複数記事要約のための
固有表現を用いた記事セットの自動分類

野畑 周 † 関根 聡 ‡ 井佐原 均 †
nova@crl.go.jp sekine@cs.nyu.edu isahara@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
‡Computer Science Department, New York University

1 はじめに

複数記事要約は、最近の自動要約の研究において主流になりつつあり、米国の評価型ワークショップ Document Understanding Conference (DUC)[2] や日本における Text Summarization Challenge (TSC)[4] においても、自動要約システムの評価を行う主な課題になっている。複数の記事から成る記事セットを一つの文章に要約する際には、その記事セットの主題を考慮する必要がある。例えば特定個人の行動や発言内容、また各地で起った地震の被害状況など、記事セットによってその主題が異なる。複数記事の自動要約をより適切に行うためには、それらの主題を判定しその主題に応じて形式を選択して要約することが必要であると考えられる。本研究では、固有表現に基いて定義した分類によって記事セットの主題を表し、無作為に作成した 100 個の記事セットについて、三人の被験者による分類付与の評価結果を示す。さらに、機械学習に基いた自動分類実験を行い、各分類についてその精度の評価と結果の分析を示す。

2 研究の背景

複数記事要約の観点から記事セットを分類する先行研究には、コロンビア大学の McKeown らによるものがある [3]。彼らは、DUC2001 のトレーニングデータに含まれる 30 記事セットについて、記事セットを single-event, person-centered, multi-event に other を加えた 4 種類に分類している。彼らの分類は、要約の対象となる記事セットによく見られる性質を効率良く分類しているが、other に分類される記事セットが 11 個と全体の 1/3 以上を占めており、分類の定義をより詳細にすることが可能であることが示唆されている。我々は、記事セットのより詳細な分類について固有表現抽出に基づく分類の定義を提案し、DUC の記事セットだけでなく日本語新聞記事から無作為に作成した 30 記事セットに対して提案した定義に基づく分類を行い、その分類の定義が有効性をもつことを示した [5, 9]。本研究では、より規模の大きなデータに対する実験として、新しく作成した 100 個の記事セットについて提案した定義に基く分類を行い、そのデータに対する三人の被験者によって付与された分類結果について、互いに比較した結果を示す。さらに、決定木を用いた自動分類実験を行い、用いた素性の種類と分類結果との比較結果を示す。

3 分類の定義

本稿で用いる分類の定義は、分類実験とは別に作成した記事セットを分析した結果作成したものである [9]。我々はそれらの記事セットを分析し、固有表現に基づく 13 種類の分類を定義した。ここで対象とする固有表現のクラスは、MUC[1] や IREX[8] において定義されたものをもとに拡張したものであり、人名 (person)・組織名 (organization)・地名 (location)・施設名 (facility)・固有物名 (product)・イベント名 (event) の 6 つのクラスを用いている。single-(class) の分類は、そのクラス内の特定の表現が記事セットの中心になっていることを示す。例えば、ある特定の人物が記事セットの中心になっている場合は single-person に分類する。multi-(class) の分類は、そのクラス内に突出する表現があるわけではないが、クラスとして記事セットの中心になっていることを示す。例えば、個々の記事が国内各地で別々に起こった地震の規模について述べているような記事セットは multi-event に分類する。

表 1: 100 記事セットの分類の正解データ

Single-Person	12	Multi-Person	2
Single-Org	8	Multi-Org	8
Single-Location	7	Multi-Location	2
Single-Facility	1	Multi-Facility	0
Single-Product	23	Multi-Product	3
Single-Event	43	Multi-Event	27
Other	0		

3.1 記事セットの生成

記事セットの偏りを避けるため、日本語新聞記事コーパス (毎日新聞 1998 年版、1999 年版) から以下の手順で記事セットを作成した。

1. 新聞記事コーパスから無作為に記事を一つ選択:
2. 選択された記事からキーワード列を取り出す: キーワードは、Juman3.61[10] 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、記事間の類似度を計算: 各記事について同様にキーワード列を取り出し、キーワード同士の類似度を Dice の係数 [6] を用いて求めた。
4. 類似した記事を取り出す: 同一の記事以外で、Dice の係数の値が一定の値 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。
5. 繰り返す: 1.〜4. の記事セット生成を 300 回繰り返してできた 300 記事セットのうち、5 記事以上でかつ記事セットの内容が重ならないものを抜き出して 100 記事セットを作成した。

