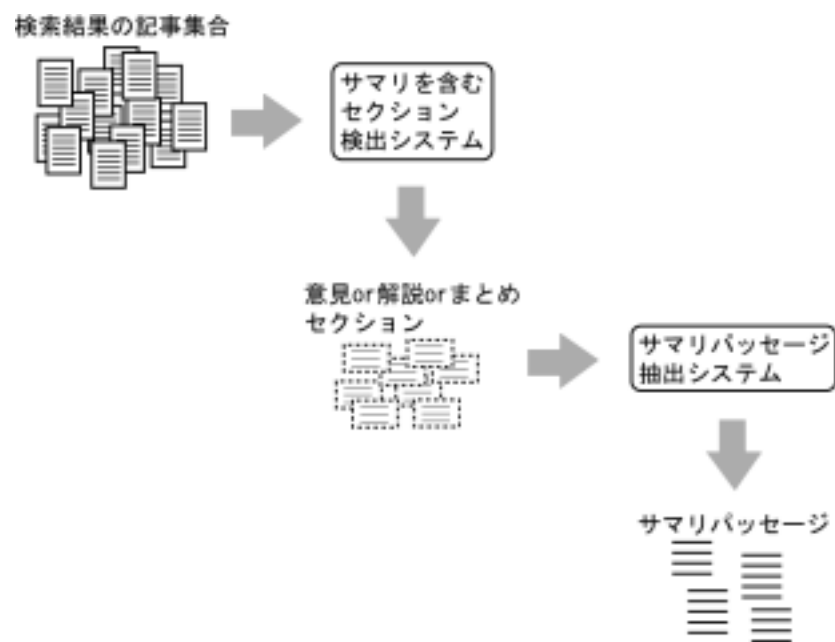


図 1.1: 全体の処理の流れ

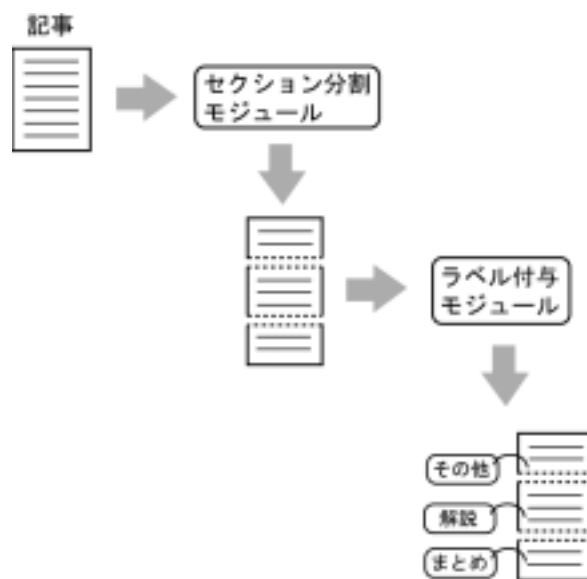


図 1.2: サマリを含むセクション検出システム

1.4 本論文の構成

2 章では, 本研究で想定している新聞記事のモデルについて説明する. 3 章では, 2 章でのモデル化に基づいた, 検索結果の記事集合からサマリパッセージを抽出するシステムについて述べる. 4 章では, そのシステムの評価結果と考察を述べる. 5 章では, サマリパッセージを用いた複数記事要約手法を提案し, 従来の要約手法と比較する. 6 章で本研究のまとめを述べ, 7 章で今後の課題について述べる.

第 2 章

本研究における新聞記事のモデル

本章では、本研究において新聞記事がどのようにモデル化されるのかについて述べる。記事集合からのサマリパッセージ抽出処理は、本章で提案するモデル化に基づく。

2.1 セクション単位の扱い

記事によっては異なるカテゴリに属する内容が 1 記事中に存在する場合がある。以下にその例を挙げる。

民事訴訟の「判決」迫り、危機感抱く - - 松本サリン事件でオウム真理教の 1
2 人逮捕

土地取得をめぐる訴訟、各地へ飛び火恐れ

松本サリン事件の標的は、オウム真理教の土地取得をめぐって争われた全国初の民事訴訟だった。長野県松本市では一九九一年から、支部建設に進出した教団と反対する住民の対立が激化。事件当時、裁判はすでに結審、約三週間後に判決が迫っており、教団側は危機感を抱いていたとみられる。

：

裁判官のサリン被害から判決言い渡しが遅れた裁判は、事件後に原告側が弁論再開を申し立て、現在は弁論が続行中で、判決の見通しは立っていない。

松本サリン事件とオウム真理教

9 1 ・ 6 ・ 1 8 オウム真理教が長野県松本市芳川野溝に 9 4 0 平方メートルの土地を買収、賃借契約

1 2 ・ 1 0 地権者が教団に賃貸した土地での建築工事禁止の仮処分を長野地裁松本支部へ申請

1 2 ・ 2 1 教団が買収した土地に松本支部完成
 9 3 ・ 2 ・ 9 オウム真理教松本進出阻止対策委員会が 1 4 万 7 1 0 2
 人分の反対署名を地裁松本支部へ提出
 9 4 ・ 5 ・ 1 0 土地明け渡し訴訟結審
 6 ・ 2 7 松本サリン事件. 7 人が死亡, 土地明け渡し訴訟の裁判
 …

この例では,

松本サリン事件とオウム真理教

の行を堺に, 前半に解説的な内容が, 後半にまとめ的な内容がある. このような記事に対し
 ては, 1 記事を単位として意見, 解説, まとめの検出やサマリパッセージ抽出を行なうので
 はなく, カテゴリ毎のまとまりを単位として行なうべきである. 従って本研究では, 1 章 3
 節で述べたように, カテゴリ毎のまとまりをセクションと呼び, セクションを単位として
 各記事进行处理する.

本研究では, セクションのカテゴリとして以下のものを考えている.

- 意見セクション
- 解説セクション
- まとめセクション
- その他セクション

以下に各セクションについて述べる.

2.2 意見セクション

意見セクションは, 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, 主張の記述があるセクションである. 以下に 1 章 3 節で挙げた意見セクションの例
 を再び挙げる.

⋮

- (1) しかし、長野県警は毒物が何であるかまだ判明していない事件発生の翌日、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搜索した。
- (2) 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。
- (3) 慎重さとは別の力、つまり思い込みが働いたと、思わざるを得ない。
- (4) 振り返ってみれば、これが教団につながる重要な情報を見落とすことになったのではないか。
- (5) やがて、捜査の目は河野さんから教団に向けられていく。
- (6) しかし、長野県警が河野さんに「謝罪」するまでに一年がかかった。
- (7) 捜査の軌道修正にこだわりを感じる。
- (8) 事件の処理が一段落した時点で、県警に捜査の検証と公表を求めたい。
- (9) 苦い経験から教訓を得るためには、不可欠の作業と考えるからだ。

⋮

この例では、(1)(5)(6)の文がサマリパッセージで、(2)(3)(4)の文が(1)に対しての、(7)(8)(9)の文が(5)(6)に対しての意見、主張である。

2.3 解説セクション

解説セクションは、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、一般的見解などの記述があるセクションである。以下に解説セクションの例を挙げる。

⋮

- (1) 長野県松本市では一九九一年から、支部建設に進出した教団と反対する住民の対立が激化。
- (2) 事件当時、裁判はすでに結審、約三週間後に判決が迫っており、教団側は危機感を抱いていたとみられる。

⋮

(1)の文がサマリパッセージで、(2)の文が(1)に対しての解説である。

2.4 まとめセクション

過去の主要な出来事が、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である。

まとめセクションには以下の3つのサブカテゴリがある。

- 箇条書によるまとめ
- 文章によるまとめ
- 週間日誌

箇条書によるまとめの例は1章で挙げた。

以下に、文章によるまとめの例を挙げる。

1963年8月、部分的核実験停止条約（PTBT）に米、英、ソ連（当時）が調印して発効（中国と仏は未加盟）。大気圏、宇宙、水中の核実験は禁止されたが、地下核実験は継続された。

68年7月、核兵器非保有国だけに国際原子力機関による査察を義務付けた核拡散防止条約（NPT）に62カ国が調印、70年発効。日本は「核保有国の核軍縮義務が不明確」として当初署名しなかったが、76年に加盟。昨年末現在の加盟国は175カ国。

CTBT交渉について当初は核保有国の反対意見が根強かったが、93年7月に米英が核実験停止措置を継続しつつ「核実験全面禁止の可能性を探る予備協議」を開始、仏露も同調した。同8月のジュネーブ軍縮会議で「核実験禁止特別委」に条約交渉権限を与えるよう、PTBT未加盟の中仏も含めた核保有国の賛成を得て全会一致で決定、94年1月から本格交渉が始まった。

昨年5月、NPT発効25年を迎え、再検討・延長会議で無期限延長を決定。

このように、それまでの経緯を、意見、解説的な内容を含まないで、文章で網羅的にまとめている個所が、数は少ないが存在する。

週間日誌は毎日新聞で連載されている記事で、その週にあった主要な出来事を簡潔にまとめている。週間日誌は毎日新聞独自の記事だが、他の新聞にも、週間日誌に相当する、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。従って週間日誌を処理対象とすることは、本研究の手法が毎日新聞に特化していることを意味するわけではない。以下に週間日誌の例を挙げる。

[週間日誌] 3 月 2 0 日 < 月 > ~ 3 月 2 6 日 < 日 >

【 2 0 日 】

東京の地下鉄でサリンを使った無差別テロ. 1 0 人が死亡, 5 4 9 3 人が治療を受け, 7 1 4 人が入院. 朝のラッシュ時, 東京の営団地下鉄日比谷, 丸ノ内, 千代田 3 線の電車 5 本の車内で, サリンを発生させる組織犯行. 警視庁と東京地検が捜査本部.

