複数文書からの重要情報の抽出と表の生成

岡﨑 健介 [*1]　村田 真樹 [*2]　馬 青 [*3]
[*1] 鳥取大学 工学部 知能情報工学科
[*2] 鳥取大学大学院 工学研究科 情報エレクトロニクス専攻
[*3] 龍谷大学 理工学部 数理情報学科
[*1,*2]{s142017,murata}@ike.tottori-u.ac.jp
[*3] qma@math.ryukoku.ac.jp

1　はじめに

関連した事柄を調査する際，重要な項目ごとに情報を表に整理することで，その情報を使う人にとって，可読性や利便性が向上すると考えられる．赤野らの研究 [1] では，word2vec を用いて，Wikipedia の城に関するページに出現する単語をベクトルで表現し，これをクラスタリングした後，表の形で整理していた．word2vec では，周辺の単語を考慮してを単語ベクトルを求めるので，例えば人名の単語であっても，周辺に出現する単語の違いによって単語ベクトルが変わるため，人名を城に果たした役割ごとに分類できる．しかし，先行研究では単語のみを表に整理するため，城に果たした役割ごとに情報を分類できていたとしても，役割についての情報情報が不足していた．本研究では，複数の文書から重要な情報を文単位で抽出し表に整理することで，単語単位の情報抽出では情報が不足していたものを改善することを目的とする．本研究での主張点は以下の 3 点である．

新規性

赤野らの研究 [1] などの従来手法では情報を「毛利豊元」のように単語単位で抽出していたが，本研究では情報を「毛利豊元が城主となった。」のように文単位で抽出をする．

有用性

文単位の情報を抽出し表に整理することで，単語単位の情報抽出では情報が不足していたものを改善できるという有用性がある．

性能

整理した表の列ごとの情報抽出の再現率は 0.62 と低い値となったが，適合率が 0.90 と高い値となった．

2　提案手法

2.1　提案手法の手順

本研究では図 1 のように，関連する内容の複数の文書から情報を抽出し，抽出された情報を文書ごとに表に整理する手法を提案する．具体的な手順を以下に示す．

手順 1 複数文書を文単位に分割する．
手順 2 手順 1 で分割された各文の文ベクトルを計算する．
手順 3 文ベクトルを x-means 法 [2, 3] でクラスタリングする．
手順 4 手順 3 で得られた全てのクラスタを見るのは大変なので，クラスタの重要度を計算し，重要度の高い順に，行を文書，列をクラスタとする表に整理することで可読性を高める．
手順 5 表の各クラスタに項目名を付与し，クラスタにどのような情報が含まれるかをわかるようにする．
手順 6 表に文のままで情報が含まれていると見づらいので，各クラスタをさらに格要素ごとに整理することで可読性を高める．

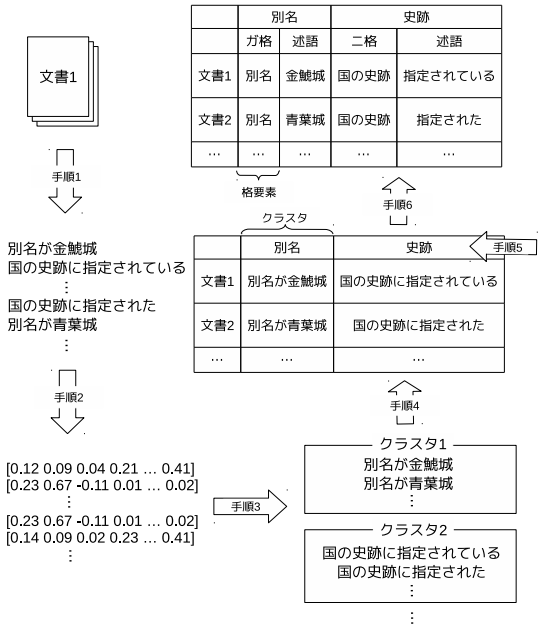

図 1　提案手法の概要図

2.2 文の分割方法

2.1 節の手順 1 における複数文書を文単位に分割する方法を説明する．文書を句点ごとに分割したものを文とした場合，例えば「人口は 98,891 人で，面積は 611.76km^2。」のような複数の情報を含む文が存在してしまう．このような文は，「人口は 98,891 人。」という人口に関する文と，「面積は 611.76km^2。」という面積に関する文に分割されることが望ましい．よって以下の手順で文を分割し，得られた短い文を本研究では 1 つの文として扱う．図 2 に分割結果の例を示す．

