E-053

Wikipedia から抽出した語彙関係リソースの小論文自動評価タスクへの適用

Extraction of lexical relation resource from Wikipedia for auto-evaluation task of short essay

藤田 央† 藤田 彬† 田村直良‡

Hiroshi Fujita† Akira Fujita† Naoyoshi Tamura‡

１． はじめに

近年，学習者により書かれた文章を教育的な目的で自動評価する技術の需要が高まっている．大学入試や就職試験等の大規模な学力試験において課される小論文試験の採点や,e-learning 等の電子的な学習システムにおいて学習者の能力を測るために出題される記述式テストの採点が例として挙げられる．このような，多数の文章を同一基準の下で迅速に評価するタスクにおいては，評価者間で評価結果に差異が生じることがある．要因として，「個々の評価者が着目する言語的要素」と「評点決定に寄与する各要素の配分（重み）」の違いが挙げられる．我々は，これらの違いを機械学習により定量化して，文章を教育的観点から自動評価する手法を提案している[8]．この手法では，語彙的結束性等に関する素性を用いて文章のまとまり(局所的一貫性と結束性)の妥当性を評価する．しかし，小論文を評価対象とする場合，扱われる話題には時事的なものが多く，新しい語彙を即時的に反映しない語彙リソース(シソーラスやオントロジー辞書等の語彙資源)では捉えきれない語彙が存在する．

そこで,本研究ではWikipediaに記載された語彙を抽出して，小論文の自動評価に用いる辞書を拡張することを目的とする．Wikipediaは，技術用語や専門用語など，様々な記事が記載されている大規模なインターネット百科事典である．時事的な語彙に関する記事についても，複数の編集者により逐次編集がなされている．

提案する手法では,Wikipediaに記載された記事の見出し語や定義文（見出し語の簡単な説明がなされる冒頭部分）から語彙の抽出を行なう．抽出した語彙は，文章中で名詞として捉えられなかった語彙が捉えられるように,形態素解析器の辞書に追加する．

実験では，小論文を対象に小論文評価における文章のまとまりの評価の精度が向上することを示す．

２． 上位下位関係(hyponymy relations)を持つ名詞群

２．１ 小論文評価に必要な意味情報

我々の手法では文章のまとまりの自動評価の精度向上を目的とする．この際に名詞と名詞の間に上位下位関係があるかないかを評価の一部に用いる．しかし，高校生の小論文では時事的な話題を扱うことが多い．そのために時事的でありかつ相互に上位下位関係のある名詞の組み合わせを任意の知識源から抽出する必要がある．

本研究では，時事的でない一般名詞間の関係についてはEDR電子化辞書を参照する．また，「上位語：メディア，下位語：インターネット」 のような時事的な名詞間の上位下位関係や「上位語：女性，下位語：女性専用車両」 のような時事的な名詞と一般名詞の上位下位関係についてはWikipediaから抽出する上位下位関係を参照する．

２．２ EDR電子化辞書とWikipediaの比較

〈名詞の数〉

EDR(v4.0)の日本語単語辞書に記載されている名詞は約 28 万語あり,Wikipedia の記事の見出し語は約 146 万語(2011 年 4 月 7 日時点)ある.表層が共通する単語の総数は約 6.7 万語である.

〈登録されている単語の種類〉

EDRは一般名詞を,Wikipediaは専門用語や固有名詞を扱っている．また，Wikipedia には一部であるが一般名詞も存在する．しかしEDRで扱われている一般名詞に比べると数が少ない.

〈上位下位関係〉

EDR の概念辞書に記載されている上位下位関係は約8万ペア（全体で43万ペアの上位下位関係があったが概念識別子に対応する単語が上位語と下位語ともにあるペアのみ利用した)，Wikipediaから抽出した上位下位関係は約 173 万ペア，両者の共通ペアは 29 ペアである．

３． Wikipediaの名詞と上位下位関係の抽出の方法

３．１ 名詞の抽出

Wikipediaからの名詞の抽出方法を以下に示す.

１．Wikipedia のダンプデータ [1] から「page.txt」を使用する.