東京協和, 安全両信組を引き継ぐ東京共同銀行が営業開始.

⋮

【 2 1 日 】

世界貿易機構 (W T O) 事務局長にイタリアのルッジェロ元貿易相が内定.

【 2 2 日 】

警視庁がオウム真理教を強制捜査. 目黒公証役場事務長, 仮谷清志さん拉致 (らち) 事件で. 警官 2 5 0 0 人を動員, 2 5 力所を捜索. 自衛隊から戦闘用防護服と防毒マスク 1 0 0 0 着を借りて. 地下鉄サリン事件で検出されたアセトニトリルなど大量の薬品を押収. 意識不明で監禁されていた 6 人を保護. 医師ら 4 人を監禁容疑で逮捕. いずれも山梨県上九一色村の教団施設で. (以下 はオウム真理教関連) .

与党 3 党が訪朝団派遣で合意. 1 9 9 0 年の「三党共同宣言」については「歴史的事実として認識する」.

⋮

【 2 3 日 】

統一地方選始まる. 1 3 知事選告示, 4 9 人が立候補. 2 6 日, 札幌市長選告示, 2 人が立候補.

山梨県上九一色村施設の捜索でサリンの原料となる薬品押収. 警視庁は施設内でサリンの生成が行われたと断定.

⋮

上の例は, 95 年の毎日新聞コーパスを「地下鉄 サリン オウム」というクエリで検索した結果の記事集合中に含まれていた記事だが, クエリの話題以外の出来事についても記述されている. このように, 週間日誌には, 記事集合が対象にしている話題 (上の例では「地下鉄 サリン オウム」) 以外の話題であっても, 例えば「統一地方選開始」など, その一週間に起きた主要な出来事であれば記述されている. そのため, 週間日誌からのサマリパッセージ抽出では, 記事集合が対象にしている話題に関する, 過去の主要な出来事の記述のみを抽出するようにしなくてはならない.

2.5 その他セクション

意見, 解説, まとめセクション以外のセクションとして, 以下の 5 つのサブカテゴリが挙げられる.

- 報道セクション
- 対象の話題に関するエピソードを述べているセクション
- 裁判記録, 外交文書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

報道セクションは最新の出来事に関する記述や, それを補う情報の記述がほとんどを占めており, 意見, 解説, まとめセクションと違い, 対象の話題に関するそれまでの経緯の中でどの出来事が重要なのかを判断する手がかりはほとんど無いと考えられる. 従って, 報道セクションはサマリパッセージ抽出の対象から除外する. 以下に報道セクションの例を挙げる.

「地下鉄サリン」と同一薬品, オウム真理教施設で押収...独自製法か, 化学反応の促進剤

地下鉄サリン事件で, 現場に残ったサリンの分析を進めている警視庁など捜査当局は三十一日までに, 特殊な化学薬品を新たに検出, 同じ薬品が山梨県上九一色(かみくいしき)村のオウム真理教施設に保管されていたことを突き止めた. この薬品は, サリン製造に必要な薬品類とは全くの別種で, 化学反応の試薬として用いられている. 捜査当局は, 犯人が独自に開発した技術として, 製造中のサリンに薬品を添加したとみて, 薬品がサリン製造犯特定の決め手となるとしている. 同じ薬品が教団施設から見付かったことで, 捜査当局は地下鉄サリン事件とオウム真理教との関連が極めて濃厚になったとの見方を強めている.(社会面に関連記事)

⋮

これは 31 日に出版された記事の一部であり, その日までに起きた最新の出来事とそれに関連した情報について記述している.

エピソードを述べているセクションでは, 対象の話題の中では周辺的な情報と考えられる内容を述べている. 以下にエピソードを述べているセクションの例を挙げる.

：

車内は座っている人のほか立っている人が数人程度の込み具合だったが、動き始めて一人の男性が顔を真っ赤にして倒れ、乗客の一人が緊急停止ボタンを押し、乗務員を呼んだという。そのころには、石井さん自身も気分が悪くなり、頭が痛くなってきたという。

赤ちゃん、来月生まれるのに…

兜町の中島病院では、近くの営団地下鉄茅場町駅から体の異常を訴えてきた乗客らが救急車で次々に運ばれ、待合室は百人以上の人であふれた。足立区、会社員、和田栄二さん（29）は死亡した。同病院では非番の医師や看護婦も緊急に呼び出し、医師は通常の倍の六人で治療に当たった。同病院は「患者が殺到しているため、軽症の人は応急的な手当てをするだけで帰ってもらい、後日改めて来てもらうよう対応している」と話した。

：

この記事では、地下鉄サリン事件発生時に現場にいた人々の状況を述べている。このようなセクションも、それまでの主な経緯を知る上で有効ではないと判断した。従って、エピソードを述べているセクションもサマリパッセージ抽出の対象から除外する。

以下に、公文書をそのまま載せてあるセクションの例を挙げる。

96年版・外交青書〈要旨〉

10日の閣議に報告された1996年版外交青書の要旨は次の通り。

第1章 総括—95年の国際社会

〈概観〉

（1）国際情勢認識＝国際情勢には依然流動的な要素が多く、いまだ新たな秩序の確立には時間を要する。しかし95年は新たな国際秩序の萌芽（ぼうが）を示すいくつかの好ましい進展が見られた。朝鮮民主主義人民共和国（北朝鮮）の核開発問題解決に向けた動き、旧ユーゴにおける和平合意、パレスチナ暫定自治拡大合意等に見られるように…

このセクションは、96年の毎日新聞コーパスを「核実験 CTBT」で検索した結果に含まれていた記事の一部で、外交文書をそのまま載せている。このような文書は、対象の話題に関する参考資料として新聞に掲載されているものと考えられ、当然、新聞記者が書いたものではなく、意見セクション、解説セクション、まとめセクションのように、対象の話題のそれまでの経緯における重要な出来事に触れるような性質のものではない。従って、公文書をそのまま載せているセクションもサマリパッセージ抽出の対象から除外する。

用語解説セクションでは、対象の話題に現われるキーワードや人物などを説明している。以下にその例を挙げる。

【株主総会】 株式会社の株主が持ち株に応じて議決権を行使し、会社の意思を決める最高機関。「特殊株主」と呼ばれる総会屋に不祥事などで付け込まれ、総会が長時間に及ぶのを恐れ、議事進行を円滑に進め、穏便に終わらせようとするのが利益供与事件の背景にある。

⋮

用語解説は、そのキーワードや人物に対する一般的な説明を与えるもので、対象の話題の経緯や重要な出来事などを記述するものではない。用語解説セクションもサマリパッセージ抽出対象から除外する。

一覧表の形式のセクションでは、ある同じ種類の大量の情報を、説明なしで列挙しているセクションである。以下にその例を挙げる。

< 計画・指示 >	起訴罪名
麻原彰晃 (4 0)	殺人
< 実行 >	
新実智光 (3 1)	殺人
遠藤誠一 (3 5)	殺人
中川智正 (3 2)	殺人
富田隆 (3 7)	殺人
⋮	

このようなセクションは、同じ形式で表すことができる、同じ種類の大量の情報を、コンパクトに伝えることが目的と考えられる。このように伝えられる情報は、対象の話題のそれまでの経緯や、対象の話題における重要な出来事に関する情報を述べるものではない。従って、これらのセクションもサマリパッセージ抽出対象から除外する。

2.6 セクション区切り

これまで意見、解説、まとめ、その他セクションについて説明してきたが、これらのセクションが一記事中に混在する場合、セクションの区切れ目がどのように現われるかを以下で述べる。以下ではセクションの区切れ目をセクション区切りと呼ぶことにする。

以下にセクション区切りの例を挙げる。

：

捜査当局は村井氏が直前物質の化合物を保管していた点を重視。土谷容疑者とは別に村井氏も独自にサリンの製造を試みるなかで、まだ直前物質が残っている可能性もあるとみている。

土谷容疑者の供述によるサリン製造の経過

1993年 8月 第7サティアン横にクシティガルバ棟が完成
11月 サンプルとしてサリン約20グラムを生成続いて化学班メンバーを使ってサリン約1キロを生成、容器に保管
12月ごろ サリン約5キロを生成
1994年 3月 科学技術省メンバー設置の「スーパーハウス」で

：

この例は、前半にその他（報道）セクションが、後半にまとめセクションがある記事だが、

土谷容疑者の供述によるサリン製造の経過

の行がセクション区切りとなっている。また、以下の例は解説セクションの次に意見セクションがくる記事で、

巧妙さに欠ける——作家、佐木隆三さんの話

の行がセクション区切りである。

：

レベルで考え得ることだ。袋が狙い通りに燃えたとしても、硫酸だけが広がり反応はごく一部分でしか起こらない。シアン化水素の大量発生が目的なら、稚拙な方法だという印象がある。しかし、入手が容易ではないし、殺人につながることから、単なるいたずらとは考えにくい」と話している。

巧妙さに欠ける——作家、佐木隆三さんの話

火をつけた点は地下鉄サリンや横浜異臭事件と手口が違い、巧妙さに欠ける気がする。（麻原彰晃代表の「予言」で）四月十五日に騒ぎになった新宿での犯行という点でも、便乗犯ではないか。だが、大量殺人につながる恐れもあり、便乗犯と片付けるには悪質すぎる。

：

このように、セクション区切りは記事中で明示されている。

これまでに、文書を意味的なまとまりに分割するテキストセグメンテーションの研究は数多くなされてきた。それらの研究では、語彙的結束性や接続詞、照応表現、文タイプの情報などの言語的な情報が用いられてきた。一方、本研究で行なう文書の分割は、意味的なまとまりへの分割ではなく、カテゴリ毎のまとまりであるセクションへの分割である。この場合、先に観たように、記事を明示的に分割している行が直接的にセクション区切りを示しているので、従来の研究で用いられてきたような言語的な情報より、記事中のレイアウト情報を用いた方が良い。

第 3 章

記事集合からのサマリパッセージの抽出

本研究では,2 章で説明した記事のモデルに基づき, 検索結果の記事集合からサマリパッセージを抽出するシステムを実装した. システム全体の処理は 1 章で示した以下の流れに従う (図 1.1).