1. 文を KNP[*1]を用いて構文解析する．
2. 条件 (a)，(b) を同時に満たす文節箇所で分割する．
   （a）文節の係り先が末尾の文節番号である．
   （b）並列構造を表す<P>が付与されている．
3. 分割された文に対しても，文を分割できなくなるまで 1，2 を行う．
4. 分割された各文を KNP で格解析する．
5. 出力された格解析結果のうち，係り先が末尾の文節番号である文節，もしくは末尾の文節に注目する．
6. 注目している格解析結果に含まれる各格要素について，述語よりも前にある場合は，格要素を格要素に係る文節と統合する．
7. 格要素と述語をまとめて文を作る．

分割前

流域には貴重な生態系が拡がっていたが、噴火によって大半の渓谷が分厚い火山堆積物の底に埋もれた。

分割後

貴重な生態系が流域に拡がっていた。
噴火により大半の渓谷が分厚い火山堆積物の底に埋もれた。

図 2　分割結果の例

2.3 文ベクトルの計算

2.1 節の手順 2 における文ベクトルの計算方法を説明する．文ベクトルは以下の手順で求める．

1. 文を格要素ごとに分割する．
2. 分割された格要素ごとに以下の手順で格要素ベクトルを求める．
   （a）文を MeCab[*2]を用いて形態素解析する．
   （b）形態素解析結果のうち，品詞が名詞で，かつ，品詞分類 1 が代名詞，数，非自立，副詞可能でない単語を抽出する．
   （c）抽出した単語のベクトルの平均を格要素ベクトルとする．
3. 格要素ベクトルの総和を文ベクトルとする．

2.4 単語ベクトルモデル

2.3 節の文ベクトルの計算で用いる単語のベクトルには，fastText[4, 5] によって学習させたものを使用した．fastText は隠れ層と出力層からなる 2 層のニューラルネットワークで，隠れ層が単語の分散表現に相当する．

今回は学習データとして，アルファベットとカタカナは全角に，英数字は半角に統一した Wikipedia の全 1,061,375 記事 (6 月 1 日時点) を使用した．また，単語ベクトルの次元数は 300 次元とした．

2.5 x-means 法

本研究では文ベクトルのクラスタリングに x-means 法を用いる．x-means 法は，k-means 法を拡張した手法である．k-means 法では，あらかじめクラスタの数を指定する必要があるが，x-means 法では以下の手順により，最適なクラスタの数を推測できる．

1. クラスタ数 $k = 2$ で再帰的に k-means 法を実行する．
2. クラスタリング前後のベイズ情報量基準 BIC を比較する．
3. BIC の値が小さくなる限りこれを続ける．

2.6 重要度の計算方法

2.1 節の手順 4 におけるクラスタごとの重要度の計算方法を説明する．

クラスタリング結果には表 1 の (a) ように関連する文だけで構成される密集率の高いクラスタもあれば，表 1 の (b) のように関連性のない文で構成される密集率の低いクラスタもある．密集率の高いクラスタほど重要であると考えられる．よって，k 番目のクラスタの密集率 d_k を式 1 のように定める．ここで，N_k は k 番目のクラスタに含まれる文の総数であり，$S_{k,l}$ は k 番目のクラスタに含まれる l 番目の文のベクトルであり，$S_{k,mean}$ は k 番目のクラスタに含まれる文のベクトルの平均である．

$$d_k = \frac{1}{N_k} \sum_{l=1}^{N} \frac{S_{k,l} \cdot S_{k,mean}}{|S_{k,l}||S_{k,mean}|} \quad (1)$$

[*1] http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP

[*2] http://taku910.github.io/mecab/

表 1　クラスタの密集率の例

（a）文の密集率が高いクラスタの例

	クラスタ 1
大阪城	国の史跡に指定された
熊本城	国の史跡に指定されている
広島城	国の史跡に指定された
鳥取県	国の史跡に指定された

（b）文の密集率が低いクラスタの例

	クラスタ 2
大阪城	大阪城は日本の城の一つ
熊本城	日本の100名城に選定された
広島城	別名は鯉城
鳥取城	天守台、石垣、堀などが残る

多くの文書の情報を含むクラスタほど重要であると考えられる．よって，k 番目の文書カバー率 c_k を式 2 のように定める．p_k は k 番目のクラスタにおいて文を抽出できた文書の数であり，P は文書の総数である．

$$c_k = \frac{p_k}{P} \tag{2}$$

k 番目のクラスタの重要度 i_k を式 3 のように定義する．

$$i_k = d_k \times c_k \tag{3}$$

2.7　クラスタの項目名の求め方

2.1 節の手順 5 におけるクラスタごとの項目名の求め方の概要を図 3 に示す．生成された表の各クラスタについて，以下の手順でクラスタの項目名を付与する．

1. クラスタに含まれるの各文について，文に含まれる単語のうち品詞が名詞のものを抽出する．
2. 1 で抽出した各単語について，文書頻度を求める．
3. 文書頻度が最大の単語をクラスタの項目名として付与する．
4. 文書頻度が最大の単語が複数ある場合は，読点で区切って全て付与する．

