多面的な用語説明を生成するためのテキスト分類手法

三條場 旭彦
筑波大学図書館情報専門学群

藤井 敦
筑波大学大学院図書館情報メディア研究科

1 はじめに

科学技術や文化の急速な発展によって，様々な用語について調べる機会が増えている．こうした調べ物を支援するツールには，事典や検索エンジンがある．事典の長所は，人手で情報が統制されており質が高い点にある．しかし，短所として，未収録語が調べられないため情報の量が少ない．検索エンジンの長所は，情報の量が多く，新語に対応できる点にある．しかし，短所として，検索された情報は統制されていないため質が低い．著者らは，両者の長所を統合することで，Web 上の雑多な情報から説明情報を抽出し，様々な用語に対する説明を事典のように構築する研究を行っている [3, 4, 5].

本研究は，ある用語に関する多面的な説明情報を生成するために，Web 上の雑多な情報を「説明の観点」に基づいて分類する手法を提案する．用語説明の観点は，動物名では「分布」や「形態」など,「病名」では「診断」や「治療」などと用語の種類によって異なる．そこで，人手で作成されたフリー百科事典 Wikipedia から用語の種類ごとに観点を抽出し，用語説明のモデルを生成する．さらに生成したモデルを使用して，Web 上の雑多な情報を観点に分類する．

2 先行研究との比較

事典検索システム Cyclone[1][3, 5] に実装されている要約手法 [4] は，観点に基づいて複数の段落から代表文を抽出する．しかし，要約できる用語の種類は,「情報処理用語」のみに限定される．観点の設定と代表文を抽出する規則を人手で生成するため，実装のコストが高いことが原因である．本研究は，観点の設定と観点への分類を自動化することによって，多様な用語に対応する点が異なる．

Blair-Goldensohn ら [2] は，評判情報を観点に基づいて要約した．例えば，レストランに関する評判では,「サービス」や「価格」などが観点である．観点は，要約の対象である評判情報そのものから抽出される．それに対して，本研究は，外部情報である Wikipedia から観点抽出し，分析対象のテキストからは得ることができない観点に対応する点が異なる．

Biadsy ら [1] は，人物情報の要約に Wikipedia を使用した．Wikipedia から人物情報に関する記事を収集し，人物の説明に使われる文のモデルを学習した．しかし，Biadsy らは観点を用いておらず，また，対象は人物情報のみに限定される．本研究は，観点を用いて分類を行い，人物情報以外の多様な用語に対応する点が異なる．

[1] http://cyclone.slis.tsukuba.ac.jp/

3 用語説明の分類手法

3.1 概要

本研究で提案するテキスト分類手法の概要を図 1 に示す．図 1 は事前に行うオフライン処理と分類対象のテキストが与えられたときに行うオンライン処理で構成されている．

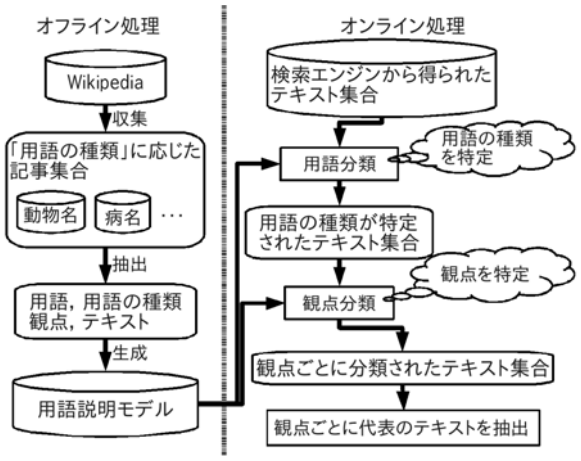

図 1: 本研究で提案するテキスト分類手法

図 1 のオフライン処理について概要を説明する．まず，Wikipeida から「動物名」や「病名」などの「用語の種類」に応じて記事集合を収集する．次に，収集した記事を観点ごとに分割し,「用語」,「用語の種類」,「観点」,「説明テキスト」の 4 つ組を抽出する．最後に，抽出した組から用語説明モデルを生成する．

次に，図 1 のオンライン処理について概要を説明する．まず,「テナガザル」などの用語をクエリとして Web を検索し，テキストを収集する．ここでは，Google や Yahoo!などの検索エンジンを使用する．次に，収集したテキスト集合中の各テキストを「テナガザル」の説明に関する観点に基づいて分類する．