4 実験

本節では、前節の方法で作成された 100 記事セットについて記事分類の正解データを作成し、このデータに対して被験者による分類の評価と、いくつかの分類に対する自動分類実験の結果を示す。100 記事セットに対する記事セットの分類の結果を表 1 に示す。この記事セットの分類は、次節に示す 3 人の被験者の結果を基に著者の一人が検討し、最終分類結果として作成したものである。記事セットの全分類結果を足すと 100 を超えるが、これは一つの記事セットが複数の分類をもつことを許しているためである。分類によってばらつきはあるが、other と multi-facility を除いて全ての分類に記事セットが分類されている。分類の中では、event に関わる分類を割り当てられた記事セットが多く、single-event と multi-event 合わせて全体の 7 割を占めている。

4.1 被験者による記事セットの分類

3 人の被験者に、100 個の記事セットについて分類を行ってもらい、その結果を正解データに対して評価した。評価は、一つの記事に複数の分類が割り当てられていた場合には、そのうちのどれか一つでも一致していたら正解と判定している。各被験者に対する評価結果を表 2 に示す。表の値は、各被験者の分類結果に対する再現率と適合率である。記事セット全体に対して判定する場合は、再現率と適合率が同じになるので一つにまとめて示している。評価結果を見ると single-person に対する結果が最も良く、次いで single-event、multi-org、single-product の分類の評価が高い。個々の評価結果は被験者によって差があるが、分類全体としては平均して F 値で 83 ポイントの評価結果を得た。

4.2 自動分類

記事セット分類の正解データのうち、頻度の高い 5 つの分類 (single-person, multi-org, single-event, multi-event, single-product) を対象として、決定木による自動分類実験を行った。各決定木は一つの記事分類のみを対象とし、与えられた記事セットに対して、それが対象とする分類にあてはまるか否かを判定する。決定木作成には、固有表現抽出の結果に基いた以下の素性を用いた。固有表現抽出には、拡張された固有表現 [7] に基づいたパタンベースのシステムを用いている:

1. 記事セット中の固有表現の頻度・記事頻度: この素性は、記事セット中最も特徴的な固有表現を見つけるためのものである。
2. 記事セット中の固有表現クラスの頻度・記事頻度: 特定の固有表現 (token) だけでなく、固有表現のクラス (type) についても、そのクラスが特徴的であるかどうかを捉えるための素性である。
3. 「クラスターム」を主辞とする語句の頻度・記事頻度: 「クラスターム」とここで呼んでいるのは、例えば「法」、「法案」や「地震」など、一般名詞だがある固有表現のクラスを示すものであ