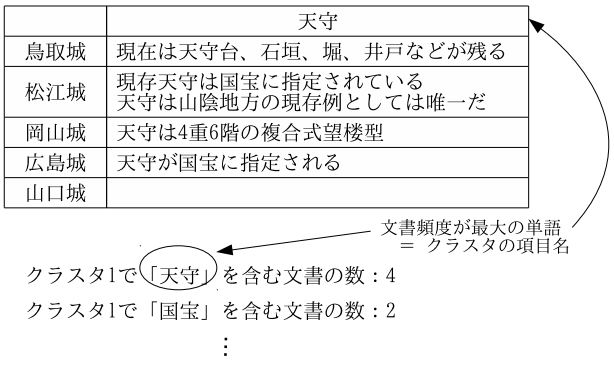

図 3　クラスタの項目名の求め方の例

3　実験

3.1　実験データ

実験で用いる複数文書データとして，日本の 100 名城のうち，中国四国地方に位置する城名を見出しとする Wikipedia(6 月 1 日時点) の記事 22 件を使用した．

3.2　実験結果

出力された表のうち，重要度が最も高いクラスタの一部を表 2 に示す．ここで，1 行目はクラスタの項目名を，2 行目はクラスタに含まれる文を，格要素ごとに列に整理した際の対応する格を示す．

表 2　出力された表の一部

	指定		
格	ガ	ニ	述語
松山城 (備)	城跡が	国の史跡に	指定され、
丸亀城	城跡の全域は	国の史跡に	指定されており
岡山城		史跡にも	指定されている
大洲城			
月山富田城	城郭跡は	国の史跡に	指定されている
津和野城	城跡は	国の史跡に	指定されている
津山城	城跡は	国の史跡に	指定されている
高知城	城跡は	国の史跡に	指定されている
宇和島城	城跡は	国の史跡に	指定されている

3.3　情報抽出の評価

式 6 より F 値を求め，情報抽出の性能を評価する．式中の「クラスタの項目名を中心とする文」の例を表 3 に示す．表 3 の 2，3 行目の文は項目名「天守」を中心とする文であるが，4 行目「天守台の石垣は拡張したことが確認できる」は，項目名「天守」という単語を含むが，「石垣」を中心とする文であるので，列の項目名を中心とする文ではないと判断した．重要度の高い上位 3 列の評価結果は表 4 のようになった．

$$\text{適合率} = \frac{\text{クラスタの項目名を中心とする文の数}}{\text{クラスタに含まれる文の数}} \tag{4}$$

$$\text{再現率} = \frac{\text{クラスタの項目名を中心とする文の数}}{\text{クラスタの項目名を中心とする文書中の文の数}} \tag{5}$$

$$F\text{ 値} = \frac{2 \times \text{適合率} \times \text{再現率}}{\text{適合率} + \text{再現率}} \tag{6}$$

表 3　クラスタの項目名に関連した文の例

	天守
岡山城	天守は 4 重 6 階の複合式望楼型○
広島城	天守が国宝に指定される○
鳥取城	天守台の石垣は拡張したことが確認できる×

3.4　単語単位の情報抽出との比較

単語単位の情報抽出と比較するため，赤野らの研究 [1] の手法で情報抽出を行う．実験で用いる複数文書データと単語ベクトルモデルには，文の情報抽出で用いたもの

表 4　情報抽出の評価結果

列番号	1 列目 (指定)	2 列目 (選定)	3 列目 (天守)	平均
適合率	1.00(10/10)	0.86(12/14)	0.85(17/20)	0.90
再現率	0.53(10/19)	1.00(12/12)	0.33(17/52)	0.62
F 値	0.69	0.92	0.47	0.69