ここで，用語の種類によって説明に必要な観点が異なるため,「用語分類」と「観点分類」の 2 段階で分類を実行する．どちらの分類もサポートベクターマシーン

（SVM）を使用し，One vs Rest 法を用いて多値分類に拡張した．

「用語分類」は，SVM のスコアが最も高い用語の種類にテキストを分類する．「観点分類」は，分類された用語の種類に対応する観点候補のいずれかにテキストを分類する．例えば，「動物名」では，「分布」や「形態」など，「スポーツ名」では「ルール」や「用具」などが観点の候補である．さらに，観点ごとに SVM のスコアが最も高いテキストをその観点の代表テキストとする．

3.2 用語の種類に応じた記事の収集

記事の収集では，Wikipedia の記事に付与されている「カテゴリ」を使用する．例えば，「ヒミズ」（モグラの一種）の記事には，「日本の哺乳類」や「食虫目」などのカテゴリが付与されている．Wikipedia のカテゴリは図 2 のように階層化されている．この階層関係を利用して，収集する用語の種類に対応するカテゴリを決定した．

Wikipedia に付与されているカテゴリの単位と，一般的な用語の種類が必ず一致するとは限らない．そのため，収集する用語の種類に対応するカテゴリを目視で特定した．Wikipedia には，図 2 の「生物」や「映画」などのように 100 の最上位カテゴリがある．まず，最上位カテゴリの中から収集する用語の種類を含むカテゴリを選択する．最上位のカテゴリでは，収集する用語の種類以外の記事集合を含むカテゴリも存在する．そのため，下位のカテゴリをたどりカテゴリに属する見出し語を調べ，目的とする用語の種類に対応するカテゴリを収集する．

図 2 では，「動物名」の記事集合を収集するために，まず，最上位カテゴリ「生物」を選択した．次に，下位のカテゴリに属する記事が「動物名」に適切かどうかを人手で調べた結果，「○」が付いたカテゴリの「哺乳類」や「両生類」などは，「動物名」として適切だった．適切なカテゴリに属する記事集合を「動物名」の記事として全て収集した．「×」が付いたカテゴリの「生物材料」は「材料名」の記事集合であり，「脊椎動物」は動物における体の構造についての記事集合だった．「魚類」は「魚類名」に関する記事集合であり「哺乳類の画像」は，動物の画像だったため，「動物名」として不適切だった．

実験では，表 1 に示すように，「動物名」，「病名」，「スポーツ名」などの一般用語 10 種類と「数学用語」，「建築学用語」，「法学用語」などの学術用語 10 種類を使用した．

3.3 用語説明モデルの作成

Wikipedia から収集した記事を用いて用語説明のモデルを生成する．用語説明モデルは，用語の種類に対応する観点と観点への分類基準で構成される．

用語に応じた観点を設定するため，Wikipedia の「セクション」を利用する．例えば，「ヒミズ」の記事は「形態」，「生態」，「人間との関係」，「近縁種」，「関連項目」，「参考文献」のセクションによって説明されている．多くの用語で使われるセクションは，用語を説明する上でよく用いられる観点であるため，収集した記事から 50 件

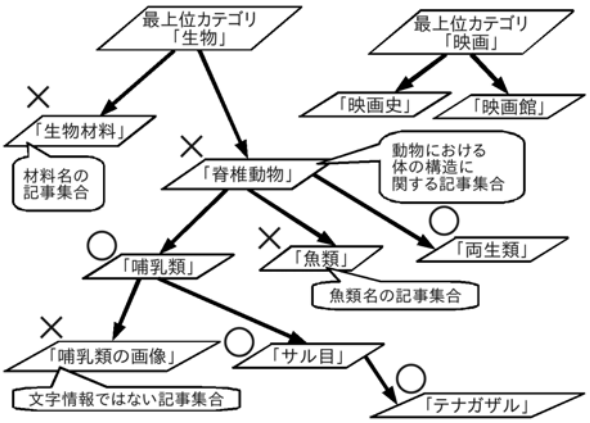

図 2: Wikipedia カテゴリの階層

以上の用語で使用されているセクションを観点とした．しかし，50 件を越える説明情報が得られない用語の種類は，使用頻度が高い上位 3 件のセクションを観点とした．ただし，以下のセクションは，説明の観点として不適切であると判断し，削除した．

> 概要，概説，概論，備考，脚注，出典，その他，外部リンク，関連，関連文献，関連図書，関連書籍，関連事項，関連記事，関連項目，関連リンク，関連カテゴリ，参考，参考書籍，参考文献，画像

ここで，「概要」，「概説」，「概論」は，様々な観点が混在するため，本研究で対象とする観点からは削除した．

次に，検索エンジンで得られたテキスト集合を SVM で適切な観点に分類するための学習データを作成する．ChaSen を使用し，観点ごとの説明情報を形態素解析する．観点分類では，名詞，動詞，形容詞，形容動詞，アルファベット，未知語を素性とした．用語分類では，観点分類で使用した素性に加え，用語自体も形態素解析し，用語に含まれる単語も素性とした．例えば，「キタゾウアザラシ」は，「キタ」，「ゾウ」，「アザラシ」に分割し，さらに，「キタ」と「アザラシ」にそれぞれ語頭と語末を表す位置情報を付けて素性とした．また，素性の値は全て 1 とした．

