Wikipedia 記事構造のモデル化による用語説明の自動編集

藤井裕也　　藤井敦　　徳永健伸
東京工業大学大学院情報理工学研究科

1 はじめに

科学技術と文化の発展により，Web 検索や事典を使って様々な用語について調べる機会が増加している．Web 検索は，膨大な情報が得られるという情報の量に関する利点がある反面，情報の質に関する欠点がある．事典は統制された情報が得られるという情報の質に関する利点がある反面，情報の量に関する欠点がある．

本研究は，Web 検索と事典の利点を統合し，Web テキストから事典的な説明記事を自動編集することを目的とする．具体的には，Web 上の百科事典である Wikipedia において用語の説明が編集される仕組みを学習し，「用語説明モデル」を構築する．用語説明モデルにおいて，例えば動物名なら「生態」や「形態」など，病名なら「症状」や「治療」などのように，用語の種類毎に説明に要する観点が異なる点に着目し，用語の種類毎に観点の集合を定義する．用語を検索して得られる Web テキスト集合を用語説明モデルに基づいて観点に分類することで，その用語に関する統制された説明記事を生成する．その結果，Wikipedia に未収録の用語に関する説明記事が得られるだけでなく，Wikipedia に記載されている用語に関しても不足している情報を補うことができる．本研究は Web ページを対象とした複数文書要約と見なすことができる．

2 関連研究

Biadsy ら [1] は Wikipedia の人物に関する記事の集合から人物記述に関するモデルを学習し，人物情報の要約に利用した．Wikipedia から要約に関するモデルを学習する点で本研究と類似する．しかし，Biadsy らは観点には着目しておらず，要約の対象も人物情報に限定されている．

Ye ら [7] は Wikipedia 記事内のテキストから要約文候補を選択することで Wikipedia 記事の自動要約を行った．そのために，Wikipedia 記事内のリンク，infobox，セクション などの情報を利用して Extended Document Concept Lattice と呼ばれるモデルを構築した．Ye らの手法は Wikipedia 記事そのものの要約を目的としているため本研究とは目的や方法が異なる．

Sauper ら [5] は「病名」などの特定のドメインに関する観点を Wikipedia 記事から抽出してテンプレートを作成し，テンプレートと Web テキストを用いることで用語に関する要約を自動生成した．説明の観点に着目している点や Web テキストから要約文の候補を選択する点で本研究と類似する．しかし，ドメインが既知の用語しか扱えず，多義語に対応できないという問題点がある．

Fujii [3] は Wikipedia の記事構造をモデル化することで，Web テキストから用語に関する説明記事を自動編集した．Wikipedia 記事の目次情報であるセクション名を観点の候補とみなすことで用語の種類毎に観点集合を定義した．この手法では用語の種類毎に人手で Wikipedia 記事を収集しているため，新しい用語の種類を追加する度に収集のためのコストがかかる（拡張性問題）．また，セクション名によって観点を定義しているため，「経歴」と「略歴」のように内容が重複する観点が定義されてしまう（冗長性問題）．本研究は Fujii の手法における上記 2 つの問題を解消する手法を提案する．

3 用語説明の自動編集手法

3.1 概要

提案手法の概要図を図 1 に示す．本手法は図 1 上部の「用語説明のモデル化」と図 1 下部の「用語説明の自動編集」から成る．

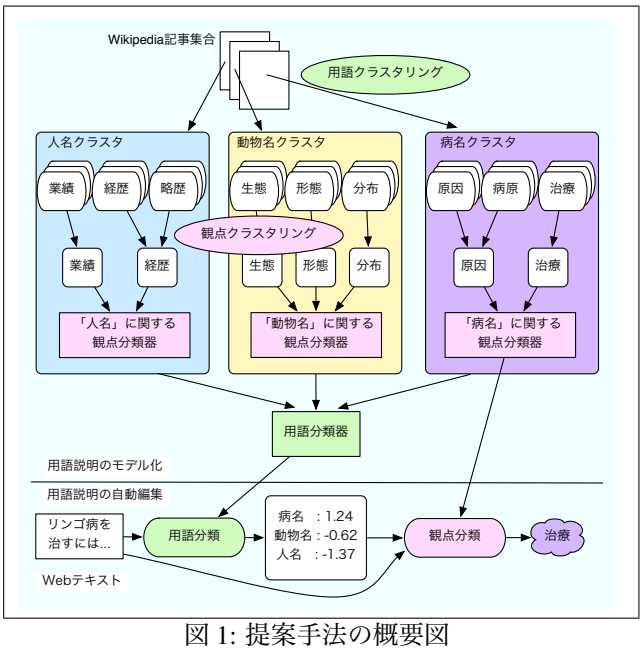

図 1: 提案手法の概要図

用語説明のモデル化では，テキスト中の用語の種類を分類するための用語分類器と，テキストの観点を分類するための観点分類器を学習する．モデル化において，「拡張性問題」を解消するために，Wikipedia 記事を観点の類似性に基づいて自動的にクラスタリングし，生成されたクラスタを用語の種類として定義する（用語クラスタリング）．また，「冗長性問題」を解消するために，