1. 記事集合からのサマリを含むセクション (意見セクション, 解説セクション, まとめセクション) の検出 (図 1.2)
 - (a) 各記事のセクションへの分割
 - (b) 分割された各セクションがどのカテゴリに属するかを表すラベルの付与 (サマリを含むセクションの検出)
2. 検出されたサマリを含むセクションからのサマリパッセージの抽出

本章では, まず, サマリを含むセクション検出システムについて述べ, その後, サマリパッセージ抽出システムについて述べる.

3.1 サマリを含むセクション検出システム

サマリを含むセクション検出システムは,1 記事ずつ入力として受けとり, まず 1) 各記事をセクションに分割し, その後 2) 分割された各セクションに対して, そのセクションが属するカテゴリを表すラベルを付与して出力する. 従ってこのシステムは, 図 1.2に示されているように,

- 1) セクション分割モジュール

2) ラベル付与モジュール

の2つのモジュールから構成される。

3.1.1 セクション分割モジュール

2章で述べたように、セクション区切りは、ほとんどの場合記事中で明示されている。観察の結果、セクション区切りには以下のようなタイプがある。

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

以下は1のタイプの例である。

⋮
は第一通報者。被害者として事情を聴き、捜査の協力者と認識している。あくまで広範な捜査の一環」と話している。

人権感覚希薄な風潮、歯止めを - - ジャーナリスト、大谷昭宏氏の話
松本サリン事件に携わった記者たちの取材上の困難を知ったうえで、あえていくつかの問題点を指摘したい。まず、証言でも証拠でも、決定的に「裏取

⋮
この例では、

人権感覚希薄な風潮、歯止めを - - ジャーナリスト、大谷昭宏氏の話
が解説セクションと意見セクションのセクション区切り行である。

以下はタイプ2の例である。

⋮
争中の土地明け渡し訴訟を担当する裁判官が住む官舎があり、捜査当局はこの裁判官らを狙ったものとみている。

< 地下鉄サリン事件容疑者一覧 >

容疑者名	年齢	教団内の所属
麻原彰晃	容疑者 (4 0)	代表 ● 計
中川智正	容疑者 (3 2)	法皇内庁 ● 医

⋮

この例では,

<地下鉄サリン事件容疑者一覧>

が解説セクションとその他(一覧表)セクションのセクション区切り行である.

タイプ3の例は以下のようなものである.

⋮

とりでに奇跡や超能力があらわれていると思います。(聞き手, 編集委員・横山真佳) = つづく

.....

1934年山形県生まれ. 仙台のカトリック系施設から上智大卒. 放送作家を経て, 72年「手鎖心中」で直木賞. 次いで奇想天外な独立国家物語「吉里

⋮

この例では, 「.....」の行が意見セクションとその他(用語解説)セクションのセクション区切り行である.

セクション分割モジュールは, 上で述べたような3つのタイプのセクション区切りを検出し, 記事をセクションに分割する.

しかし, 記事中にセクション区切りが連続して現われる場合があり, その場合には, 中身が無い(文が0)のセクションが出来てしまう. 例えば以下のような場合である.

⋮

ゲドンをつくり出さなければならなくなっただけです. そこで国家の権威の象徴である「霞が関」を狙った.

若者論

現実と虚構同一視——朝倉氏

ゲーム感覚と同じ——佐々氏

司会 現代若者論風に分析するとどのようなことなのでしょう.

朝倉 ぼくらの世代は現実と神秘的な世界は分けて考えるが, 若い人は二

⋮

そこで, 1つ前の行がセクション区切りかどうかを調べて, もしそうであれば現在の行をセクション区切りとしないようにしている.

従来のテキストセグメンテーションの手法が文書中の言語的な情報を用いるのに対し, 本研究の手法は, 記事中のレイアウト情報を利用する手法と言える. 2章で述べたように,

カテゴリ毎のまとまりは記事中で明示的に分割されるようにレイアウトされているので、本研究のように記事をカテゴリ毎のまとまりに分割する場合は、言語的な情報を用いるよりレイアウト情報を用いた方がよい。

3.1.2 ラベル付与モジュール

ラベル付与モジュールでは、1 セクションずつ入力として受け取り、各セクションに対して、意見、解説、まとめ、その他のいずれかのカテゴリに対応するラベルを付与する。図 3.1 で示されているように、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、意見、解説、まとめ、その他のいずれかのカテゴリに対応している。入力されたセクションは、図 3.1 で示されている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て、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。

9 つのフィルタは以下の 3 タイプに分けられ、タイプ 1、タイプ 2、タイプ 3 の順に起動するようにしている。

1. 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)
2. 内容から判断するセクション用のフィルタ (5 ~ 7)
3. 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)

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 (333 セクション)
- 96 年の記事を「核実験 CTBT」で検索した結果の上位 100 記事 (114 セクション)

意見セクション

新聞記事の中には社説記事がある。社説記事は対象の話題の過去の出来事に対して意見、主張を述べているので、社説記事は全体で 1 つの意見セクションである。社説記事には、以下のように、見出しに必ず文字列「社説」が含まれている。

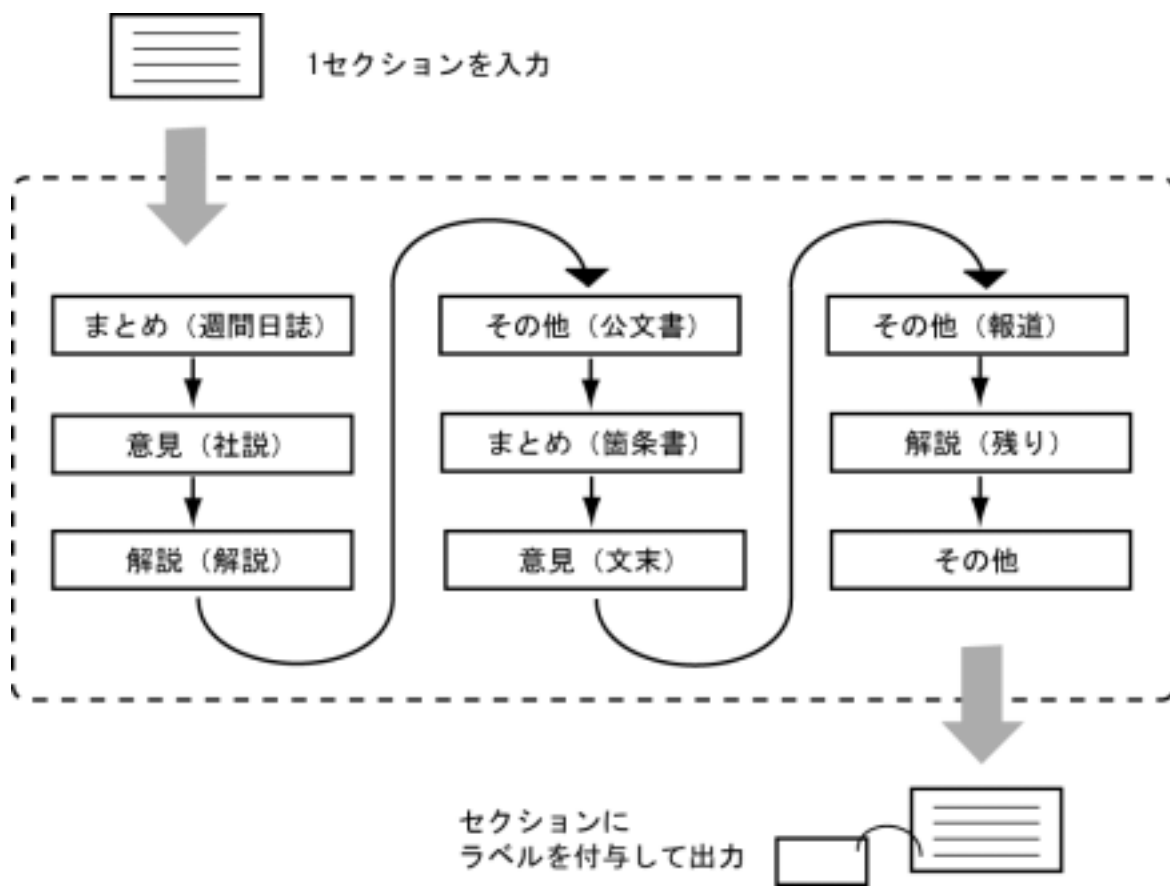


図 3.1: ラベル付与処理の流れ

[社説] 社会 教団をめぐる疑惑解明を オウム真理教

山梨県上九一色村にあるオウム真理教の施設に対する捜索は、拉致容疑から殺人予備容疑に切り替わった。捜査の流れが新たな段階に入ったことを示す

：

よって、見出しに文字列「社説」が含まれているセクションには、2 番目のフィルタ (意見 (社説)) によって意見ラベルが付与される。

また、意見や主張を述べている文が占める割合が高いセクションは意見セクションなので、6 番目のフィルタ (意見 (文末)) によって意見ラベルが付与される。本研究では、意見や主張を述べている文を意見文と呼ぶ。例えば、以下の文は意見文である。

1. しかし、うろたえる必要はない。
2. この事実は、一連の事件の本質を象徴しているように思えてならない。
3. 仮谷さんら致事件の捜査だから人命にもかかわるため全国一斉にやらざるを得ない。

意見 (文末) フィルタでは、ある文が意見文かどうかは、その文の文末表現から判断する ([6])。上の例では、それぞれ「必要はない」、「思えてならない」、「を得ない」が意見を述べていることを示す文末表現である。

