様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　１９　日現在

機関番号：　１２６０８
研究種目：　基盤研究(B)（一般）
研究期間：　2015～2017
課題番号：　１５Ｈ０２７４７

研究課題名（和文）ウィキペディアのモデル化に基づく体系的・連想的な解説記事の自動生成

研究課題名（英文）Modeling Wikipedia to Automatically Generating Coherent and Associative Expository Articles

研究代表者

藤井　敦（Fujii, Atsushi）

東京工業大学・情報理工学院・准教授

研究者番号：３０３０２４３３

交付決定額（研究期間全体）：（直接経費）　　12,500,000円

研究成果の概要（和文）：本研究の目的は，ある用語の解説記事をウェブ上の情報から自動生成することであり，二種類の対照的なモデルを探求する．体系的解説モデルは，動物や病気といった用語のカテゴリに応じて解説する観点を切り替えて記事を生成する．連想的解説モデルは，既存語との共通点や相違点に基づいて直感に訴える解説を生成する．両モデルを使い分ける方略の仕組みを通して，自然言語理解の本質に迫ることを目指す．

研究成果の概要（英文）：The purpose of this research is intended to automatically generate expository text for an input term, for which two types of contrastive models are explored. First, because the viewpoints used for explanation can be determined depending on the type of the term in question. Second, the target term is compared with existing terms to generate intuitively understandable explanation. The nature of natural language understanding is explored through the strategic use of these models.

研究分野：　自然言語処理

キーワード：　自然言語処理　情報検索　人工知能　百科事典　ウィキペディア　情報の組織化　連想

１．研究開始当初の背景

科学技術や文化の発展によって新しい用語が次々と生み出され，検索の重要性が増している．検索エンジンと事典は，それぞれ情報の量と質に優れており，私的な調べ物から公的機関による調査活動に至るまで幅広い領域で必需品となっている．

研究代表者は，検索エンジンと事典の長所を統合するために，ウェブ情報や特許情報から事典情報を自動構築するための研究開発に二十年近く取り組んできた．

折しもウィキペディアが登場し，情報の量と質を備えたツールとして期待された．しかし，人類の歴史が続く限り，用語の誕生や用法の変化を止めることはおよそ不可能である．森羅万象を網羅することは困難である．

我々は，ウィキペディアを計算機上でモデル化し，当モデルに基づいて新語の解説記事を自動生成することが根本的な解決に近いのではないか？ という発想に至り，科研費基盤Ｂ(H22～H24)で基盤技術を確立した．

具体的には，用語の種類に応じて必要な観点を自動的に選択して「体系的な解説」を指向した．例えば，「デング熱」は病名なので，症状，予防，治療等の観点から解説するのに対して，映画のタイトルであれば，監督，出演者，あらすじ等が観点となる．「ハブ」のように複数の意味で使われる用語は各々の意味に対応した観点に基づいて解説する．

他方において，用語の解説とはただ単に正確に解説すれば十分という訳ではない．聞き手の事前知識を確認し，相手が理解できる語句や専門知識の水準を把握する必要がある．また，始めのうちは末節を切り捨てることで相手の直感に訴えて大筋を把握させて，その後に体系的な解説によって不備を正すような方略もある．

直感に訴える一つの例として，解説対象の用語を他の用語と対比させて直感的な理解を支援する「連想的」な解説がある．「ハクビシンは，ネコのような体型」，「iPS 細胞は生命のタイムマシン」，「選択ソートはバブルソートよりも値の交換回数が少ないので高速」のように下線付きの用語を「基準語」として，それらとの共通点や相違点に基づいて対象語を解説する．また，基準語の解説を引用することで，複数の解説記事で繰り返される記述の量が体系的な解説よりも少ない．両モデルの得失を以下にまとめる．