Wikipedia 記事のセクションをクラスタリングすることで内容の類似するセクションを統合し，用語の種類毎に観点集合を定義する（観点クラスタリング）．最後に記事内のセクション本文を学習データとして用語分類器及び観点分類器を学習する．

用語説明の自動編集では．用語を検索して得られた Web テキストの集合を用語説明モデルに基づいて観点に分類することで，用語が持つ観点毎に代表的なテキストを引用し，説明記事を自動編集する．

以降の各項で提案手法の各手順について詳説する．

3.2 用語クラスタリング

1 節で説明したように，用語の種類によって説明に要する観点の集合が異なる．また，セクション名は観点を表すと見なすことができる．そこでセクション名の形態素（名詞，形容詞）を素性として用語の種類をクラスタリングする．対象記事内に含まれるセクション名だけではセクション数が少ないため十分な情報が得られない可能性がある．そこで，対象記事と同じ Wikipedia カテゴリに属する記事に含まれるセクション名の形態素も素性として加える．この素性を「拡張セクション名」と呼ぶ．カテゴリ名自体も有用な情報と考えられるので，カテゴリ名の形態素（名詞，形容詞）も素性として用いる．更に，記事本文の類似性を反映させるために対象記事内に含まれる形態素（名詞，動詞，形容詞）を素性に用いる．ここで，カテゴリ情報によって用語の種類を直接決定する方法もある．しかし，カテゴリはノイズが多いためクラスタリングにおける素性の 1 つとして利用する．

用語クラスタリングは観点の類似性を基準としているため，観点クラスタリングの結果に影響を受ける．そこで観点クラスタリングの結果を素性に加えることを考える．観点クラスタリングの結果は，(1) のように各セクションが各観点クラスタに所属する度合が付与された形(所属度ベクトル）で表わされる．対象記事内の各セクションの所属度ベクトルを合計したベクトルを素性として用いる．所属度の低い観点クラスタの素性はノイズとなる可能性があるため，所属度の高い上位 5 件の観点クラスタ以外の所属度は 0 とする．

(1) ライオン//生態	1:0.133	2:0.341	3:0.277	...
ライオン//形態	1:0.485	2:0.197	3:0.085	...
記事名//セクション名	観点クラスタ番号:所属度 ...			

用語クラスタには名前がついていないため，カテゴリを利用して各用語クラスタに名前を付ける．用語クラスタ内の各記事が属するカテゴリ名を集計する．そして用語クラスタ内で頻出し，かつそのクラスタに特有なカテゴリ名をそのクラスタの名前として採用する．具体的にはカテゴリ名の自己相互情報量と頻度の積をスコアとし，スコアが最も高いカテゴリ名を採用する．例えば，動物名の記事が多く含まれる用語クラスタには「脊椎動物」というカテゴリ名が用語の種類名として採用される．

3.3 観点クラスタリング

セクション名が完全一致または部分一致するセクション同士は類似した内容である可能性が高いため，セクション名の形態素（名詞，形容詞）を素性に用いる．セクション本文の形態素（名詞，動詞，形容詞）も素性に用いる．

2 つのセクションが類似した内容かどうかを判定する際，それらが同じ用語の種類であるという情報は重要な手がかりとなる．そこで用語クラスタリング結果を素性に加える．用語クラスタリング結果は (2) のように各記事が各用語クラスタに所属する度合いが付与された所属度ベクトルで表わされるので，この記事の所属度ベクトルを素性に用いる．この際，ノイズを避けるために所属度の最上位の用語クラスタ以外の所属度は 0 とする．

(2) ライオン	1:0.034	2:0.898	3:0.247	...
ハクビシン	1:0.152	2:0.825	3:0.328	...
痛風	1:0.743	2:0.153	3:0.212	...
記事名	用語クラスタ番号:所属度 ...			