解説セクション

新聞記事の中には、以下のように、見出しに文字列「解説」が含まれている記事がある。

[解説] 初動のつまずき響く - - 松本サリン事件、オウム真理教の 1 2 人逮捕

「信教の事由」も壁に

容疑者逮捕まで約一年を要した松本サリン事件。化学兵器という国内では未知の凶器が使われた前例のない大量殺人に、警察の捜査は混乱、長期化し、

：

これらの記事は対象の話題についての解説を述べているので、1 記事全体で解説セクションである。よって、見出しに文字列「解説」が含まれているセクションには、3 番目のフィルタ (解説 (解説)) によって解説ラベルが付与される。

しかし、解説セクションの中には見出しで判断できないものもある。現在のところ、そのような解説セクションの特徴が掴めていない。そこで、1~7 番のフィルタにかからずに残ったセクションを解説セクションの候補と考え、以下の 2 つの条件を満たすセクションに対して、8 番目のフィルタ (解説 (残り)) が解説ラベルを付与するようにした。

1. 1～7 番のフィルタにラベルを付与されず、残ったセクションの中で、
2. セクション中に閾値 (現在は 5 文) 以上の文がある。

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

2 章で述べたように、まとめセクションには以下の 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

週間日誌の見出しには、以下のように、文字列「週間日誌」が必ずある。

[週間日誌] 3 月 2 0 日 < 月 > ~ 3 月 2 6 日 < 日 >

【 2 0 日 】

東京の地下鉄でサリンを使った無差別テロ. 1 0 人が死亡, 5 4 9 3 人が治
 ……

従って、1 番目のフィルタ (まとめ (週間日誌)) によって、見出しに文字列「週間日誌」が含まれるセクションに対してまとめラベルが付与される。

箇条書によるまとめの一行一行には以下の特徴がある。

1. 行頭に日付を表す数字がきて、
2. その次に、その日に起きた出来事の説明がくる。

例えば以下のようなものである。

……

1 ・ 2 7	仏が再開 6 回目の核実験実施
2 9	シラク大統領が仏の核実験終了を宣言
3 0	中国外務省が「 C T B T 締結・発効まで核実験を継続」と言明

- 3・25 米英仏が南太平洋非核地帯条約（ラロトンガ条約）に調印
- 4・11 アフリカ非核化条約（ペリンダバ条約）にアフリカ49カ国
が調印
- ⋮

また、箇条書によるまとめセクションには、以下のように、日付を表す数字の前に などの記号が先にくる場合がある。

- ⋮
- 1994・4・27 講演の中で麻原彰晃代表が、自分の…
- 95・1・4 記者会見で教団幹部は「オウム真理教の諸…
- 1・7 教団のラジオ放送が「（一部テレビが）儀…
- ⋮

そこで、5 番目のフィルタ（まとめ（箇条書））によって、以下のようなセクションにまとめラベルが付与される。

先頭に 0 か 1 個の文字がきて、次に 0 個以上の空白、次に数字がくる行が、閾値以上連続して現われるセクション。

現在は閾値を 3 としている。

文章によるまとめには、現状のシステムでは、それと認識することができず、まとめラベルの付与は不可能である。

その他セクション

2 章で述べたように、その他セクションには以下の 5 つのサブカテゴリがある。

- 報道セクション
- 裁判記録や外公文書などの公文書をそのまま載せてあるセクション
- 事件当日の一日の流れや、その話題にまつわるエピソードを述べているセクション
- 用語解説セクション
- 一覧表セクション

観察の結果、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ほとんどである。特徴の1番目は[11]を参考にした。

1. セクションの先頭の閾値 (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) 日以内の日付とする)
3. 「(～面に関連記事)」を含む

以下に報道セクションの先頭の数文を例として挙げる。

長野県松本市内の住宅街で昨年六月, サリンがまかれ七人が死亡, 六百人近くが重軽症を負った事件で, 警視庁・長野県警合同捜査本部は十六日 午後, オウム真理教代表の麻原彰晃被告 (4 0) = 本名・松本智津夫 = ら十二人を殺人, 殺人未遂容疑で再逮捕した。捜査本部は, 長野地裁松本支部で係争中の教団を被告とする民事訴訟の妨害が直接の動機だったとみている。教団はこの事件の後, 武装化路線を一層強め, 地下鉄サリン事件などを引き起こしており, 一連の事件の「原点」にもあたるとして, 徹底した解明を急ぐ。(3 面に解説, 社会面に関連記事)

⋮

第1文において, 特徴の1と2が現われている。特徴1を太字で, 特徴2を下線で示した。また, 例の最後に特徴3が「(3 面に解説, 社会面に関連記事)」として現われている。

この3つの特徴のいずれかがあれば, そのセクションは報道セクションである可能性が高いので, 7番目のフィルタ (その他 (報道)) によって, その他ラベルが付与される。

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。以下に公文書をそのまま載せてあるセクションの見出しの例を挙げる。

- [オウム裁判] 中川智正被告への検察側冒頭陳述要旨 / …
- 96年版・外交青書<要旨>
- オウム真理教に対する破防法適用に向け、官報に公示 (全文) …

そこで、見出しにそのような文字列があるかを調べ、もしあれば 4 番目のフィルタ (その他 (公文書)) がその他ラベルを付与する。

どのラベルも付与されず、8 番目のフィルタによって解説ラベルも付与されなかった記事に対しては、9 番目のフィルタ (その他) が最終的にその他ラベルを付与する。

現状のシステムでは、エピソードと用語解説、一覧表を、それと認識することはできない。

システムの出力例

システムは、例えば以下のように出力する。

```
...
950706165.txt
S0:解説
950606443.txt    ← 記事
S0:その他        ← 第 1 セクションに付与されたラベル
S1:まとめ        ← 第 2 セクションに付与されたラベル
950717151.txt
S0:解説
S1:まとめ
951227027.txt
S0:意見
...
```

この例の記事 950606443.txt にはセクションが 2 つあり、それぞれ、その他ラベルとまとめラベルが付与されている。

3.2 サマリパッセージ抽出システム

サマリを含むセクション検出後、検出された各セクション毎に、サマリパッセージ抽出システムに入力される。抽出方法は、セクションの形式 (文章形式か、箇条書形式か、週間日誌か) に応じて異なる。

3.2.1 意見、解説、文章によるまとめ

意見、解説、文章によるまとめセクションは通常の記事形式で構成されている。文章形式のサマリを含むセクションにおける過去の出来事の記述 (サマリパッセージ) の例を以下

に挙げる.

- 国際的孤立を避けるため, 今月 13 日, C T B T 交渉で「平和的核爆発」の禁止除外問題で妥協の用意があると明言した.
- 国連加盟国 (185 カ国) 全体の 85 % に上り, 圧倒的多数による条約成立となった.
- 5 核保有国の支持を受けたオーストラリアや日本が国連総会本会議での多数決採択に持ち込んでいた.

これらの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.

- 叙述文 過去
- 叙述文 状態 (現在 | 過去)

文のタイプは文末表現をもとに判定する. 表 3.1 に「叙述文 過去」の文末表現の例を, 表 3.2 に「叙述文 状態」の文末表現の例を挙げる.

表 3.1: 「叙述文 過去」の文末表現の例

きた	した	えた
じた	だった	あった
なった	られた	なかった

表 3.2: 「叙述文 状態」の文末表現の例

ている	でいる	ていない
てこない	でいます	でいた
ってきた	できた	ていなかった

例えば, 以下の意見セクションの一部からは,

中国が今年 6 月に核実験を再開したさい, 私たちは核実験をきっぱりとやめ, 不名誉な記録をこれ以上塗り替えないように訴えた. その声が中国指導者の耳に届かなかったことは誠に残念だ.

以下のサマリパッセージが抽出される。

中国が今年 6 月に核実験を再開したさい、私たちは核実験をきっぱりとやめ、
不名誉な記録をこれ以上塗り替えないように訴えた。

3.2.2 箇条書によるまとめ

箇条書によるまとめセクションからのサマリパッセージ抽出では、箇条書全体を抽出する。以下に箇条書によるまとめセクションの例を示す。

⋮
1 9 8 9 . 8 . 2 5 <オウム真理教宗教法人認証>
 1 1 . 3 弁護士一家失跡
 横浜市の弁護士、坂本堤さん（当時 3 3 歳）の
 一家 3 人が行方不明に。
9 0 . 2 . 1 8 <総選挙で 2 5 人全員落選>
 1 0 . 2 2 波野村国土法違反
⋮

この例にあるように、箇条書によるまとめセクションの全ての行が、先頭に数字（日付）を持っているわけではないので、単純に、先頭に数字のある行を抽出するだけではうまくいかない。そこで、先頭に数字のある行だけでなく、空白 2 個以上でインデントされている行も出力するようにした。但し、以下のような間違いが起きるので、インデントされている行を出力するには、それ以前に、先頭に数字がある行が出現していなければならない、という条件をつけた。

⋮
先進医学だけでは救えない患者を前に、患者のためというより、自らの
心のよりどころとして、「魂の救済」を目指す宗教が必要だったのだろ
うか。

■ ■ ■ (← このような行が出力されてしまう)

妻（ 4 6 ）も、林被告と一緒に教団幹部の指紋を消す手術に加わり、執
行猶予付き有罪が確定した。

⋮

また、先頭が数字の行やインデントされている行が1, 2行しか続いてなかったら、それが箇条書によるまとめセクションである可能性は低いので、先頭が数字の行やインデントされている行が、閾値 (現在は3行) 以上連続していなくてはならない、という条件もつけた。

3.2.3 週間日誌

週間日誌は以下のような構造になっている。

[週間日誌] ○月 日<月> ~ ○月▲日<日>

【 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

...

...