【体系的解説】
利点：各語について深く正確に解説できる．
欠点：用語間で解説の内容に重複がある．
　　細部に固執すると本質を見失う．

【連想的解説】
利点：複数の用語を広く浅く概観できる．
欠点：良い基準語の存在に強く依存する．
　　正確性に欠ける．

２．研究の目的

ウィキペディアのように人手で統制された事典情報が大規模化した一方で，それを凌駕する大量の「統制されていない情報」がウェブには存在する．本研究の目的は，ウィキペディアをモデル化して未統制のテキスト群を統制し，解説記事を自動生成する点にある．ウィキペディアの記事集合から「人間が用語を解説する仕組み」のモデルを作ってテキスト群の自動編集に応用する．

様々な観点から系統立てて詳説する「体系的解説」と既知語との関連付けや対比によって直感的に概説する「連想的解説」のモデルを構築して使い分けを可能にする．その結果，「自分自身が物事の本質を体系的に捉えてから，それを噛み砕いて直感的に説明する」あるいは反対に「まず直感的に概説して理解の素地を作ってから，体系的に詳説する」といった解説モデルの探求を通して自然言語理解の本質に迫る．

ウィキペディアから有用な言語資源を自動構築する研究活動は国内外に数多く存在する．対して本研究の特色は，「ウィキペディア記事が如何にして編集されたのか」という原理をモデル化し，新たな言語資源を生成するための触媒として利用する点にある．

その触媒によって異種のテキスト群に化学反応を起こし，解説記事としての生命を吹き込む，いわば，個別の情報にはない新たな価値を創造するための取り組みである．

現時点において，本研究の手法が上手く機能するのは，病気，人物，映画のように解説するときに押さえるべき要点（観点）が固定化されている用語であり，これらのカテゴリに関する大量の用語に対する解説文書から当該カテゴリに関する標準的な観点の集合や階層関係を機械学習できる場合である．

中長期的には，本研究の成果をウィキペディアのコミュニティーに還元して，ウィキペディアの品質が向上することに伴って本研究のモデルがより一段と改善される好循環が実現するかもしれない．

３．研究の方法

１．および２．で説明した体系的解説と連想的解説のうち，前者は過去の科研費で確立した基盤技術の拡張に焦点を当てた．対して，後者の連想的解説は，複数の異なるモデルの構築や実装の手段について試行錯誤した．

４．研究成果

(1)体系的解説モデルの概要

体系的解説モデルの構築では，まず，ウィキペディアから用語カテゴリごとに記事集合を集めて分類器を機械学習する．さらに各カテゴリに頻出するセクション名を観点として観点ごとの記述から分類器を機械学習

する．「キーウィ」に関するテキスト群を与えると，それらを「動物」や「植物」の用語カテゴリに分類して対応する観点のいずれかに細分類する．最後に，各観点から代表性が高いテキストを選択して，それらを連結することで解説記事とする．

(2)非自然言語に基づく解説への対処

ウィキペディアは，専門用語に関する記事の件数および内容の密度において依然として不足がある．具体的には，解説記事のセクション構造が単調で解説の観点が不明確になりやすい．特に数式のような非自然言語テキストによる簡素な解説になりやすい傾向にある．その結果，本研究で提案した体系的解説のモデル構築において，用語カテゴリと観点の対をテキスト情報に高精度で対応付けることが困難であった．

数式の構造解析に基づく文書分類に関する基礎研究を進展させたことによって，上述のような状況においても数式情報を手掛かりとして，用語カテゴリや観点の分類精度を向上させることや，当該数式に関連する自然言語テキストをウェブから検索することで情報の不足を補うための新たな切り口を得ることができた．

(3)手順に関する解説モデル

ある目的を達成するための手段をいくつかの単純な動作のまとまりとして認識することがある．道案内をするときには，広域地図で大雑把に説明して，相手に土地勘がない場所について詳細地図で説明する方が全てを克明に説明するよりも効率的である．コンピュータのプログラムは頻繁に使用する典型的な一連の操作系列をモジュールとして独立させ，さらに目的や機能を表す名前を付けることで，プログラムの可読性や生産性を向上させることに成功している．