3.4 テキスト分類の実行例

「ハクビシン」を Yahoo!で検索して得られた上位 100 件のスニペットを分類した．作成した用語説明モデルには，「動物名」に対して，「生態」や「分布」など合計 7 種類の観点が対応していた．以下，観点ごとに分類された代表スニペットを示す．括弧内の数字は，左から順に，「該当する観点に正しく分類された件数」，「該当する観点に分類された件数」，「Yahoo!検索における順位」である．太文字の箇所が観点に対応する記述である．

- 生態（4/5，74 位）

  ハクビシンは雑食性であり、特に果実類を好むことから秩父地域の観光果樹園のぶどうを ... ハクビシ

ンは、食肉目ジャコウネコ科で東南アジア. を中心とした亜熱帯から熱帯地域に生息している。 ... これは、ハクビシン. が持つ ...

- 分布（6/55，4 位）

  白鼻心】、 masked palm civet、［学名：Paguma larvata］ 哺乳（ほにゅう）綱食肉目ジャコウネコ科の動物。**中国南部、チベット、インド、マレー半島、ボルネオ島、スマトラ島、 ...**

- 形態（5/7，16 位）

  ハクビシン・アライグマ・アナグマ・イタチ・イノシシ 害獣でお困りならお任せください! ... ハクビシン ジャコウネコ科 **体長/50〜70cm 体重/3〜5kg** ... 年々増え続けていますが、ハクビシンやアライグマなどは、 ...

- 人間との関係（15/23，22 位）

  ハクビシン. 飼育園館. 多摩動物公園・井の頭自然文化園. 生息地. 東南アジア、中国など。 **日本では四国、静岡など。最近は東京でも見られることがある。** ... ハクビシンがむかしから日本にいたのか、それとも帰化動物なのか、 ...

- 特徴（2/2，37 位）

  ハクビシンも実は国外から持ち込まれた外来種です。 ... ハクビシンはその名前の由来のとおり、**鼻筋にとおる白いくっきりとした線と、長くしなやかなしっぽが特徴的です。 ... 足の指の数を比べてみると、タヌキは 4 本ですが、ハクビシンでは 5 本あります。 ...**

- 分類（4/5，1 位）

  トップ **せきつい動物亜門 哺乳類 (哺乳綱) 食肉目 裂脚亜目 ジャコウネコ科 ハクビシン. ハクビシン (C)Kojo TANAKA. 分類. 哺乳類 食肉目裂脚亜目 ジャコウネコ科 ...** インドのカシミール地方、中国、台湾、マレー半島、 ...

- 歴史（1/3，84 位）

  ハクビシンは、明治維新以前に既に国内に持ち込まれていたと言われています。 江戸期の鳥獣戯画にハクビシン ... 本市では、ハクビシンが平成１４年８月、環境省による移入種対応方針に移入種として掲載されたことから、以来移入種として対応しています。...

この実行例では，本手法で分類された 7 件のスニペットを読めば,「ハクビシン」について多面的な説明を得られることが分かった．それに対して，Yahoo!検索の結果を上位から読んだ場合には，上記 7 つの観点に関する情報を得るために，33 件のスニペットを読む必要があった．

括弧内の数字を見ると,「分布」に分類されたスニペットの多くが誤りであった．これは，ハクビシンの目撃情報などに地名がかかれていると,「分布」に分類されやすいことが原因である．

Wikipedia における「ハクビシン」の記事には,「形態」,「歴史」,「分布」,「分類」に関するセクションはない．しかし，本手法は動物名に関する記事の集合から代表的なセクションを観点として抽出する．その結果，Wikipedia において個別の記事では欠落している観点を補うことができた．

分類したスニペットには，上記 7 種類以外の観点による説明として,「ハクビシン」をテーマにした「著作」や「名称の由来」があった．

4 評価実験

4.1 実験方法

本研究の目的は，Web 上の雑多なテキストを分類することである．しかし，今回は「整った」テキストである Wikipedia の記事を分類し，その正解率を評価した．収集した記事をセクションごとに分割したテキストを 1 件のデータとした．5 分割の交差検定によって用語分類と観点分類の正解率を計算し，結果を表 1 に示す．

4.2 用語分類の考察

表 1 より，用語分類の正解率は平均で 86.54%と高かった．特に，記事数が多い上位 5 種類の用語に対する正解率は約 90%もしくは 90%を越えた．

記事数が少ない用語の種類に対する正解率は低かった．記事数が少ない下位 3 種類の「獣医学用語」,「地質学用語」,「物性物理学用語」は正解率が 0%だった.「物性物理学用語」では，誤った記事数 23 件のうち，7 件が「化学用語」に分類された.「物理化学」と「化学用語」の両方を「物性物理学用語」として収集したためである．