生成した用語クラスタ及び観点クラスタを用いて用語の種類毎に観点集合を決定する．Fujii の手法では用語の種類毎に記事内に含まれるセクション名を集計し，閾値以上の件数が得られたセクション名を観点として定義していた．本手法はセクション名を集計する代わりに，セクションが属する観点クラスタを集計することで観点を決定する．

観点クラスタ内には「経歴」と「来歴」のように内容が類似していて表現が異なるセクション名が含まれている．これらをまとめるために日本語語彙大系 [6] の意味カテゴリを用いる．例えば,「経歴」と「来歴」は共に【伝承】という意味カテゴリに属している．観点クラスタに含まれるセクション名を一度意味カテゴリ名に置き換えてから集計し，最も頻度の高い意味カテゴリ名をその観点クラスタのラベルとする．ラベルが一致した観点クラスタをマージした後，ラベルを意味カテゴリに置き換える前のセクション名で最頻出のセクション名に戻す．このとき元のセクション名が同じ観点クラスタ同士もマージする．

最後に，用語の種類毎に観点集合を決定する．用語クラスタに含まれる記事内のセクションが属する観点クラスタを集計し，頻度が 50 以上または上位 3 位以内の観点クラスタを観点とする．

3.4 用語分類器及び観点分類器の学習

3.2 で得られた用語クラスタと 3.3 で得られた観点集合を用いて用語分類器及び観点分類器の学習を行い，用語説明モデルを構築する．学習には Support Vector Machine（SVM）[2] と one-vs-rest 法を用いて多値分類を行う．用語クラスタに含まれる記事内のセクション本文を訓練事例として用語分類器を学習する．素性として，セクション本文の形態素（名詞，動詞，形容詞）及びそのセクションが属する記事名の形態素（名詞，形容詞）を用いる．次に用語の種類毎に観点分類器を学習する．各観点を構成する観点クラスタ集合に含まれるセクション本文を訓練事例とし，セクション本文の形態素（名詞，動詞，形容詞）及びセクション名の形態素（名詞，形容詞）を素性として用いる．

3.5 用語説明の自動編集

3.4 で構築した用語説明モデルを用いて用語説明の自動編集を行う．まず，自動編集の対象語を Web 検索エンジンを用いて検索し，スニペットを収集する．スニペットを形態素解析して，名詞，動詞，形容詞を素性とする特徴ベクトルを作る．次に，スニペットの特徴ベクトルを用語分類器及び観点分類器に入力し，用語の種類毎に各観点へのスコアを求める．用語分類の結果，最も多く分類された用語の種類を対象語の種類とする．そして対象語の種類が持つ観点毎にスコアの高い順に N 件ずつスニペットを代表文として引用することで説明記事を自動編集する．このとき，異なる観点で同一のスニペットが選ばれた場合，観点内での順位が高い方を優先し，同率の場合はスコアの高い方を優先する．予備実験により $N = 3$ とした．

各観点の代表文として用いるスニペットの選び方として，「決定的手法」と「非決定的手法」の 2 通りを考える．決定的手法は Fujii の手法と同様に，まず用語分類器によってスニペットの用語の種類を一意に決定し，その用語の種類に対応する観点分類器を用いて各観点へのスコアを求める．この方法は用語分類を誤ると観点分類の結果も誤るという問題点がある．非決定的手法は，用語分類器によって用語の種類を一意には決定せず，全ての用語の種類に対する観点分類器を用いて各観点へのスコアを求める．そして用語分類器と観点分類器のスコアの合計を各観点へのスコアとする．スニペットに対して全ての用語の種類の各観点にスコアが付くので，用語分類の誤りに対して頑健である．

4 評価実験

4.1 実験データ

評価実験には Fujii [3] が使用した評価用データを用いた．このデータは Wikipedia 記事に用語の種類を人手でラベル付けしたデータであり，10 種類のラベルが付与されている．データを使用する前に，「外部リンク」や「概要」などの観点を表さないセクションを除外してフィルタリングを行った．また，本文が 10 形態素未満の短いセクションも除外した．この操作によって有効なセクションがなくなった用語を除外した．フィルタリング後，実験データの記事数は 4,870 であり，種類毎の内訳は以下の通りである．

人名（910），企業名（691），料理名（559），映画名（600），植物名（511），動物名（554），病名（461），虫名（257），スポーツ名（172），魚類名（155）

