複数記事に対する要約や情報抽出に関する一考察

野畑 周
nova@crl.go.jp
独立行政法人 通信総合研究所
けいはんな情報通信融合研究センター

関根 聡
sekine@cs.nyu.edu
Computer Science Department
New York University

1 はじめに

近年、自動要約の研究においては、Document Understanding Conference (DUC)[1] や Text Summarization Challenge (TSC)[2] などの要約コンテストにも見られるように、単一文書の要約だけでなく複数文書の要約が主要なタスクの一つとして扱われるようになりつつある。複数の新聞記事を要約したり、それらから適切な情報を抽出するシステムを開発するためには、現実の複数の新聞記事がどのような軸でまとめられるかを、分類・整理する必要がある。本研究では、恣意性を避けるために、ランダムに一記事を選択しその類似記事として記事群を生成し、各々の記事群について、それらの分類を行った結果を報告する。合わせて、考えられる要約や情報抽出の出力形式を考察する。

2 研究の背景

複数新聞記事を対象として要約・情報抽出を行うときには、対象とする記事群の性質によって望ましい手法・出力形式が変わると考えられる。例えば、ある特定の事件の経過について報道した一連の記事と、各記事がそれぞれ異なる交通事故について報道しているが全体としては交通事故についての報道という点で共通しているような記事群とでは、望ましい要約や情報抽出の出力は異なってくると考えられる。

McKeown らは、要約生成に先立って記事群の性質を判定するモジュール (Router) を設けて記事群を分類し、その後に各々の種類に応じた要約手法を適用する手法をとっている [3]。彼らは DUC2001 のトレーニングデータ 30 記事群を以下の 4 種類に分類している：

Single-Event (2 記事): 一つの出来事に関する記事群。記事間のタイムスパンが最も短い。例: ピナツボ火山の噴火についての記事

Person-centered (Biography) (10 記事): ある人物に関する記事群。例：グリーンスパン議長の経歴

Multi-Event (7 記事): 同種の複数の出来事に関する記事群。例：各地で観測された日食現象

Other (11 記事): それ以外の記事群。例：各国の南極開発政策、イラク・クウェート戦争の概要

Router モジュールが分類を行う際には、(1) 記事間のタイムスパン、(2) 記事セット中の同じ日に書かれた記事の割合、(3) 大文字で始まる語の頻度、(4)he, she などの人称代名詞の頻度、の 4 つの情報を用いている。

彼らの分類は、要約の対象となる記事群によく見られる性質を効率良く分類している。しかしながら、彼らの分類では Other に分類される記事群が多く、これらには他に適切な分類があるのではないかと考えられる。さらに彼らは、実際に要約を行うときには Multi-Event と Other の分類を同一視して同じ要約手法を適用しており、Router モジュールによる分類では Single-Event, Person-Centered, Other の 3 種類に分類することになる。

本研究では、McKeown らによる分類よりも網羅的な分類の定義を与え、それが現実の新聞記事群に対して適用可能なことを示すことを目的とする。

3 実験と結果

本節では、記事群の分類についての実験方法とその結果について述べる。

3.1 記事群の生成

我々は、恣意的に記事群を作成することをさけるために、京大コーパスに含まれる記事から無作為に

表 1: 記事群の主題とその分類

記事群	記事数	記事群の主題	分類
s01	14	ロシアのチェチェンとの紛争	Single-Event
s02	10	日本各地の地震速報	Multi-Event
s03	8	愛犬家連続失踪事件	Single-Event
s04	7	民主リベラル新党の結成	Single-Event
s05	6	書初め大会、競技大会等	Multi-Event
s06	6	各国の元首の動向	Multi-Person
s07	6	日米首脳会談	Single-Event
s08	4	大相撲の記事。特に貴ノ花について	Single-Person
s09	4	伊東議員の北海道知事選出馬	Single-Person
s10	4	インドネシア・ポルトガル外相会談	Single-Event
s11	3	浄土真宗本願寺派の記事	Single-Org
s12	3	市立船橋の高校サッカー大会での活躍	Single-Org
s13	3	ビールメーカー各首脳のコメント	Multi-Org
s14	3	動物が関与した事件の記事	Other
s15	3	レコード大賞、ノーベル賞に関する記事	Multi-Artifact
s16	3	中央アルプス・木曽駒ケ岳付近で発生した雪崩	Single-Event
s17	3	関西国際空港に関する記事	Single-Location
s18	3	サッカーのインタコンチネンタル選手権	Single-Event
s19	3	野茂投手の大リーグ挑戦	Single-Person