手順の解説を指向して，料理レシピの構造解析に関する基礎研究を行った．調理手順をまとまりのある構成要素に分割して，中心的な動作を簡潔に表す見出しを生成することによって，手順の骨格であるアウトラインを抽出する手法を提案した．

ただし，「切る」と「煮る」という二つの動作が継続性において異なることを認識できないと，依然として継続しているにもかかわらず，「煮る」を含んだ見出しの候補が削除されてしまう．そこで，レシピテキストに対する動作の継続性解析を提案することで同問題を解決した．

(4)連想的解説モデルに関する基礎研究

同種の対象を複数並べて，他との比較によって得失について解説する手法の実現を試みた．解説の対象は種々の観点から数値的に評価されており，自然言語テキストによるレビューが利用可能であることを前提とする．数値とテキストによる評価をそれぞれ「体系的解説」および「連想的解説」と見なして，一方から他方への変換について探求した．一つの対象を種々の観点で評価する場合は，特に優れた観点と特に劣った観点に関するレビューは注目に値する．しかし，ある観点について賞賛もしくは批判する記述があったときに，「ある対象が持つ特徴の中で最も優れている（もしくは最も劣っている）から」なのか「他の対象と比べて優れている（もしくは劣っている）から」なのかを特定して書き手の意図を明確にすることが好ましい．特に，レビューの総合点として最低点が付けられている場合は，非常に不愉快な思いをしたために，その事故が起因する項目に最低点を付けて，さらにその最低点を強調するために，他の項目は一律に高い点数を付ける事例が複数存在した．こうした記述は，レビュー（評価）というよりも「ネガティブ投票」と呼ぶべきであり，通常のレビューとは区別する方が合理的である．本研究は，映画や宿泊施設に対するレビュー集合の分析を通して，上述した種々の違いを識別するための素性を提案した．

(5)実体験の解説としての評判分析

評判情報の分析に関する既存の研究は，レビュー等の意見性を有するテキスト情報から，意見の最小単位として，「評価の対象(T)，評価の属性(A)，評価の極性(E)」という三つ組を抽出する．例えば，宿泊施設の利用客によるレビューからは，「○×ホテル(T)の立地(A)はとても便利だった(E)」のような意見が抽出される．しかし，こうした評価は往々にして種々の前提条件に依存する．

先ほどの例では「小さい子供連れにとっては」という条件が付いており，それゆえ出張目的の利用客には当てはまらない可能性がある．こうした差異の抽出に取り組んだ研究は国内外に存在しない．すなわち，既存の評判分析手法によって得られる「○×ホテル」に関する総合評価には一定の割合で誤りが含まれることになる．

実データに基づく我々の調査では，この誤差は最大 30%であり決して無視できる数値ではない．本研究は，単語，構文，意味に基づく自然言語解析結果と教師あり機械学習手法を併用して，実用的な精度で評価に関する条件の抽出に成功している．なお，当該成果は著名な国際会議に採択されるとともに，筆頭著者が情報処理学会の山下記念研究賞を受賞した．

(6)視線追跡に基づく可視化手法の評価手法

生成された解説テキストやそこから抽出された語句に基づいて構成されたキーワードマップの品質をユーザの閲覧行動に基づいて評価することを目的として，視線追跡の機能を持つディスプレイを用いた基礎研究を行い，研究の継続に値する有望な評価手法を特定した．

５．主な発表論文等

〔雑誌論文〕（計４件）

① Sidik Soleman, Atsushi Fujii. Plagiarism Detection based on Citing Sentences, **査読有**, 2017. 21st International Conference on Theory and Practice of Digital Libraries.