【▲日】

その日の出来事 1

その日の出来事 2

...

訃報

その週に亡くなった人について

語録

その週にあった著名人のセリフ

その日の出来事では、その週に起きた種々雑多な話題を一行ずつ記述している。その中には、記事を検索した際の検索意図から外れている話題も含まれている。そこで、記事を検索した時に使用したクエリを含む行だけを出力するようにした。また、出力される各行の先頭に、月日の情報を補うようにした。例えば、クエリが「地下鉄 サリン オウム」の場合、以下のような週間日誌からは、

[週間日誌] 3月20日<月> ~ 3月26日<日>

【20日】

東京の地下鉄でサリンを使った無差別テロ. 10人が死亡, 5493人が治療を受け, 714人が入院. 朝のラッシュ時, 東京の営団地下鉄...

東京協和, 安全両信組を引き継ぐ東京共同銀行が営業開始.

大阪府知事選に横山ノック(山田勇)参院議員が出馬表明.

⋮

【21日】

世界貿易機構(WTO)事務局長にイタリアのルッジェロ元貿易相が内定.

【22日】

警視庁がオウム真理教を強制捜査. 目黒公証役場事務長, 仮谷清志さん拉致(らち)事件で. 警官2500人を動員, 25カ所を搜索. 自衛隊から...

与党3党が訪朝団派遣で合意. 1990年の「三党共同宣言」に...

シンガポールでフィリピン人のメイドに死刑が執行された問題...

⋮

【23日】

統一地方選始まる. 13知事選告示, 49人が立候補. 26日,...

山梨県上九一色村施設の搜索でサリンの原料となる薬品押収. 警視庁は施設内でサリンの生成が行われたと断定.

⋮

以下のようなサマリパッセージが抽出される.

3月20日: 東京の地下鉄でサリンを使った無差別テロ. 10人が死亡, 5493人が治療を受け, 714人が入院. 朝のラッシュ時, 東京の営団地下鉄...

3月22日: 警視庁がオウム真理教を強制捜査. 目黒公証役場事務長, 仮谷清志さん拉致(らち)事件で. 警官2500人を動員, 25カ所を搜索. 自衛隊から...

3月23日: 山梨県上九一色村施設の搜索でサリンの原料となる薬品押収. 警視庁は施設内でサリンの生成が行われたと断定.

⋮

3.3 関連研究

[4]では,¹記事を談話セグメントに分割し, 各セグメントの役割を解析する手法を提案している. セグメントの役割として, Support segment, Summary segment, Anecdotal

segment の 3 つを挙げている。Summary segment は、その 1 記事の重要なポイントを要約しているセグメントである。[4] では、システムが処理対象の 1 記事から検出した Summary segment の情報を、その 1 記事からの重要文抽出タスクにおいて利用することで、重要文抽出の精度が上がることを示している。

[4] は、要約処理対象の記事の中から、その処理対象全体の要約的な個所 ([4] においては”Summary segment”，本研究においては「サマリパッセージ」) を検出し、記事の要約処理に利用する、という点で本研究と関連しているが、以下の点で本研究と異なる。

1. 検出すべき要約的な個所の性質と検出方法
2. 記事の分割方法

3.3.1 要約的な個所の性質と検出方法

[4] では、分割されたセグメント集合から Summary segment を検出する際、以下の情報を用いている。

- セグメントの出現位置
- セグメント中出现する語の重要度

[4] では「Summary segment は 1 記事中の前半 20 %か後半 20 %に出現する」と仮定している。また「Summary segment では重要な内容が述べられており、重要な語を多く含む」と仮定している。語の重要度は、「重要な内容は繰り返し述べられる」と仮定し、出現頻度などをもとに計算する。従って、記事の前半 20 %か後半 20 %に出現しているセグメントで、重要度の高い語が多く含まれていれば Summary segment と見なす。

1 記事を対象とする場合、記事のリード部分に要約的な内容が書かれていることは広く知られており、1 記事の要約的な個所の検出に位置情報を使うことは有効である。また、1 記事中で述べられる重要な話題は 1 つであることが多いので、記事内でその話題に関する語が繰り返し現われる可能性は高く、語の出現頻度から重要な内容が述べられている個所を検出する方法も有効であると考えられる。しかし複数記事を対象とする場合、複数記事全体の内容を要約しているような個所が決まった位置に現われることはない。また、頻出する内容が重要な内容という近似も、うまくいかないことが多いと考えられる。なぜなら、複数記事の場合、述べられている話題の数が多く、重要な内容であっても、複数記事内で、他の内容と比べて、顕著なほど繰り返されることがない場合もあると考えられるからである。

また、複数記事を対象に、語の出現頻度などの統計的な情報を得ようとするとなると大変なコストがかかる。

本研究では、1 章で述べたように、意見、解説、まとめセクションが要約的な個所と考え、表層的な特徴を手がかりに、これらのセクションの検出を行なう。これらのセクション中では、対象の話題を熟知している新聞記者が、話題のそれまでの経緯の中で何が重要かを考えて過去の出来事をまとめている。従って、意見、解説、まとめセクションが複数記事全体の内容の要約的な個所と考えられ、これらのセクションの検出が要約的な個所の検出になる。これらのセクションは、その文章中の表現や記述形式に特徴があるので、それを手がかりに検出することが可能である。また、表層的な特徴を手がかりとして検出するので計算コストも低い。

3.3.2 記事の分割方法

[4] では、記事を談話セグメントに分割する際、語彙的結束性の情報を用いている。それに対し本研究では、記事中のレイアウト情報を用いる。2 章で述べたように、記事を意味的なまとまりである談話セグメントに分割する際は語彙的結束性などの言語的な情報が有効だが、本研究のように、意見、解説、まとめなどのカテゴリ毎のまとまりに分割する際は、明示的に記事を区切っているレイアウト情報がカテゴリ毎のまとまりを直接的に反映しているため、レイアウト情報を用いた方が良い。

第 4 章

システムの評価

この章では、サマリを含むセクション検出システムとサマリパッセージ抽出システムの評価について述べる。サマリを含むセクション検出システムは、ラベル付与モジュールのみを、実験によって定量的に評価し、それに基づき考察する。サマリパッセージ抽出システムについては処理結果を定性的に論じる。

4.1 サマリを含むセクション検出システム

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み評価を行った。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。ラベル付与の精度を以下の式で測定する。

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

(正解セクション … 正解のラベルが付与されているセクション)

評価に用いたデータは以下の3つである.

- 毎日新聞'95 を“地下鉄 サリン オウム”で検索した結果の上位 300 記事 (333 セクション)
- 毎日新聞'96 を” 核実験 CTBT”で検索した結果の上位 100 記事 (114 セクション)
- 毎日新聞'97 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)

これらのデータのうち,1,2 はラベル付与モジュールのルール作成に用いた. 各データの内訳をそれぞれ表 4.1, 表 4.2, 表 4.3に示す. 「意見」「解説」「まとめ」「サマリ」はそれぞれ意見セクション, 解説セクション, まとめセクション, サマリを含むセクションを意味する.

表 4.1: 地下鉄 サリン オウム '95

	意見	解説	まとめ	サマリ
セクション数	58	64	23	145

表 4.2: 核実験 CTBT '96

	意見	解説	まとめ	サマリ
セクション数	19	49	10	78

表 4.3: 山一 野村 証券 '97

	意見	解説	まとめ	サマリ
セクション数	13	50	17	80

4.1.2 評価結果

各データ毎の評価結果を表 4.4, 表 4.5, 表 4.6に示す.

表中の () 内の数値は再現率, 精度の分母, 分子に相当する.

表 4.4: 地下鉄 サリン オウム '95 評価結果

	意見	解説	まとめ	サマリ
Recall	66 % (=35/58)	81 % (=52/64)	87 % (=20/23)	90 % (=130/145)
Precision	93 % (=38/41)	56 % (=52/93)	80 % (=20/25)	82 % (=130/159)

表 4.5: 核実験 CTBT '96 評価結果

	意見	解説	まとめ	サマリ
Recall	79 % (=15/19)	63 % (=31/49)	80 % (=8/10)	79 % (=62/78)
Precision	88 % (=15/17)	79 % (=31/39)	89 % (=8/9)	95 % (=62/65)

表 4.6: 山一 野村 証券 '97 評価結果

	意見	解説	まとめ	サマリ
Recall	85 % (=11/13)	82 % (=41/50)	59 % (=10/17)	83 % (=66/80)
Precision	85 % (=11/13)	82 % (=41/50)	77 % (=10/13)	87 % (=66/76)

4.1.3 考察

意見セクション

間違いの多くは、解説セクションを意見セクションと判断したものだった。このようなセクションの多くは、人間による判断においても、意見と解説のばらつきが生じる可能性が高いと思われるセクションである。例えば以下のセクションはばらつきが生じる可能性が高いと思われる。

⋮

20日は李登輝・台湾総統就任式があり、これに合わせて実施すれば台湾威嚇になりかねない。「『自衛のために核兵器を保有している』との中国の主張にそごをきたす可能性が高いうえ、強烈な国際的非難を浴びる」(外務省幹部)ことを覚悟しなければならない。また、22日までは江主席がアフリカ歴訪中であるためだ。

⋮

他には、報道、用語解説といったその他セクションを意見セクションと判断した間違いがある。この間違いは、意見文の文末表現リストに、意見文の文末表現として適切かどうか今後さらに検討が必要と思われる表現が含まれていたために起こった。例えば以下のセクションは用語解説だが、文末表現「考えられる」が含まれているために意見セクションとされた。

簿外債務

有価証券報告書の財務諸表に記載されていない債務。子会社の借金に対する債務保証や、含み損が生じている非上場有価証券などを明らかにしていないケースが考えられる。また、“飛ばし”を繰り返して時価をはるかに上回る投資商品を「買い戻し条件」付きで、ある企業に保有してもらっている場合も、簿外に巨額の債務を抱えることになる。