２．Wikipediaにある各記事の見出し語を抽出する.

３.記号や画像ファイルなど不要な見出し語は除く.

（例：!，#，¥，samsung LE26R41BD.jpg）

４．定義文から後述の方法で上位下位関係にある上位語と下位語を抽出する．これらの内訳は下位語：

† 横浜国立大学大学院環境情報学府
Graduate School of Environment and Information Sciences, Yokohama National University
†† 横浜国立大学大学院環境情報研究院
Graduate School of Environment and Information Sciences, Yokohama National University

[1] http://dumps.wikimedia.org/jawiki/latest/の URL にある jawiki-latest-pages-articles.xml.bz2 をダウンロードするとpage.txt と text.txt が得られる.

約 61.8 万語，上位語：約 23.4 万語　見出し語：約 18.3 万語(うち約 50 万語は下位語と重複する) 計約 133.5 万語である．Wikipedia から抽出した名詞を「Wikipedia の名詞群」と呼ぶことにする．

３．２　上位下位関係の抽出

Wikipedia から上位下位関係を抽出する先行研究には隅田らの研究がある[5]．Wikipedia の記事の定義文には各記事の見出し語の上位語が記述されている．例えば見出し語「ウメ」の記事の定義文は下記の通りである．

例：「*ウメ（梅，学名：Prunus mume）は，バラ科サクラ属の落葉高木，またはその果実のこと*」

見出し語に対する上位語を以下のように抽出した．

１．Wikipedia のダンプデータから「text.txt」をダウンロードする．

２．text.txt の各記事の最初の一文が「ＡはＢ。」「ＡとはＢ。」の２つのパタンいずれかに合う文章を取り出す．このときＡを下位語としＢを上位部とする．上位部とは上位語を含む文字列である．

３．上記の２つのパタンに合った文章の上位部から[[と]][2]で囲まれた部分を上位語とする．

例えば

「*あだち充は[[日本]]の[[漫画家]]。*」の場合は

上位語：日本，下位語：あだち充

上位語：漫画家，下位語：あだち充

のように抽出される．

また，「*言語(げんご)とは，[[コミュニケーション]]のための[[記号]]の体系である。*」の場合は

上位語：コミュニケーション，下位語：言語

上位語：記号，下位語：言語

のように抽出される．

このようにして Wikipedia から上位下位関係を約 183.5 万ペア抽出する．

３．３　上位下位関係の拡張

前節で抽出された上位下位関係がある名詞ペアのセットは相互に関係を持たないことを説明した．これは，

上位語：人，下位語：作曲家

上位語：作曲家，下位語：バッハ

の２つの上位下位関係のセットがあった場合，２つの上位下位関係は推移律(この場合作曲家)が入っていても相互に関係がないことを示している．しかし，実際は「作曲家」を間にとって

上位語：人，下位語：バッハ

という上位下位関係がある．そこで，このように２つの上位下位関係のセットにおいて以上のような関係があるような場合には関係があるとする．

[2] Wikipedia における他記事へのリンクを表す書式

４．　複合名詞の扱い

複合名詞は語基に分ける(以下,語基に分けることを語基分けと呼ぶことにする)．このとき語基は「名詞・接尾・一般」，「名詞・接尾・形容動詞語幹」，「名詞・形容動詞語幹」，「名詞・接尾・サ変接続」，「名詞・接尾・特殊」と品詞分類される形態素とする．一つの複合名詞から考えられるすべての語基の分け方のパタンを扱う．

例えば「外国人旅行者」を語基に分ける場合，この複合名詞には，「外国人旅行者」「外国人旅行」「人旅行者」「外国人」「人旅行」「旅行者」「外国」「人」「旅行」「者」のような名詞が含まれる．このとき，明らかに名詞でない語基(「人旅行者」「人旅行」「外国人旅行」)は人手で除外する．