② Sidik Soleman, Atsushi Fujii. Toward plagiarism detection using citation networks, **査読有**, 2017. 2017 Twelfth International Conference on Digital Information Management.
DOI: 10.1109/ICDIM.2017.8244682

③ Tokinori Suzuki, Atsushi Fujii. Mathematical Document Categorization with Structure of Mathematical Expressions. **査読有**, pp.119-128, 2017.
Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries.
DOI: 10.1109/JCDL.2017.7991566

④ Yuki Nakayama, Atsushi Fujii. Extracting Condition-Opinion Relations Toward Fine-grained Opinion Mining, **査読有**, pp.622-631, 2015.
Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing.
DOI: 10.18653/v1/D15-1074

〔学会発表〕（計 14 件）

① **笛田 剛, 藤井 敦. 複数ユーザの視線情報モデルに基づくキーワードマップの自動評価手法, 情報処理学会情報基礎とアクセス技術研究会,** 2018-IFAT-130, 2018. Mar.

② **伊藤 歩未, 藤井 敦. ビジネス文書を対象とした自動分類手法, 情報処理学会情報基礎とアクセス技術研究会,** 2018-IFAT-130, 2018. Mar.

③ **中山 祐輝, 藤井 敦. 宿泊者レビューに対するホテルの返信から何が見えるか？, 情報処理学会情報基礎とアクセス技術研究会,** 2018-IFAT-129, 2018. Feb.

④ Sidik Soleman, Atsushi Fujii, A Method for Plagiarism Detection over Academic Citation Networks, **情報処理学会情報基礎とアクセス技術研究会,** 2018-IFAT-129, 2018. Feb.

⑤ **中山 祐輝, 藤井 敦. ラウンドトリップディスカッション：宿泊客のレビューとホテル側の返信から評価における論点を探る,** IDR **ユーザフォーラム** 2017. Dec. **奨励賞受賞**

⑥ **中山 祐輝, 藤井 敦. 宿泊施設の顧客レビューに対する応対文書の自動分類手法, 情報処理学会情報基礎とアクセス技術研究会,** 2017-IFAT-126, 2017. Mar.

⑦ Sidik Soleman, Atsushi Fujii. Modeling Plagiarism in Citation Networks for Academic Publications, **情報処理学会情報基礎とアクセス技術研究会,** 2017-IFAT-126, 2017. Mar.

⑧ Tokinori Suzuki, Atsushi Fujii. A Structure-based Method for Mathematical Document Classification, **情報処理学会情報基礎とアクセス技術研究会,** 2017-IFAT-125, 2017. Mar.

⑨ **二階堂 学, 藤井 敦. 確率的言語モデルに基づく効率的閲読のモデル化, 情報処理学会情報基礎とアクセス技術研究会,** 2017-IFAT-124, 2017. Feb.

⑩ Soleman Sidik, Atsushi Fujii. Plagiarism Detection Based on Citation Contexts, **情報処理学会情報基礎とアクセス技術研究会,** 2017-IFAT-124, 2017. Feb.

⑪ **大沼 俊輔, 藤井 敦. 情報検索における視線情報を用いた適合性推定, 情報処理学会情報基礎とアクセス技術研究会,** 2016-IFAT-122, 2016. Mar.

⑫ **杉本 憲哉, 藤井 敦. 動作の継続性に着目した料理レシピのアウトライン自動生成, 情報処理学会情報基礎とアクセス技術研究会**, 2016-IFAT-122, 2016. Mar.

⑬ Xinliang Zhao, Atsushi Fujii. Sentence Selection for Language-gap Reduction in Cross-lingual Sentiment Classification, **情報処理学会情報基礎とアクセス技術研究会**, 2016-IFAT-121, 2016. Dec.

⑭ Sidik Soleman, Atsushi Fujii. Authorship Segmentation for Retrieving Source Documents in Plagiarism Detection, **情報処理学会情報基礎とアクセス技術研究会**, 2016-IFAT-120, 2016. Jan.

６．研究組織

(1)研究代表者

藤井　敦（FUJII, Atsushi）
東京工業大学・情報理工学院・准教授
研究者番号：３０３０２４３３

(2)研究分担者

徳永　健伸（TOKUNAGA, Takenobu）
東京工業大学・情報理工学院・教授
研究者番号：２０１９７８７５