表 2: 被験者の結果に対する評価

	被験者 A		被験者 B		被験者 C	
分類	再現率	適合率	再現率	適合率	再現率	適合率
s-person	0.92 (11/12)	0.92 (11/12)	0.92 (11/12)	0.92 (11/12)	1.00 (12/12)	1.00 (12/12)
m-person	0.50 (1/2)	0.50 (1/2)	0.50 (1/2)	0.13 (1/8)	0.50 (1/2)	1.00 (1/1)
s-org	0.50 (4/8)	0.67 (4/6)	0.88 (7/8)	0.58 (7/12)	0.75 (6/8)	0.55 (6/11)
m-org	0.75 (6/8)	0.67 (6/9)	1.00 (8/8)	0.67 (8/12)	0.88 (7/8)	0.54 (7/13)
s-event	0.67 (29/43)	0.81 (29/36)	0.67 (29/43)	0.91 (29/32)	0.72 (31/43)	0.82 (31/38)
m-event	0.41 (11/27)	0.73 (11/15)	0.67 (18/27)	0.60 (18/30)	0.30 (8/27)	1.00 (8/8)
s-product	0.57 (13/23)	0.93 (13/14)	0.61 (14/23)	1.00 (14/14)	0.35 (8/23)	1.00 (8/8)
m-product	0.67 (2/3)	1.00 (2/2)	0.00 (0/3)	0.00 (0/2)	0.00 (0/3)	0.00 (0/0)
s-facility	1.00 (1/1)	0.33 (1/3)	0.00 (0/1)	0.00 (0/1)	0.00 (0/1)	0.00 (0/0)
m-facility	0.00 (0/0)	0.00 (0/0)	0.00 (0/0)	0.00 (0/0)	0.00 (0/0)	0.00 (0/0)
s-location	0.86 (6/7)	0.75 (6/8)	0.29 (2/7)	1.00 (2/2)	0.43 (3/7)	0.75 (3/4)
m-location	0.50 (1/2)	0.50 (1/2)	0.00 (0/2)	0.00 (0/3)	1.00 (2/2)	1.00 (2/2)
全体	0.84 (84/100)		0.86 (86/100)		0.79 (79/100)	

る。これらは、固有表現を補完するために導入している。クラスタームのリストは、質問応答システムのために作成されたもので、約 16000 語の単語から成る。同じクラスタームを主辞とする語句は、同一のものを示すとみなして一つにまとめた。固有表現と同様に、各表現 (token) の頻度とクラス全体 (type) の頻度の双方を用いる。

4. クラス中の表現の異なり数: クラスの頻度が大きい場合でも、そのクラス中に様々な固有表現が含まれる場合と、特定の固有表現が頻出している場合があるため、それを明示する素性を導入した。固有表現に対するクラスと、クラスタームに対するクラスの異なり数それぞれを素性として用いた。

さらに、上記の頻度・記事頻度・異なり数を求める範囲を見出し・一文目・記事全体の 3 種類に分け、それぞれ別の素性とした。表現の頻度は記事中の全固有表現数や全クラスターム数で、記事頻度と異なり数は記事セット中の記事数でそれぞれ正規化し、全て値域が $[0, 1]$ となるようにしている。

決定木の評価は、leave-one-out 方式によって行った。すなわち、99 個のデータについて決定木を作成し、残りの 1 個についてテストすることを 100 回繰り返してその評価の平均を取った。

4.3 評価結果

表 3 に決定木による分類の評価を、被験者の評価の平均値とともに示す。各分類に対する評価に加えて、対象とした 5 つの分類に対する結果をまとめて正解データと比較し、全体では F 値 56 ポイントの評価を得た。以下、各分類について評価結果と誤りに対する考察を述べる。

Single-Person single-person は被験者による評価結果が最も高い分類だったが、決定木による分類結果も single-person に対する結果が最も評価が良かった。判定を誤った記事セットについて見てみると、missing errors(付与すべきなのに与えなかったもの) の原因としては、共参照関係を取れていないものが 2 例あった。具体的には、人名の姓名と姓だけのものとを同一視する必要がある (例：松坂大輔→松坂)。spurious errors(付与すべきでないのに与えたもの) の原因としては、固有表現タグのミス (人名でないものを人名とタグ付け) が 2 例、single-event の動作主の名前を取ってきているものが 1 例あった。

Multi-Org 判定を誤った記事セットについて見てみると、missing errors の原因としては、固有表現やクラスタームのミスによるものが 2 例あった。また、multi-org の中でのタイプの違いによるもの (記事ごとに異なる組織名が現われる/複数の組織名が記事セット全体について出ている) が 1 例あった。この二つを分けて扱うためには、記事分類の定義の再検討が必要となる。spurious errors の原因としては、single-product が適切な分類であるものが 1 例あった。これは、ある特定の法案に関する議論が中心になった記事セットであるが、議論をしている政党名を取ってきた結果 multi-org に分類された。

Single-Event, Single-Product クラスタームと固有表現の異なり数の素性を入れない場合には single-event, single-product に対する分類結果はランダムな分類結果と変わらなかったが、素性を追加