５．　実験・考察

５．１　設定

実験システムの概要を説明する．まず，文章(小論文)を MeCab[3] で形態素解析をして，CaboCha[4] で係り受け解析を行なう．次に，文節の主辞である「名詞_サ変接続」「名詞_一般」「名詞_固有名詞」「名詞_形容動詞語幹」を集め，EDR と Wikipedia の上位下位関係と照合する．そして，上位下位関係がある名詞数をカウントする．

教師データには高校生小論文を電子化したものを用いる．これらの小論文は，(論題 A)「小学校の授業における，英語の早期教育は必要であるか否かに対して意見を述べよ」，(論題 B)「グラフと説明文を読み，日本人の子育ての態度に関してどのような特色が読み取れるかに関して述べよ」という 2 種類の論題に沿って書かれている．また，400 字以内と 800 字以内の 2 種類の字数制限が存在する．事例は合計で 584 事例あり，論題 A を 400 字以内で記述するものが 153 事例,論題 B を 400 字以内で記述するものが 140 事例，論題 A を 800 字以内で記述するものが 147 事例，論題 B を 800 字以内で記述するものが 144 事例という内訳である．これらの 584 事例に対して 4 人の評価者が一貫性の観点においてつけた 5 段階の評点を各教師データのラベルとする．各評価者の小論文に対する評価の平均とその標準偏差は表２にある．

実験１．上位下位関係がある名詞の数

【目的】

「小論文から抽出した名詞の数」と「抽出した名詞の中で上位下位関係がある名詞の数」を検討する．

【手法】

１．小論文中にある名詞を表１の３つの条件(α，β，γ)で抽出する．

[3] http://mecab.sourceforge.net/

[4] http://sourceforge.net/projects/cabocha/

α：提案手法を適用しない
β：Wikipedia から抽出した名詞群を加える
γ：βの条件でさらに複合名詞を語基に分ける

２．抽出された名詞の中で上位下位関係がある名詞の数を調べる．

３．α，β，γの条件で小論文中にある名詞ペアを以下の４つに分類する．

a：EDR のみに上位下位関係があるペア
b：Wikipedia のみに上位下位関係があるペア
c：EDR および Wikipedia に上位下位関係があるペア
d：EDR，Wikipedia ともに上位下位関係がないペア

【結果】

α：全小論文中から 2397 タイプの名詞が抽出され，そのうち，他の名詞と関係(上位下位関係)がある名詞は 58 タイプある．

β：全小論文中から 2628 タイプの名詞が抽出され，そのうち，他の名詞と関係がある名詞は 661 タイプある．(EDR は 58 タイプ，Wikipedia は 603 タイプ)

γ：全小論文中から 2755 タイプの名詞が抽出され，そのうち,他の名詞と関係がある名詞は 1220 タイプある(EDR は 617 タイプ，Wikipedia は 603 タイプ)

図２は横軸に全小論文 584 編にある名詞数，縦軸に抽出する３つの方法をとり，全小論文中に存在する「全名詞タイプ数」と「上位下位関係のある名詞タイプ数」を表している．

【考察】

〈αとβの比較〉

MeCab のユーザ辞書に Wikipedia の名詞群を加えることで新たに抽出された名詞があった．これらの名詞の中には，小論文内で他の名詞と上位下位関係があるものがあった．

〈βとγの比較〉

複合名詞の語基分けをしない場合に比べて語基分けをする場合の方が，一般名詞として抽出されるものが多い．しかし，抽出される名詞数の増加幅はαとβの増加幅に比べて小さい．これは，高校生小論文中に Wikipedia に記載されている語彙が多くあることによるものと思われる．

〈上位下位関係がない名詞〉

かな混じりの名詞や誤字，もしくは EDR や Wikipedia の上位下位関係にエントリがなかったことが要因と思われる．

実験２．小論文評価システムの精度

【目的】

Wikipedia の名詞群を MeCab のユーザ辞書に追加した場合としない場合とで，小論文自動評価システムの文章のまとまりに関する評価精度を比較する．

【手法】

文献[7]で使われている文章のまとまりの素性群と,Barzilay らが提案した局所的な一貫性のモデル「entity grid モデル」[5]において扱われる構文役割の 4 種類(S:主語，O:目的語，X:その他，－:出現せず)に横野らが提案した結束性に寄与する要素(H:主題,R:述部要素)を加えたものを用いて，entity grid 上の下記の素性を計算する．