⋮

この間違いに対応するには、より多くのルール作成データを用いて、意見文の文末表現リストを適切にチューニング(追加, 削除)する必要がある。

見落とされた意見セクションのほとんどは、以下のような原因により意見文の割合が低くなり、解説セクションと判断されたものだった。

1. 文末表現リストにない表現が使われている意見文が多い

2. 全体として意見を述べているのだが、最終的に意見を述べるまでの前置きが長い

1 の見落としとしては、ルール作成用データをもとに文末表現のリストを充実させることである程度対応できると考えられる。表 4.7 にリストに無かった意見を表す文末表現の例を挙げる。

表 4.7: リストにない意見文文末表現の例

思いに駆られる	感服した	油断はできない
心を砕いている	ひどい話だ	問題点も多い
痛感する	気がする	悲しい

また、数は少ないが、体言止めで意見を述べる場合もある。

そのためには、村山富市首相も国会で合憲と明言した破防法をオウム教団に適用するのは当然。

米国という一極だけでなく、次世紀には多くの極が必要。

他には、文末表現リストにない表現が使われている意見文しかなく、またセクション全体の文数も少ないため、その他ラベルを付与された意見セクションがある。以下にその例を挙げる。

安どの気持ち——青島幸男東京都知事の話

一応の決着がついたことに世間のみなさん同様、私も安どの 気持ちを持った。
一刻も早く国民のみなさんに不安や焦燥を抱かせる事態が収拾することを
心から祈っている。(犯罪の) 事実がかなり明らかになっており、年内には決
着するのではと 期待している。

この例において意見を述べている文末表現は「気持ちを持った」「祈っている」「期待している」だが、いずれもリストには無かった。また、3 文しか無いので、8 番目の解説 (残り) フィルタにもかからず、9 番目のフィルタによってその他ラベルが付与された。

解説セクション

間違って集めたセクションには以下のものがある。

1. 意見セクションだが、意見 (文末) フィルタにかからなかったもの
2. 箇条書によるまとめだが、まとめ (箇条書) フィルタにかからなかったもの
3. 報道だが、その他 (報道) フィルタにかからなかったもの
4. 文章によるまとめ
5. その他 (エピソード, 一覧表, 用語解説)

1, 2, 3 の間違いに対応するには、それぞれ、意見 (文末) フィルタ, まとめ (箇条書) フィルタ, その他 (報道) フィルタの性能を上げる必要がある。4, 5 の間違いは、そのセクションに特徴的な手がかりを把握する必要があるが、現在に対応できていない。

見落としとしては以下のものがある。

1. 解説だが、その他 (報道) フィルタにかかってしまったもの
2. 解説だが、意見 (文末) フィルタにかかってしまったもの

1 の見落としは、その他 (報道) フィルタが解説セクションを間違って報道と認識しているのが原因だが、この間違って認識した際に適用された報道の特徴は、” 記事の先頭 1 文以内に「～? 日～(た | る)。」を含む” が多かった。しかしこのルールがないと、核実験'96 のデータでは、報道をあまり除去できなくなる。

現在は 7 番目までのフィルタによってラベルが付与されずに残ったセクションから解説セクションを特定しているが、解説セクション独自の特徴を把握することが望まれる。解説セクションの特徴として、背景や因果関係を表すと考えられる「～からである」や「～わけである」、推量を表すと考えられる「～だろう」や「～ようだ」などの表現が有効な手がかりとして使える可能性がある。

まとめセクション

間違いとしては以下のものがある。

1. 一日の流れをエピソード的に記述しているセクション (先頭に時刻, 次にその時の内容が述べられている, という形式の箇条書)

2. 一覧表

以下に 1 の例を挙げる.

1・30 山梨県上九一色村では機動隊のバスがオウム真理教施設の近くを巡回, 緊張が高まる.

3・00 警察庁が防衛庁に家宅捜索について「官庁間協力」を正式に要請. 防衛庁は陸上自衛隊化学学校の化学兵器専門の幹部 4 人を, 警視庁と山梨県警に派遣.

5・00 防衛庁が映像伝送装置を搭載したヘリコプターを 30 分以内に離陸できる待機態勢をとるよう陸自東部方面航空隊に指示.

5・30 警視庁 17 階に「警備連絡室」が設置される.

5・45 富沢暉陸幕長が防衛庁に出勤. 東京周辺の陸自化学防護部隊が待機態勢に入る.

6・00 東京都世田谷区の教団世田谷道場に, 警視庁の捜査員と機動隊員 20 人が到着. インタホン越しに「扉を開けなさい」.

⋮

「一日の流れ」を形式的な手がかりだけで除去するのは難しいが, 箇条書の先頭にある数字の表しているものが日付なのか時刻なのかの区別をつけることが出来れば除去可能である.

また, 一覧表を除去するには, 一覧表と箇条書を区別する情報を探さなくてはならない. 箇条書によるまとめでは, 先頭に日付がきて, 次にその日の出来事を説明する文がある. 一覧表にはそのような説明文はほとんどない. 説明しているような文があるかどうかを判定できれば箇条書と一覧表を区別できる可能性がある.

見落としとしては以下のものがある.

1. 箇条書によるまとめだが, 先頭に日付がある行が閾値 (現在は 3 行) 以上連続していないためにフィルタにかからなかったセクション
2. 日付が先頭以外の位置にあるためフィルタにかからなかったセクション
3. 文章によるまとめ

以下に 1 の例を挙げる.

- ：
- 10月 損失補てんは独禁法違反にあたるとして、公正取引委員会
が野村に排除勧告、他の大手3社に警告
 - 92年 1月 損失補てん、一任勘定取引などを禁止する改正証券取引法
が施行
 - 95年 11月 米国野村証券が自己資本規制ルールに違反したとして、
ニューヨーク証券取引所が100万ドルの罰金命令
 - 96年 9月 総会屋親族企業「小甚ビル」への利益提供が発覚
 - 97年 3月 野村が、小甚ビルへの利益供与についての社内調査結果を
発表
- ：

以下に2の例を挙げる。

【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79
00万円を付け替え

【日興証券】93年10月に購入した東京電力株の値下がりです約8000万円
の損失を出し、その後1000万円を付け替え

【大和証券】93年12月に購入したJR東日本株の値下がりです3億4000
万円の損失を出し、数千万円を付け替え

【野村証券】95年3月、一任勘定取引の損失の穴埋めとして、現金3億200
0万円を供与し、約5000万円を付け替え

このような、見落としした箇条書によるまとめに対応するには、箇条書かどうかの判定を
もっと柔軟に行う必要がある。例えば、以下のような条件を設定すれば良いと考えられる。

1. 日付を表す数字を含む行が連続していなくても、また、日付を表す数字が行頭になく
ても、局所的に頻出していて、

2. それらの日付が昇順（古い日付から新しい日付という順）で現われるセクション

文章によるまとめには、以下のようなセクションを集めることで対応できる可能性が
ある。

1. 1番目のフィルタから7番目のその他（報道）フィルタまでの間、どのラベルも付与
されなかったセクションで、

2. 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

4.2 サマリパッセージ抽出システム

サマリパッセージ抽出システムの評価では、サマリを含むセクション検出システムの評価のような定量的な評価ではなく、システムの処理結果を観察し、それに基づき考察した。

4.2.1 文章によるまとめ、意見、解説セクション

これらのセクションからのサマリパッセージ抽出では、ある文の文末表現を調べて、その文が過去の出来事を記述しているか文 (サマリパッセージ) どうかを判断するが、以下のようなサマリパッセージは抽出することができなかった。

1. 体言止めで過去を記述しているもの
2. 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もの

以下に 1 の例を挙げる。

- 救急車出動の理由について、警視庁は「栄養失調の複数の信者が倒れたらしい」と説明。
- この時の参加者は、たった 4 人。
- 野村証券の業務管理部は、9 1 年の証券不祥事の反省で新設されたが、実際は不正防止どころか隠ぺいに関与。

前述したように、体言止めで意見を述べる場合もあるので、体言止めの文を全て抽出するという単純な対処法ではうまくいかない。

以下に 2 の例を挙げる。

- 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。
- 村山首相は横浜異臭事件の後で「サリンでなくてよかった」と漏らしてひんしゅくを買ったが、私たちが「危機管理」に対して感じている不安は、政府や行政当局者の座標軸が国民の側になくことに由来することを知るべきだ。

これらに対しては係り受け解析や構文解析により対応できる可能性がある。

4.2.2 箇条書によるまとめ

まとめセクションの考察で挙げた以下の場合では失敗する.

1. 先頭に日付がある行が閾値以上連続していない箇条書によるまとめセクション
2. 日付が先頭以外の位置にある箇条書によるまとめセクション

これらの箇条書に対応するには, 前述したような柔軟な方法で箇条書を認識するようにしないといけない.

4.2.3 週間日誌

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.