4.2 用語/観点クラスタリングの評価

クラスタリングには bayon[1] を用いた．bayon はクラスタリング手法として Repeated Bisection 法を採用している．Repeated Bisection 法は，k=2 とした時の k-means 法によるクラスタリングを繰り返す事でクラスタを 2 分割していき，あらかじめ定めたクラスタ数まで分割された段階でクラスタリングを終了する．クラスタリングの評価には，純度と逆純度及びその調和平均を取った F 値を用いた．

用語クラスタリングの評価では人手でラベル付けした用語の種類を正解クラスタとした．生成する用語クラスタ数は，正解クラスタと同じ 10 とした．観点クラスタリング結果を用いない場合の結果を表 1 に示す．拡張セクション名（exs）を用いる場合が全体的に良く，その他の素性を加えても大きな差は見られなかった．すなわち，拡張セクション名は対象記事が持つ潜在的な観点集合を表していることが分かる．また，カテゴリ名（cat）は有効ではなかったことから，カテゴリにはノイズが多く含まれていることが分かる．

表 1: 用語クラスタリングの結果（観点クラスタなし）

素性セット	bow	cat	sec	exs	cat sec	cat exs	bow cat exs
純度	0.866	0.797	0.672	0.903	0.843	0.902	0.895
逆純度	0.847	0.773	0.614	0.921	0.779	0.917	0.903
F 値	0.857	0.785	0.642	0.912	0.810	0.910	0.899

bow:本文内の形態素，cat:カテゴリ名の形態素，
sec:記事内のセクション名の形態素，
exs:拡張セクション名の形態素

表 2: 用語/観点クラスタリングの繰り返しによる F 値の変化

回数	1	2	3	4
用語	0.912	0.928	0.911	0.911
観点	0.360	0.344	0.327	0.351
用語	0.922	0.916	0.912	0.915
観点	0.243	0.357	0.361	0.336

3 節で説明したように，用語クラスタリング及び観点クラスタリングは互いに影響を与える．従って互いの結果を素性に用いながら両クラスタリングを交互に行うことで互いにクラスタが洗練されることが期待できる．用語クラスタリングから始めた場合と観点クラスタリングから始めた場合それぞれについて，各クラスタリングを 4 回まで交互に実行した際の F 値の変化を表 2 に示す．なお，観点クラスタリングの評価ではセクション名に正解ラベルが付与されていないため純度と逆純度を計算することができない．間接的な評価として，各セクションが属する記事に付与されている用語の種類をそのセクションの正解ラベルとして純度と逆純度を計算した．また，生成する観点クラスタ数は 100 とした．

用語クラスタリングと観点クラスタリングのどちらから始めても，最初に互いを素性に用いた際は独立にクラスタリングするよりも結果が良くなった．しかし，2 または 3 回以上繰り返しても結果は悪くなった．そこで，用語クラスタリングから開始し，2 回目の用語クラスタリングで止めた．

4.3 用語説明の自動編集の評価

評価用の用語について，提案手法で自動生成した記事がその用語に関する Wikipedia 記事にどの程度類似するかを評価した．具体的には，ROUGE-1 [4] を使い，

[1] http://code.google.com/p/bayon/wiki/Tutorial_ja

Wikipedia 記事と提案手法で生成した記事との間の unigram 一致率に基づいて，再現率，精度，F 値を評価した．ROUGE は本来再現率のみを計算する．しかし，要約文の長さが一定でない場合のために精度，F 値も考慮する．

4.1 で説明した 4,870 語を評価に用い，5 分割交差検定を行った．用語分類器及び観点分類器の学習には Tiny SVM[2]を線形カーネルで使用した． スニペットの取得は Web 検索エンジン Bing[3]を用いた．各評価用記事について記事名をクエリとして検索し，得られたスニペットの上位 100 件を要約文の候補として用いた．ただし，Wikipedia 記事を引用しているサイトからのスニペットは除外した．スニペットが 100 件得られなかった記事については評価対象から外した結果，評価対象の用語は 4,648 語になった．

用語説明モデルに提案手法と Fujii の手法をそれぞれ用いた場合の結果を比較した．また，スニペットの選択方法として決定的手法と非決定的手法をそれぞれ用いた場合で比較を行った．スニペットの選択方法におけるベースラインとして，まずスニペット 100 件を用語分類し，最も多く分類された用語の種類 i を求める． 次に，用語分類で i に分類されたスニペットを検索結果の上位から順番に（i の観点数）$\times$ 3 件だけ引用する．