「虫名」は,「数学用語」や「料理名」のように記事数が近い用語の種類に比べて正解率が低かった．誤った記事 63 件のうち 48 件は「動物名」に分類された.「虫名」の観点である「生態」,「特徴」,「分類」は全て「動物名」の観点と同じであり，説明の内容が似ていることが分類誤りの原因だった．

4.3 観点分類の考察

観点分類の正解率は平均で 79.66%だった．用語分類と同様に，記事数が多いと正解率が高かった.「動物名」は，正解率が 91.04%と高かった．表 1 で観点分類の正解率が最も高かった「企業名」の内訳を表 2 に示す.「企業名」は，観点ごとに説明の内容が本質的に異なるため，分類の正解率が高かった．

記事数が多いにもかかわらず観点分類の正解率が低かった「人名」の誤りについて考察する．表 3 に「人名」の内訳を示す.「経歴」と「略歴」はどちらも人物の生い立ちについて書かれているにもかかわらず，記事の著者によってセクション名が異なった．その結果,「経歴」と「略歴」は相互に誤分類が多かった．今後は，セクション名の表層だけにとらわれず，内容に基づいて観点を設定する必要がある．

表 1: 用語の種類における記事数と観点数と分類の正解率

用語の種類	記事数	観点数	用語分類		観点分類	
			正解記事数	正解率	正解記事数	正解率
動物名	1317	7	1201	91.19	1199	91.04
映画名	1169	5	1154	98.72	938	80.24
病名	878	8	853	97.15	674	76.77
企業名	618	3	556	89.97	602	97.41
人名	500	4	454	90.80	302	60.40
植物名	276	3	212	76.81	240	86.96
数学用語	228	3	198	86.84	140	61.40
虫名	203	3	140	68.97	171	84.24
化学用語	201	3	156	77.61	169	84.08
料理名	190	3	137	72.11	146	76.84
情報工学用語	103	3	61	59.22	61	59.22
魚類名	96	3	44	45.83	62	64.58
スポーツ名	86	3	66	76.74	65	75.58
法学用語	56	3	40	71.43	30	53.57
建築学用語	55	3	17	30.91	30	54.55
電気工学用語	49	3	7	14.29	18	36.73
天文学用語	38	3	21	55.26	20	52.63
獣医学用語	34	3	0	0	15	44.12
地質学用語	24	3	0	0	7	29.17
物性物理学用語	23	3	0	0	5	21.74
平均	307.2	3.6	265.85	86.54	244.7	79.66

表 2: 「企業名」に関する観点分類の内訳

観点	記事数	分類された観点			正解率
		沿革	事業所	主な商品	
沿革	469	437	0	2	99.54%
事業所	116	2	112	2	96.55%
主な商品	63	8	2	53	84.13%

表 3: 「人名」に関する観点分類の内訳

観点	記事数	分類された観点				正解率
		経歴	略歴	著書	人物	
経歴	219	142	48	28	1	64.84%
略歴	133	69	55	7	2	41.35%
著書	83	0	0	83	0	100%
人物	65	20	9	23	13	20%

5 おわりに

Wikipedia から用語説明モデルを生成し，Web 上のテキストを説明の観点に基づいて分類する手法を提案した．現在は，テキストを分割せずに分類しているため，複数の観点が混在し，重複する場合がある．今後は，文などの単位で分割して分類することで，冗長性の少ない説明を生成する必要がある．

謝辞

本研究の一部は，文部科学省科研費特定領域研究「情報爆発時代に向けた新しい IT 基盤技術の研究」（課題番号： 19024007）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] Sasha Blair-Goldensohn, Kerry Hannan, Ryan McDonald, Tyier Neylon, George A.Reis, and Jeff Reynar. Builing a sentiment summarizer for local service reviews. *In WWW2008 Workshop on NLP challenges in the information Explosion Era*, 2008.

[3] Atsushi Fujii. Producing an encyclopedic dictionary using patent documents. *In Proceedings of the 6th International Conference on Language Resources and Evaluation*, 2008.

[4] Atsushi Fujii and Tetsuya Ishikawa. Summarizing encyclopedic term descriptions on the web. *In Proceedings of the 20th International Conference on Computational Linguistics*, pp. 645–651, 2004.

[5] Atsushi Fujii, Katunobu Itou, and Tetsuya Ishikawa. Cyclone: An encyclopedic web search site. *In Special Interest Tracks & Posters of The 14th International World Wide Web Conference*, pp. 1184–1185, 2005.