第 5 章

サマリパッセージを用いた複数記事要約

本章では、検索結果の記事集合からサマリパッセージを抽出した後、そのサマリパッセージを用いてどのように記事集合を要約するのかについて述べる。また、従来の要約手法について述べ、本章で提案する手法と比較する。

5.1 要約アルゴリズムの概要

以下に、サマリパッセージを利用した複数記事要約アルゴリズムの概要を述べる。

1. 検索結果の記事集合からサマリパッセージを抽出する (図 1.1).
2. 抽出されたサマリパッセージを、それぞれ、記事集合中の過去の記事の最も関連の強い個所と対応づける。
3. 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。

1 は本研究の目的であり、詳細はすでに述べた。1 の段階を終えると、意見セクション、解説セクション、まとめセクションからサマリパッセージが抽出されている。

2 の段階では、抽出されたサマリパッセージを、それぞれ、記事集合中の過去の記事の最も関連の強い個所と対応づける。方法としては、抽出されたサマリパッセージをそれぞれクエリとして、過去の記事集合に対して検索を行い、内容的に最も関連の強い個所を同定する。サマリパッセージをクエリとする際は、例えば、形態素解析して、その中から名詞だけを抜き出し、それらをクエリにする。検索の方法としては [8] が有効であると考えられる。[8] では、一文書全体ではなく、クエリに強く関連する文書中の一部分を検索すること

が可能である。図 5.1に、抽出されたサマリパッセージを過去の記事の最も関連の強い個所と対応づける様子を示す。

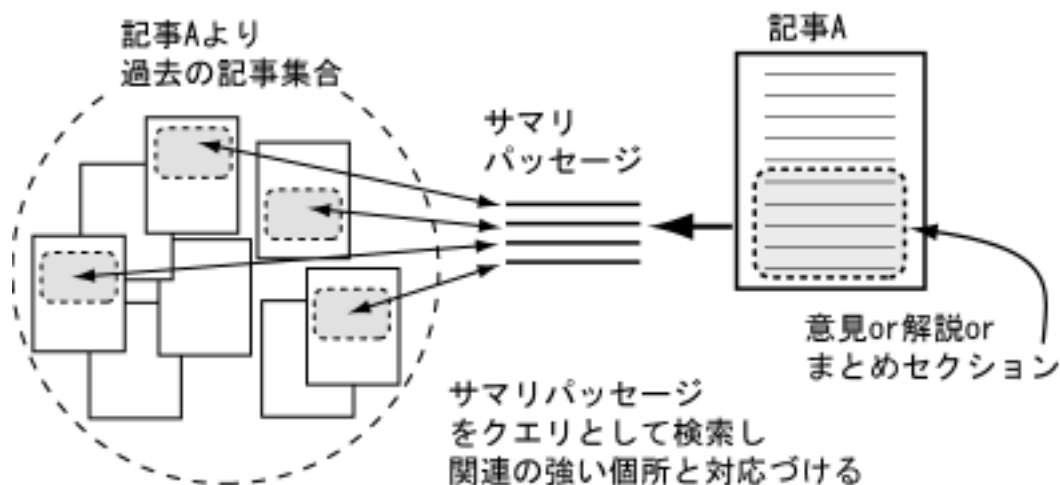


図 5.1: サマリパッセージと過去の記事の対応付け

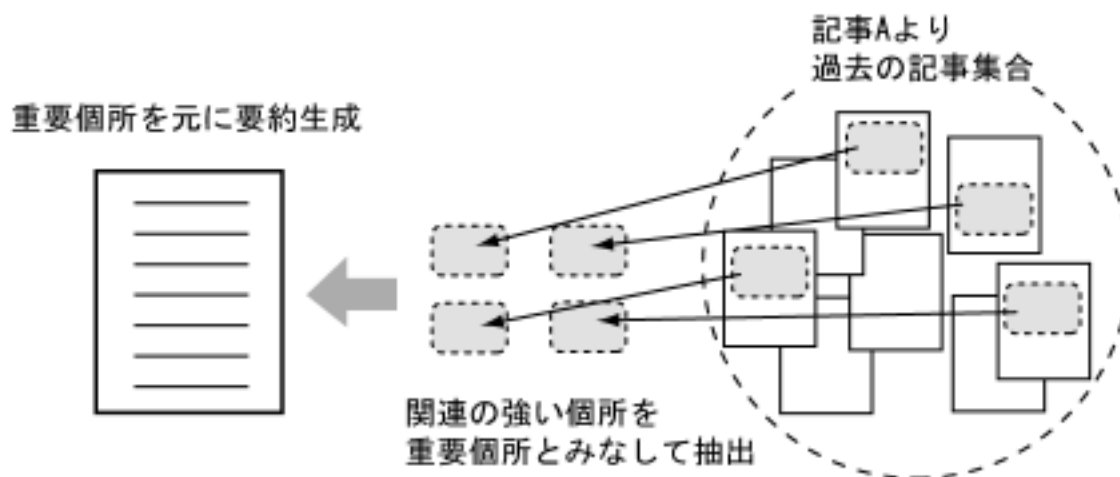


図 5.2: 要約生成

3では、2の段階で同定したサマリパッセージと関連の強い個所を、複数記事中の重要個所と見なして要約を生成する。単純には、同定された重要個所を記事の出版日時順に並べるという方法が考えられる。この処理の際に重要なのは、冗長的な個所を検出し、除外することである。また、同定された個所を単純に並べただけでは、文章として不自然になること

も考えられる。従って、要約の生成の段階では、並べられた重要個所を文章として自然になるように書き換える処理が必要である [10]。図 5.2 に、抽出された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、要約を生成する様子を示す。

以上で述べた複数記事要約アルゴリズムにおいて、サマリパッセージは、複数記事要約の、言わば「種」として機能する。

5.2 関連研究

複数の新聞記事を要約する際には、どの記事のどの内容が重要かを同定する必要がある。その点に関して従来研究を概観し、本研究と比較する。

[1] は、記事集合をあらかじめクラスタリングし、各クラスタ毎に要約を生成する手法を提案している。この手法では、クラスタ中の記事の各文の重要度を計算し、要約率に応じて記事集合から重要度の高い数文を抽出し、それらの文の記事の出版日順に並べて出力する。文の重要性は、クラスタ centroid(クラスタの特徴を表す語の集合) 中の語を文が含む割合と、文の位置 (リードにあるかどうか) を基に判断している。

[5] は、検索結果をクラスタリングし、クラスタ毎に、そのクラスタの話題を表す語とそれらを含む文を抽出する。各クラスタから抽出された文は、焦点-主題連鎖を考慮して並べ換えられ、そのクラスタの要約として出力される。文の重要性は、文に含まれる単語の評価値から決まる。単語の評価値は、クラスタの話題を示す尖度、歪度という統計的な情報から決まる。

[1][5] の手法では、ともに統計的な情報をもとにクラスタ全体の話題を表す文を抽出し、それを重要な内容としている。統計的情報をもとに重要な内容を同定する手法はドメインに依存せず適用できる手法といえるが、語の出現頻度などの統計的な基準では記事集合中の重要な内容を常に正確に反映しているとはいえない。また、統計的な情報を得るには計算的にコストがかかる。

[3] は、複数記事中の類似する情報を統合することで要約を生成している。この手法では、まず、記事をパラグラフに分割し、類似するパラグラフをクラスタリングする。次に、各クラスタ中に頻出するフレーズを抽出し、類似するフレーズ同士を、述語項構造を調べることで統合し、その後、統合されたフレーズをもとに文生成器により要約を出力する。この場合、繰り返し出現する、クラスタ間で共通しているフレーズを重要と考えている。

[12] は、複数の情報源から得られる新聞記事の共通部分を融合することで要約を生成する手法を提案している。この手法では、まず、共通部分を有する記事の候補として、何らかの関連がある記事を検出する。次に、検出した記事を共通部分と差分に分け、利用目的に応

じて、共通部分と差分の両方を用いたり、共通部分のみを用いたりしながら、それらを文章へと再構成する。この手法では、関連する記事の間で共通する個所を重要と考えている。

[3][12] の手法では、複数記事間で共通して現われる個所が重要な内容と考えているが、この手法も統計的な手法と同様、ドメインに依存しない点が有利だが、「重要な内容は繰り返し現われる」という前提は常に正しいとは限らない。

[7] は、記事の第一段落を用いて、その重複個所、冗長個所を削除することにより、複数の関連記事を要約する手法を提案している。重複個所、冗長個所を削除する際は、推量文は重要でないとして削除したり、続報記事の導入部の特徴的な表現「～したが」「～事件で」「～について」などを重複個所として削除するなどのヒューリスティックスを用いている。この手法では、一般に新聞記事のリードにその記事の概要が述べられていることから、リードの部分が重要な個所と考えている。

複数記事をリードを用いて要約する手法は計算コストも低く、新聞記事の性質を反映しており有効と考えられるが、複数記事中のリードを全て同等に重要と見なすのは、重要性の基準として粗い。なぜなら、ある話題に関する複数記事の中には、全体においてはそれほど重要でない内容も含まれているにもかかわらず、全ての内容のリードを重要と見なしてしまうからである。

[2] は、テロリストの記事だけに限定して、情報抽出の手法を用いた要約手法を提案している。まず、情報抽出手法を用いて、記事集合から犯人、犠牲者、事件のタイプなどの情報を抽出し、テンプレートにそれらの情報を埋める。次に、そのテンプレートを用いて、新しい情報を優先させたり、不完全な情報同士を組み合わせでより完全な情報にしたりしながら要約を生成する。

ドメインの知識をテンプレートとして予め与えており、重要な内容を正確に同定できると考えられるが、ドメインに強く依存していて汎用性に欠ける。また、テンプレートを作成するコストが高いと思われる。

本研究の提案する手法では、意見、解説、まとめセクションにある、新聞記者の視点に基づいた過去の出来事の記述（サマリパッセージ）を抽出し、その情報に基づいて重要個所を同定する。記者はそれまでの話題の流れを熟知しており、どの出来事が重要であるかを考えて、過去の主要な出来事を要約し、意見、解説、まとめセクションに盛り込んでいる、と考えられる。従って、サマリパッセージを用いた重要個所同定は、対象の話題を熟知している記者の知識を利用した方法といえる。

サマリパッセージの情報は、それまでの経緯の中で何が重要かに関する記者の判断を直接的に反映しており、統計的な情報や出現位置の情報よりも正確な重要個所同定を可能にする。また、計算コストも低く、新聞記事であればどのメインにも適用できると考えられる。