記事を一つ選び、その記事に類似した記事を検索することを繰り返して記事群を作成した。具体的な生成方法を以下に示す。

1. 京大コーパスに含まれる記事から無作為に記事を一つ選択する
2. 選択された記事 a_i からキーワード列を取り出す：
   キーワードは、Juman3.61[4] を用いた形態素解析結果から、時相名詞・副詞的名詞を除いた名詞で頻度 2 以上のものとした。
3. キーワード列を用いて、記事間の類似度を求める：
   各記事 a_j について同様にキーワード列を取り出し、キーワードどうしの類似度を Dice の係数 [5] を用いて求めた。すなわち、記事 a_i のキーワードの数を $w(a_i)$、記事 a_j のキーワードの数を $w(a_j)$、両記事に共通するキーワードの数を $w(a_i, a_j)$ として以下の式で定義される $D_{i,j}$ の値を求めた。

$$D_{i,j} = \frac{2 * w(a_i, a_j)}{w(a_i) + w(a_j)}$$

4. 類似した記事を取り出す：
   同一の記事以外 $(i \neq j)$ で、Dice の係数 $D_{i,j}$ の値が一定の値 C 以上となる記事を類似記事と見なして取り出した。ここでは $C = 0.5$ とした。

以上の記事群生成を 50 回繰り返してできた 50 記事群のうち、3 記事以上のものを選ぶと 26 記事群となった。それらの中からさらに記事群の内容が同じと考えられるものを除くと、19 記事群になった。表 1 に記事群のリストを示す。

3.2 記事群の分類

表 1 に示した各記事群の内容を、実際に読んで分類を行った。記事群に与える分類づけとしては、以下の 11 分類を用いた。

出来事:	Single-Event, Multi-Event
人名:	Single-Person, Multi-Person
組織名:	Single-Org, Multi-Org
固有物名:	Single-Artifact, Multi-Artifact
地名:	Single-Location, Multi-Location
その他:	Other

これらは、出来事 (Event) を除いては、IREX ワークショップ [6] で用いられた固有表現のクラスに基づいている。Single の場合は固有表現の具体的な値 (「貴ノ花」「関西国際空港」など)、Multi の場合は複数の固有表現、すなわち固有表現のクラス (「人名」や「地名」) に注目するという意図を示している。McKeown らの分類の一つである Person-centered(Biography) は上の定義中の Single-Person に相当するので、この分類は McKeown らの用いた分類を含んだより詳細なものになっている。これに基づいて 19 記事群を分類すると、表 1 の右端の列に示すように、一つを除いて全ての記事群について分類づけを行うことができた。分類の頻度は表 2 のようになった。McKeown

表 2: 記事群の分類

分類	数
Single-Event	7
Single-Person	3
Single-Org	2
Multi-Event	2
Single-Location	1
Multi-Person	1
Multi-Artifact	1
Multi-Org	1
Other	1

らが用いた分類 (Single-Event, Multi-Event, Single-Person) が上位に来ている。

Single の分類では、その記事群中での特定の固有表現の頻度や文書頻度 (Document Frequency) が大きくなり、Multi の分類では、特定の固有表現についての頻度・文書頻度は大きくないが、固有表現のクラス全体としての頻度・文書頻度が大きくなる傾向にある。

4 考察

本節では、各記事群に対して、分類に沿ってどのように要約・情報抽出を行うかを考察する。Single-Artifact, Multi-Location は今回の記事群では当てはまるものはなかったので、ここではふれない。

Single-Person, Single-Org

Single-Person, Single-Org では、記事群 s08 の「貴ノ花」や s09 の「伊東秀子」、s12 の「市立船橋」のように、その記事群中での特定の人物・組織を指す固有名詞の頻度や文書頻度が大きい。記事中の主要な記述には、その特定の人物・組織を E とすれば、(1) E が主体となって行った活動、(2) E を対象として他者が行った活動、の 2 つがある。従って、E を軸として、主に活動・発言を示す動詞の変化に注目して要約・情報抽出を行えばよいと考えられる。例として、記事群 s09 から抽出した情報の一部を表 3 に示す。

Multi-Event, Multi-Artifact

一方、Multi-Event では行動の内容が限定される。従って、行動を示す動詞ではなく主体・対象を示す表現の変化に注目してまとめるのが良いと考えられる。例えば、記事群 s05 中の表現では、「出来事が開かれる・開幕する」という表現がよく現れる。また、出来事の種類がさらに限定された場合、その出来事特有の情報が記述されるので、それらを捉えることも重要と考えられる。例えば地震に関する記事群 s02 では、地震が起った時刻・場所に加えてその震度・被害などが記述されることが多く、それらの情報は記事群に対して要約・情報抽出を行うときには注目すべきであろう。つまり、類似度の高い表現が複数ある場合には、それらに共通する要素を重要だと見なし、それらを軸として個々の表現を列挙する必要がある。

Multi-Artifact には、今回の「賞」に関する記事群以外に、例えば「新製品の発売」について述べた記事の集合も当てはまる。この場合も、販売元や製品の金額など、記事群に特有な情報を認識する必要があると考えられる。