表 3: 各手法の ROUGE-1 による比較

用語説明モデル	スニペット選択	再現率	精度	F 値
Fujii の手法	決定的	0.289	0.179	0.175
Fujii の手法	非決定的	0.292	0.180	0.176
Fujii の手法	ベースライン	0.281	0.181	0.174
提案手法	決定的	0.275	0.178	0.172
提案手法	非決定的	0.283	0.181	0.175
提案手法	ベースライン	0.271	0.183	0.174

各手法を比較した結果を表 3 に示す．スニペットの選択方法を比較すると，どちらの用語説明モデルを用いた場合でも非決定的手法が決定的手法及びベースラインよりも高い F 値となった．両側 t 検定を行ったところ，非決定的手法と決定的手法を比較すると，用語説明モデルに提案手法を用いた場合は有意水準 0.1%で有意差あり，Fujii の手法を用いた場合は有意水準 1%で有意差ありとなった．また，非決定的手法とベースラインを比較した場合，用語説明モデルに提案手法を用いた場合は有意水準 5%で有意差あり，Fujii の手法を用いた場合は有意水準 1%で有意差ありとなった．このことから，スニペットの選択手法としては非決定的手法が最も良かった．

次に用語説明モデルによる違いを比べると，全体的に Fujii の手法の方が F 値が高い結果となった．しかし，スニペット選択に非決定的手法を用いた場合，提案手法と Fujii の手法との間には有意水準 5%で有意差は見られなかった．

以上の結果から，提案手法は Fujii の手法の拡張性問題及び冗長性問題を解消しつつ，Fujii の手法に必要な人手の負荷を軽減することができた．

[2]http://chasen.org/˜taku/software/TinySVM/
[3]http://www.bing.com/

提案手法によって「スイギュウ」をクエリとして実際に自動編集を行った結果の一部を以下に示す．用語分類によって「スイギュウ」は動物名に分類され，次のような観点が得られた．

・分布
動物図鑑・偶蹄目ウシ科、アフリカスイギュウの生態や習性、分布域や生息地などについて、アフリカスイギュウの写真と一緒に詳しく紹介しています。 ... アフリカスイギュウはサハラ以南のアフリカ大陸に広く分布し、主に水辺の草原地帯などで生活している。
・分類
安佐動物公園のアフリカスイギュウ。偶蹄目ウシ科アフリカスイギュウ属に分類され、アフリカのサハラ以南に分布しています。天敵はライオンだけだとも言われる強い植物食の動物ですが、子供の頃はブチハイエナなどに捕食されることも多いそうです。
・形態
アフリカスイギュウは、サハラ以南のアフリカに生息し、サバンナや疎林（そりん）にすんでいます。体は黒色、大型で、断面が三角形の大きな角を持っています。オスは体長 300cm、体重 600kg を超すものが ...

5 おわりに

本研究は，Wikipedia 記事集合から用語説明モデルを構築し，Web テキスト集合から用語に関する説明記事を自動編集する手法を提案した．評価実験の結果，提案手法によって Fujii の手法における「拡張性問題」と「冗長性問題」を解消しつつ，Fujii の手法に必要な人手の負荷を軽減することができた．今後の課題として，用語説明モデルに含まれていない新しい観点を自動的に発見する点がある．

謝辞

本研究の一部は，科学研究費補助金基盤研究（B）（課題番号 22300050）によって実施された．

参考文献

[1] Fadi Biadsy, Julia Hirschberg, and Elena Filatova. An unsupervised approach to biography production using Wikipedia. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, pp. 807–815, 2008.

[2] Corinna Cortes and Vladimir Vapnik. Support-vector networks. In *Machine Learning*, Vol. 20, pp. 273–297, 1995.

[3] Atsushi Fujii. Modeling Wikipedia articles to enhance encyclopedic search. In *Proceedings of LREC*,2010.

[4] Chin-Yew Lin ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the ACL-04 Workshop "Text Summarization Branches Out"*, pp. 74-81, 2004.

[5] Christina Sauper and Regina Barzilay. Automatically generating Wikipedia articles:A structure-aware approach. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics*, pp. 208–216, 2009.

[6] 白井諭, 大山芳史, 池原悟, 宮崎正弘, 横尾昭男. 日本語語彙大系について. 情報処理学会研究報告.IM,[情報メディア] 98(106), pp. 47–52, 1998.

[7] Shiren Ye and Tat-Seng Chua and Jie Lu. Summarizing Definition from Wikipedia In *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, pp. 199–207, 2009.