１．隣接文間でつながりのある箇所／全隣接数
２．最終文とつながりのある文／最終文を除く文数
３．冒頭文とつながりのある文／冒頭文を除く文数
４．(－:出現せず以外)要素の数／全要素数

このとき「４名の評価者がつけた評点」と「教師データを用いて構築した判定器の評点推定結果」の間の差について実験１のα，β，γの条件で検討する．判定器には SVR[6]を用いる．SVR による評点推定の評価指標は MAE(Mean Absolute Error)を用いる．MAE は値が 0 に近ければ近いほど両者の評価が一致していることを示している．

【結果】

αとβにおいて MAE は最大 0.084 下がった．
βとγにおいて MAE は最大 0.023 下がった．

【考察】

表３において Wikipedia を加えた場合と複合名詞の語基分けを行なった場合で MAE の値は前者がより 0 に近づいた．

６　おわりに

本研究では高校生の小論文自動評価システムの一貫性に関する精度向上について Wikipedia から抽出した上位下位関係を用いる手法について提案した．実験１では Wikipedia を追加することで小論文中にある名詞数および上位下位関係数を多く認識することができた．また，複合名詞を語基に分けることでさらに上位下位関係を認識することができた．

実験２では取得された上位下位関係は文章のまとまりの評価において MAE 値が最大 0.084 下がり精度向上した．この結果から Wikipedia の名詞群を MeCab のユーザ辞書に加えることで小論文評価における文章のまとまりの評価において精度向上することが示せた．今後の課題として

・扱う小論文を変えて実験をする．
・Wikipedia の上位下位関係は上位下位関係の羅列になっていたが今後は木構造にする必要がある．
・Wikipedia からの語彙の抽出をすすめるとともに，分類語彙表，日本語語彙体系など他の語彙リソースの活用を検討する．

謝辞

本研究については，公益財団法人博報児童教育振興会の児童教育実践事業についての研究助成事業，「学習指導要領に立脚した児童作文自動点検システムの実現」（助成番号：11-B-081，研究代表：藤田

[5]文を行，文章中の語句要素を列，文における語句要素の構文役割を成分とする行列を用いて，語句要素の分布パタンを表現するモデル．行列から構文役割の遷移確率と構文役割の出現確率を成分とするベクトルを導出し，局所的一貫性の評価等に用いる．

[6] http://svmlight.joachims.org/

彬）の援助を受けた．また，高校生の小論文答案をお貸しいただき,研究利用を認めて下さった揚華氏,宇佐美慧氏，東京工業大学大学院社会理工学研究科の前川眞一教授に感謝の意を表す.

参考文献

[1] M. A. K. Halliday, and Hasan R, ”Cohesion in English “, Longman, London(1976).
[2] Barzilay R and Lapata M,“ Modeling Local Coherence An Entity-based Approach”, Computational Linguistics, 34(1), pp.1-34(2008).
[3] Yigal Attali, Don Powers, ”A Developmental Writing Scale”, ETS Research Report(2008).
[4] 石岡恒憲，亀田雅之，”コンピュータによる小論文の自動採点システムjessの試作”,計算機統計学，Vol.16，No.1(2003).
[5] 隅田飛鳥,吉永直樹,島澤健太郎，“Wikipediaの階層構造を知識源とする上位下位関係の自動獲得“, 情報処理学会全国大会講演論文集 第70回平成20年(5)，pp.51-52(2008).
[6] 藤田彬，田村直良，”文章構造解析に基づく小論文の論理性についての自動採点”，第9回情報科学技術フォーラム(FIT2010)，RE-004，Vol.2，pp.41-44(2010).
[7]藤田彬，藤田央，田村直良，”多様な教育的観点を考慮した機械学習による日本語文章の評価と評価モデルの顕在化”，情報処理学会研究報告 NL-202(2011)
[8]M.A.K.Halliday,”An Introduction to Functional Grammer”,(1994).(邦訳：機能文法概説，山口登，筧壽雄 訳(2001)).
[9]田窪行則,西山佑司,三藤博,亀山恵,片桐恭弘,”談話と文脈”.(2004).
[10] 横野光，奥村学,“テキスト結束性を考慮したentity gridに基づく局所的一貫性モデル”，自然言語処理，Vol.17，No.1，pp.161-182(2010).
[11] 石岡恒憲,”小論文およびエッセイの自動評価採点における研究動向”,人工知能学会誌，Vol.23，No.1(2008).
[12] Attali Yigal, Burstein Jill,”Automated essay scoring with e-rater v.2”, Journal of Technology, Leaning and assessment(2006).