Multi-Person, Multi-Org

Multi-Person, Multi-Org では要約・情報抽出の軸になる特定の表現が明確でないが、クラスとしては人名・組織名に固定することができる。従って、要約・情報抽出を行う際には、個々の人物・組織 ($E_1, E_2, \ldots$) を対象として、(1) $E_1, E_2, \ldots$ が主体となって行った活動、(2) $E_1, E_2, \ldots$ を対象として他者が行った活動、を主に記述することになると考えられる。

Single-Location, Single-Event

Single-Location や Single-Event には、特定の地名や出来事によって、Multi-Person, Multi-Org にさらに新たな軸を付加する場合もあるが、表 4 に示す記事群 s17 の内容のように、日時の他に抽出すべき軸を持たない場合もある。従って、分類の範囲としては Multi-Person, Multi-Org より広いと考えられる。同じ出来事を示す表現が複数現われうるので、類似度の高い表現が複数ある場合には、日時の情報などを利用して、それらのうち代表的な表現一つを出力する必要がある。

日付情報

19 記事群に含まれる 140 記事において、IREX で定義された各固有表現クラスの文書頻度を表 5 に示す。日付情報はほぼ全ての記事に出現することが分かる。また、各新聞記事には<DOCNO>など、記事が掲載された日付が分かる情報が添えられている。従っ

表 3: Single-Person の例

主体	対象	日時	行動
伊東秀子	-	1月6日	今春の北海道知事選に出馬する意向を固めた
〃	-	1月6日午後	離党届を提出
社会党道本部	伊東秀子	1月6日	除名を決議した
さきがけ北海道	〃	1月7日	推薦する方針を固めた

表 4: Single-Location の例

場所	日付	活動
関西国際空港	三十一日夜から一日朝	イベントや初日の出を楽しむたくさんの若者や家族連れが繰り出した
関西国際空港	5日	展望デッキの入場者が、開港4カ月で１００万人に達した
関西国際空港	年末年始	外国に出かけた人は十七万九千七百人

表 5: 19 記事群中の各固有表現クラスの文書頻度

クラス	文書頻度	クラス	文書頻度
組織名	107	**日付表現**	**135**
人名	103	時間表現	53
地名	122	金額表現	10
固有物名	17	割合表現	9

て、日付の情報は特別であり、日付を軸にして要約・情報抽出を行うことは、デフォルトとして全ての記事群について可能であるといえる。日付・時間のみが共通するような記事群を作成することも可能ではあるが、そのような記事群は、全く互いに独立の記事群から成るか、他の視点からの分類づけが可能であるかのどちらかになる。全く互いに独立の記事群からの要約・情報抽出を行うことは現実的ではないので、除外してよいと考えられる。

5 まとめと今後の課題

複数文書の要約を行うには、対象とする文書群の性質によって、望ましい手法や出力形式を選択すべきだと考えられる。実際に要約システムの中には、文書群の性質を判定し、その後に各々の種類に応じた要約手法を適用するシステムも提案されている。

本稿では、複数文書要約のために、新聞記事群の性質を示す分類の定義を与え、それが現実の新聞記事群に対して適用可能なことを示した。具体的には、以下の項目の内容を示した。

- 分類の定義: 固有名詞とそのクラスに基いた分類 (single, multi) 11 分類を提案
- 記事群の作成：無作為に一記事を選択し、それに類似した記事を取り出すことで恣意性の少ない新聞記事群を作成
- 記事群の分類：作成された記事群の性質を分類し、提案した分類の定義が適切であることを提示

今後は、実際に記事群の性質を自動的に分類するシステムを作成し、さらにその分類結果に応じて適切な要約を出力するシステムを構築していく予定である。また、分類の定義についても、今回提案した中で実際には現われなかった分類の性質を調べるために、より大規模な記事セットの集合について実験するなどして、さらに検討していきたいと考えている。

参考文献

[1] DUC. http://www-nlpir.nist.gov/projects/duc/, 2001. Document Understanding Conference.

[2] TSC. http://oku-gw.pi.titech.ac.jp/tsc/, 2001. Text Summarization Challenge.

[3] K. R. McKeown, R. Barzilay, D. Evans, V. Hatzivassilogou, M. Yen Kan, B. Schiffman, and S. Teufel. Columbia Multi-Document Summarization: Approach and Evaluation. In *Online Proceedings of DUC2001*, 2001.

[4] 黒橋禎夫, 長尾真. 日本語形態素解析システム JUMAN version 3.61. 京都大学, 1999.

[5] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill, 1983.

[6] IREX. http://cs.nyu.edu/cs/projects/proteus/irex, 1999. Information Retrieval and Extraction Exercise.