表 3: 決定木による分類の評価結果

	決定木による分類の評価			被験者の平均評価値		
分類	再現率	適合率	F 値	再現率	適合率	F 値
s-person	0.83 (10/12)	0.77 (10/13)	0.80	0.95	0.95	0.95
m-org	0.63 (5/8)	0.83 (5/6)	0.71	0.88	0.62	0.72
s-event	0.65 (28/43)	0.76 (28/37)	0.70	0.69	0.84	0.76
m-event	0.22 (6/27)	0.38 (6/16)	0.28	0.46	0.78	0.55
s-product	0.63 (10/23)	0.44 (10/16)	0.51	0.51	0.98	0.66
全体		0.56			0.83	

することで結果の向上が見られた。single-event の分類に対する誤りの原因としては、イベントを示す表現が固有表現抽出では十分捉えきれていないことがあるが、それ以外にも同一のイベントを示す表現に多様性があり、それらの間での同一性を捉えられていないことが挙げられる (例：チェチェンへの攻撃とチェチェン空爆)。single-product に対しては、「内閣支持率」や「東証平均株価」などの指数に関する表現が、固有表現でもクラスタームでも捉えられなかったことが主な原因の一つであった。

Multi-Event　multi-event に分類されている記事セットは single-event に次いで 2 番目に多いにもかかわらず、F 値の結果が非常に低くなっている。実際、分類に当てはまるかどうかの判定自体の精度においても 0.69 と、ベースライン (全ての記事セットがその分類に当てはまらないとする場合) の 0.73 よりも低くなっており、分類のための素性が不足していると考えられる。multi-event の分類に対する誤りの傾向としては、クラスタームを主辞とする語句では、single-event と multi-event とを区別することができていないことが挙げられる。例えば、国内各地での複数の事故や選挙の話を含む記事セットと、ある特定の事故や選挙について述べている記事セットが同様に扱われている。クラスタームを主辞とするより詳細な語句の扱いを検討する必要がある。

5　まとめと今後の課題

本稿では、複数記事要約のために記事セットの主題を分類することを目標として、無作為に作成した 100 個の記事セットについて固有表現に基いて定義した分類を与え、それに対する三人の被験者による分類付与の評価結果と、機械学習に基いた自動分類実験の結果を示した。被験者による分類付与では、一つの記事セットに複数の分類の付与を許した場合平均して F 値で 83 ポイントの評価結果が得られた。決定木を用いた自動分類実験では、頻度の高い 5 つの分類について固有表現の頻度などを素性として決定木を作成し、F 値で 56 ポイントの評価を得た。また個々の分類に対して、その分類結果の評価と分類誤りの分析を示した。今後の課題としては、クラスタームリストの拡張や日付情報の導入を行って分類の精度を向上させることを考えている。また、イベント表現や製品名については、同じ対象を表すのに固有表現の範囲におさまらない多様な表現が用いられており、これらを処理する方法を見出すために分析をさらにすすめたいと考えている。

参考文献

[1] DARPA. *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, Fairfax, VA, USA, May 1998.

[2] DUC. http://duc.nist.gov, 2001–. Document Understanding Conference.

[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 Proc. of DUC2001*, 2001.

[4] NII, editor. *Proceedings of the Third NTCIR Workshop*, Tokyo, Japan, October 2002. National Institute of Informatics.

[5] C. Nobata, S. Sekine, and H. Isahara. Evaluation of Features for Sentence Extraction on Different Types of Corpora. In *Proceedings of the MSQA Workshop in conjunction with ACL2003*, pp. 29–36, 2003.

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

[7] S. Sekine, K. Sudo, and C. Nobata. Extended Named Entity Hierarchy. In *Proceedings of the LREC-2002 Conference*, pp. 1818–1824, 2002.

[8] IREX 実行委員会（編）. IREX ワークショップ予稿集. IREX 実行委員会, 9 月 1999.

[9] 野畑周, 関根聡, 井佐原均. 複数記事要約のための記事セットの分類. 情報アクセスのためのテキスト処理シンポジウム, 2 月 2003.

[10] 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.