第 6 章

まとめ

本研究では、意見、解説、まとめセクション（サマリを含むセクション）中にあるサマリパッセージが、複数記事要約処理で必要になる重要個所同定において有効な情報であることを論じ、検索結果の記事集合からのサマリパッセージの抽出方法と、そのシステムの評価結果、サマリパッセージを用いて記事集合の要約を生成するまでの処理について示した。

記事集合からのサマリパッセージ抽出処理では、いきなり記事集合からサマリパッセージを抽出するのではなく、まず、記事集合から意見、解説、まとめが述べられている個所を検出する。この際、意見、解説、まとめを記事単位ではなく、セクション単位で検出する。これらのセクションの検出には、各セクションの特徴を反映したヒューリスティックスを用いる。その後、検出されたサマリを含むセクションからサマリパッセージを抽出する。サマリパッセージの抽出方法はセクションの形式によって異なる。

システムの評価は、サマリを含むセクション検出システムと、各セクションからのサマリパッセージ抽出システムを分けて行なった。また、サマリを含むセクション検出システムの評価では、セクション分割処理は人手で行ない、(意見、解説、まとめのカテゴリを表す)ラベル付与処理の評価のみを行なった。その結果、ラベル付与モジュールが高い精度でカテゴリ判定することを定量的に示した。また、サマリパッセージ抽出システムは、おおむね正確に各セクションからサマリパッセージを抽出することを定性的に示した。

最後に、抽出されたサマリパッセージを用いて記事集合を要約する方法について論じた。また、この手法と従来の複数記事要約の手法との違いを述べた。

第 7 章

今後の課題

7 章では本研究における今後の課題を論じる。今後の課題として以下の点を挙げる。

1. 記事集合からのサマリパッセージ抽出の精度向上
2. セクション分割アルゴリズムの精緻化と評価
3. サマリパッセージを用いた要約処理の詳細の検討
4. より客観性の高いデータ, ルールの作成

7.1 記事集合からのサマリパッセージ抽出の精度向上

4 章において, 記事集合からのサマリパッセージの抽出がおおむねうまくいくことを示したが, さらなる精度向上が必要である。

サマリを含むセクション検出システムでは, 以下のセクションを認識することができず, システムの精度を低下させた。

- 解説セクション
- 文章によるまとめセクション
- 以下のその他セクション
 - － エピソード
 - － 用語解説
 - － 一覧表

精度向上には、これらのセクションを認識することが必要である。解説セクション、文章によるまとめセクションの検出の可能性については4章で述べた。今後は上に挙げたその他セクションの特徴の把握するため、まず、エピソード、用語解説、一覧表セクションを大量に収集する必要がある。

また、今回はシステムを構成する各フィルタの閾値を手で決定したが、大量のデータを用いた機械学習により閾値を決めることで、さらなる精度向上が期待できる。

また、今回は3つのデータを用いてルール作成、評価を行なったが、さらにデータを追加してルールの検討、評価を行なう必要がある。

検出された各セクションからのサマリパッセージ抽出システムの開発に当たっては、十分な量のルール作成用データと評価用データを用意できなかった。今後はまず、十分な量のデータを作成する必要がある。

7.2 セクション分割アルゴリズムの精緻化と評価

2章でセクション分割アルゴリズムについて述べたが、これについても十分な量の評価用データを用意できなかった。そのため、精度的にまだ不十分なこともあり、今回は評価を行なわなかった。

現在のルールでは、分割してほしい所はほぼ全て分割してくれるが、分割しないでほしい所でも過剰に分割してしまう傾向がある。以下にその例を挙げる。

⋮

オウム真理教による地下鉄サリン事件などに加わったとされる「法皇内庁」トップ、中川智正被告（32）の初公判が二十四日、東京地裁（池田修裁判長）で開かれた。教団代表の麻原彰晃被告（40）も起訴された一連の事件で、核

⋮

サリンを発散させた経緯を詳述した。中川被告が両事件への関与を大筋で認めたことで、二十六日に初公判が予定される麻原被告は苦しい立場に立たされた。（3面に冒頭陳述の要旨、社会面に関連記事）

サリン発散、共謀は否定

起訴事実の認否で中川被告は地下鉄事件について「サリン発散計画を尊師と話したことはない」と麻原被告との直接の共謀は否定したが、教団「科学技術省」大臣の村井秀夫幹部（故人）からの指示でサリン生成には加わったこと

⋮

本来は上の例全体で1つの報道セクションなのだが、セクション分割モジュールは「サリン発散、共謀は否定」の行もセクション区切りと見なしてしまう。今後は十分な量のルール作成データを用いて、セクション区切り行の特徴を掴み、より知的な方法でセクション分割をする必要がある。また、評価用データを作成し、定量的な評価を行わなければならない。

7.3 サマリパッセージを用いた要約処理の詳細の検討

5章で、サマリパッセージを用いた複数記事要約アルゴリズムの概要を述べた。しかし、このアルゴリズムはまだ大雑把なものであり、実装レベルの詳細までには至っていない。5章では1) サマリパッセージ抽出した後2) サマリパッセージをクエリとして過去の記事集合をパッセージ検索 [8] し、最も関連の強い個所と対応づけ3) 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、と述べた。

2) の段階では以下の点を検討する必要がある。

i. 検索する際、クエリとして、サマリパッセージのどういう情報を用いるか

ii. 検索結果の中からどうやって関連の強い個所を絞り込むか

i. に関して、サマリパッセージ全体をクエリとするのか、それとも、サマリパッセージを形態素解析して名詞のみをクエリとするのか、などの選択肢が考えられる。これは実際に試しながら検討すべきである。そのためには、正確に抽出されたサマリパッセージと、それに対応する過去の記事集合中の関連の強い個所を手で決定してあるデータを作成する必要がある。ii. に関して、サマリパッセージを用いて過去の記事集合を検索した場合、検索結果は当然複数になるので、それらの中から最も関連の強い個所を絞り込む方法を検討する必要がある。単純には、検索結果のスコアが最も高い個所を最も関連の強い個所と見なす方法が考えられるが、それでうまくいくのかは、i. と同様、実際に試して検討すべきである。

3) の段階では以下の点を検討する必要がある。

a. 重要個所のどういう情報を要約に用いるか

b. 文章として自然な要約を生成する際に問題となる点は何か

a. に関して、最も関連の強い個所を重要個所と見なすのだが、その重要個所を全て要約に用いるのか、それとも、その中からさらに情報を絞り込むのかを検討する必要がある。少なくとも、5章で述べたように、重要個所間で冗長な個所があれば、それを検出し削除する

必要がある。これに関しても、実際にどういう個所が最も重要な個所として検索されるのかを見ながら検討すべきである。b. に関して、最終的に要約として出力される文章を自然なものにするためには、どういうことが問題となり、それに対してどういう解決法があるのかを検討しなくてはならない。[9] では抄録の読みやすさの向上について検討している。本研究で提案する要約手法により出力される要約は複数記事の抄録といえるものであり[9] をもとに、複数記事の抄録における読みやすさの問題とそれに対する解決法を検討すれば良いと考えられる。

7.4 より客観性の高いデータ, ルールの作成

本研究で用いたルール作成用データと評価用データは一人によって作成されたデータであり、客観性に問題がある。今後は、システムのルールの洗練、評価の客観性の向上のためにも、客観性の高いルール作成用データ、評価用データを作成する必要がある。そのためには、複数の被験者を募り、ある新聞記事集合の各記事に対する意見、解説、まとめ、その他のカテゴリ判定の実験を行なう必要がある。

謝辞

本研究を進めるにあたり、終始熱心な御指導を賜りました島津明教授、奥村学助教授に心から感謝致します。

中間審査などの折には、諸先生方から貴重な御意見を頂きました。感謝致します。

新聞記事データの利用を承諾して下さった毎日新聞社の方々に感謝致します。

さらに、貴重な御意見、討論をしていただいた自然言語処理学講座 望月源 助手ならびに研究室の皆様、明星大学 大石亨助教授に感謝致します。

最後に、筆者を支えてくれた両親と兄に厚く御礼申し上げます。

参考文献

- [1] D.R.Radev, H.Jing, and M.Budzikowska. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. Proc. of the ANLP/NAACL2000 Workshop on Automatic Summarization, pp.21-30, 2000.
- [2] K.McKeown and D.R.Radev. Generating summaries of multiple news articles. Proc. of the 18th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, pp.74-81, 1995.
- [3] K.McKeown, J.L.Klavans, V.Hatzivassiloglou, R.Barzilay, and E.Eskin. Towards multidocument summarization by reformulation: Progress and prospect. Proc. of the 14th National Conference on Artificial Intelligence, pp.453-460, 1999.
- [4] Min-Yen Kan, Judith L.Klavans and Kathleen R.McKeown. Linear Segmentation and Segment Significance. Proc. of the Sixth Workshop on Very Large Corpora, pp.197-205, 1998.
- [5] T.Fukuhara, H.Takeda, and T.Nishida. Multiple-text Summarization for Collective Knowledge Formation. Workshop on Social Aspects of Knowledge and Memory, IEEE Systems, Man and Cybernetics Conference, 1999.
- [6] 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報告 85-11, 1991.
- [7] 船阪貴浩, 山本和英, 増山繁. 冗長度削減による関連新聞記事の要約. 情報処理学会自然言語処理研究会報告, pp.39-46, 1996.
- [8] 望月源, 岩山真, 奥村学. 語彙的連鎖に基づくパッセージ検索. 自然言語処理, Vol.6, No.3, pp.101-126, 1999.

- [9] 難波英嗣, 奥村学. 書き換えによる抄録の読みやすさの向上. 情報処理学会自然言語処理研究会報告, pp.53-60, 1999.
- [10] 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.
- [11] 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, 1998.
- [12] 柴田昇吾, 上田隆也, 池田裕治. 複数文章の融合. 情報処理学会自然言語処理研究会, pp.77-82, 1997.