表１：文章中の名詞の３つの抽出方法

	MeCab ユーザ辞書への Wikipedia の追加	複合名詞の語基分け
α	×	×
β	○	×
γ	○	○

表２　各評価者の一貫性の評価における平均と標準偏差

評価者	平均	標準偏差
A	4.613	0.826
B	4.620	0.636
C	3.709	1.120
D	3.804	0.976

表３：３つの抽出方法ごとの文章のまとまりに関する自動評価の MAE

	評価者			
	A	B	C	D
α	1.120	1.089	1.903	1.259
β	1.112	1.029	1.819	1.240
γ	1.109	1.013	1.786	1.233

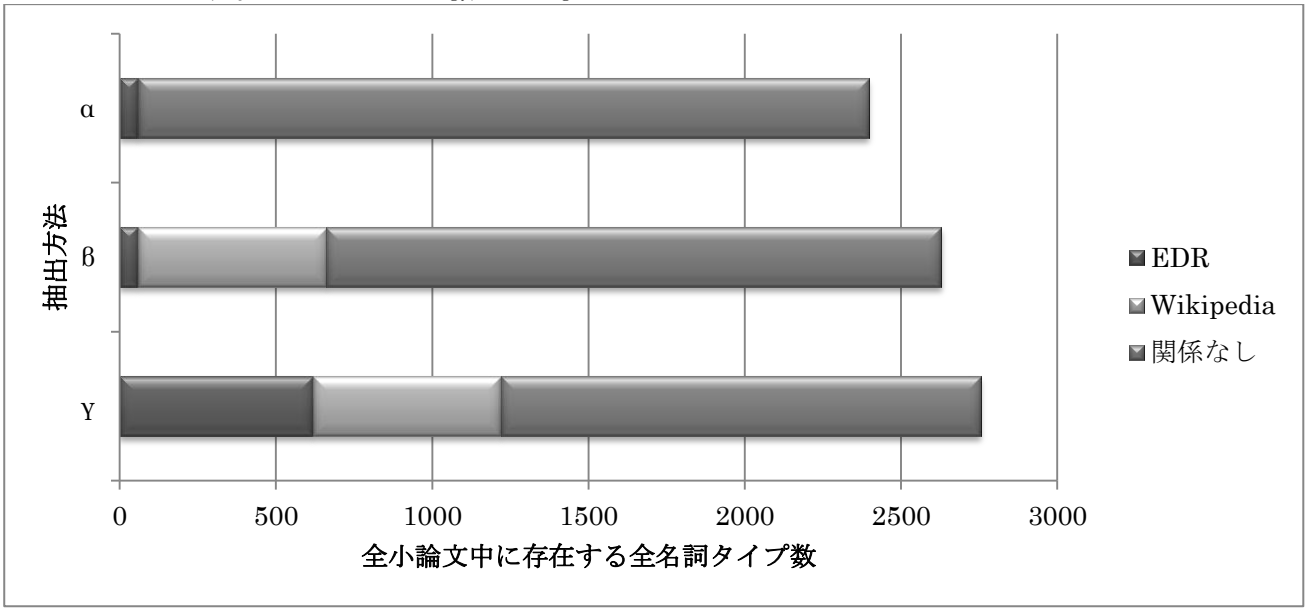

図２：全小論文中に存在する全名詞数(異なり)と上位下位関係のある名詞数