と同じものを使用した．単語単位の情報抽出では表 5 のように人名のみのクラスタができてしまい，それぞれの人名と城との関係性がわからなかったが，文単位の情報抽出では表 6 のように入封した人名の形で整理され，人名と城との関係性がわかるようになった．また，表 5 には含まれるが，表 6 に含まれない他の人名についても，項目名が「入城」,「城主」などのクラスタに含まれており，人名と城との関係性がわかるようになった．

表 5　単語単位の情報抽出の結果の一部

	クラスタ 16
松山城 (備)	秋庭，三郎，重信，政一，勝美，水谷，小堀，政次
丸亀城	
岡山城	豊，忠雄，池田，光政，新田
大洲城	直之，加藤
月山富田城	堀尾
津和野城	三上，織部
鳥取城	光政，久松，誠，潤，宮部，長吉

表 6　文単位の情報抽出の結果の一部

	封					
格	ガ	ヲ	ニ	デ	時間	述語
松山城 (備)	水谷勝隆が			5 万石で		入封
	安藤重博が	入封		6 万 5000 石で	元禄 8 年	するが
	石川総慶が	入封		6 万石で	同年	した
丸亀城						
岡山城	忠継の弟・忠雄が	入封		31 万 5 千石で	元和元年	した
	池田光政が	入封		31 万 5 千石で		した
大洲城	その小早川隆景が	入封	伊予に	35 万石で		し、
	加藤貞泰が			6 万石で	元和 3 年	入り、
月山富田城						
津和野城	亀井政矩が			4 万 3 千石で	元和 3 年	入城
鳥取城						

3.5　考察

表 4 から，重要度の高い上位 3 列の適合率の平均が 0.90，再現率の平均が 0.62 となり，適合率に比べ再現率が低い傾向にあった．この原因としては，内容が関連する文が正しく同じクラスタに割り当てられないことが考えられる．「天守が 1950 年に重要文化財に指定される」と「江戸時代に建造された天守、二重櫓、土塀の一部が重要文化財に指定されている」のように，同じ内容を表す文同士であっても，含まれる単語が大きく異なる場合は，2.3 節の方法で文ベクトルを求めると，文ベクトル同士の違いが大きくなる．文ベクトルの違いにより，これらの文が異なるクラスタに割り当てられることで，再現率が低い値となってしまう．これらの文が同じクラスタに含まれるようにするには，文ベクトルを文中の単語ベクトルだけでなく，文の構造も考慮して求める必要があると考えられる．

3.4 節の単語単位の情報抽出との比較では，単語単位の情報抽出では城と人名との関係性がわからなかったものが，文単位では改善された．しかし，「また，重要文化財にも指定されている」といった文のように，文中の主語や目的語が省略されている文では依然として情報の不足が見られた．この問題は，事前に文を照応解析することによりある程度改善できると考えられる．

4　おわりに

本研究では，関連する複数の文書から重要な情報を文単位で抽出し表に整理した．文単位の情報抽出では表 6 のように，単語単位の情報抽出ではわからなかった「人物が入封した日時」や「人物が入封した土地の生産性」などの情報を取り出すことが可能となった．

一方で，主語や目的語が省略されている文では，文単位で情報を抽出しても情報が不足してしまった．また，文の内容が関連しているにも関わらず，文の書きぶりの違いから文ベクトルの差異が大きくなり，結果としてこれらの文が複数のクラスタに分裂するケースがいくつか見られた．よって，文単位の情報抽出での情報不足を解消することと，内容の関連する文の分裂を抑えることを今後の課題としたい．

謝辞

本研究は科研費 (26330252) の助成を受けたものである．ダットジャパン株式会社の羽田典久氏らとの議論が参考になった．

参考文献

[1] Hokuto Akano, Masaki Murata, and Qing Ma. Detection of inadequate descriptions in wikipedia using information extraction based on word clustering. IFSA-SCIS 2017, pp. 1–6, 2017.

[2] D. Pelleg and A. Moore. X-means: Extending k-means with efficient estimation of the number of clusters. In *ICML '00 Proceedings of the Seventeenth International Conference on Machine Learning*, pp. 727–734, 2000.

[3] 石岡恒憲. x-means 法改良の一提案 : k-means 法の逐次繰り返しとクラスターの再併合. 計算機統計学, 第 18 巻, pp. 3–13, 2006.

[4] Piotr Bo-janowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. In *Transactions of the Association for Computational Linguistics*, Vol. 5, pp. 135–146, 2017.

[5] Armand Joulin, Edouard Grave, Piotr Bo-janowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In *arXiv preprint arXiv:1607.01759*, 2016.