2F4-2

文書ジャンル毎の特徴を用いた文書校正支援

Development of a proofreading support system adapted to the document genre

久保田 敦[*1] 横野 光[*2] 高村 大也[*2] 奥村 学[*2]
Atsushi Kubota Hikaru Yokono Hiroya Takamura Manabu Okumura

[*1]東京工業大学 総合理工学研究科
Tokyo Institute of Technology, Interdiciplinary Graduate School of Science and Engineering

[*2]東京工業大学 精密工学研究所
Tokyo Institute of Technology, Precision and Intelligence Laboratory

We propose a method for genre-dependent automatic proofreading of Japanese text. In the proposed method, the formality and the writing style of the genre are taken into consideration. For this purpose, we also propose a method for automatically assigning the formality and the writing style to each content word and make use of the assigned information to construct a classifier. The proposed method has another advantage that it does not depend much on context features, which could be wrong themselves.

1. はじめに

近年パーソナルコンピュータ・携帯電話等の発展によりデジタル文書を作成する機会は非常に多く，作成される文書のジャンルはメールや技術文書，広告，ブログなど様々である．それに伴い，書き手の文書作成を支援するために，文書校正支援という技術も発展した．文書校正とは言い換え手法の一つで，文書中の誤入力した文字や末尾表現の揺れなどの不適切な表現を適切な表現へ言い換える手法である．文書校正支援は Web 上のサービスや Microsoft Office の文書校正機能などで使用することが可能である．

これらの文書校正支援システムでは，誤字脱字や表記のゆれ等の一般的な日本語として不適切な表現の指摘が可能である．しかし，これらのシステムは文書がどんなジャンルで使用されるかを考慮していない．そのため，どのような文書でも同一の指摘を行うので，ジャンル毎の表現の特徴を活かした柔軟な校正は難しい．通常，文書はその目的や対象とする読者のためにある統一的な難易度で書かれるべきである．例えば “増える” という表現は日常で頻繁に使用されるが，この表現は論文などのジャンルでは “増加する” という，よりフォーマルな表現を使用すべきである．そのため，文書校正においても適切な難易度等を考慮するべきである．

そこで本研究の目的は，文書が使用されるジャンルを考慮した文書校正支援である．そのために各単語に難易度情報を付加しそれを文書校正に利用することで，より適切な単語の言い換えを実現する．具体的には，機械学習の手法を用いて，同じ意味の単語の集合から，適切な単語の分類を行う分類器を作成する．従来の手法で多く使用されている文脈を基にした素性と同様に，付加情報を基にした素性が，同じ意味である各単語を分類するための素性として有用であることを示す．そして，文脈を基にした素性だけでは分類出来ない問題について対処するだけでなく，その性能をさらに向上できることを示す．

連絡先: 久保田敦，東京工業大学精密工学研究所奥村研究室，〒 226-8503 横浜市緑区長津田町 4259 R2 棟 7 階 728 号室，kubota@lr.pi.titech.ac.jp

2. 関連研究

文書中の不適切な表現を適切な表現へ言い換える研究は数多くある．Yu ら [7] は周辺の文脈に含まれる語を利用し，同じ意味の単語である Near-Synonym の集合の中で文脈に最適な単語を選択できる手法を提案した．Liu ら [6] は Semantic Role Labeling の結果を基にした素性を，従来の手法で使用されている前後の文脈を基にした素性と組み合わせることで，その性能を向上させることが出来ることを示した．

また，各単語に対し情報を付加し，それを基に言い換えを推薦する研究が同様に存在する．松吉ら [4] は日本語の機能語に対して，難易度や文体等の情報を付加した辞書を構築し，その情報を利用して難易度や文体を制御した日本語機能表現の言い換えが可能であることを示した．また，千田ら [1] は日本語の単語がどのようなドメインの文書で多用されるかの情報を基にその単語の難易度を求める手法を提案した．ドメインとしては，高等教育機関ドメイン (ac.jp)， インターネットサービスプロバイダなどに発行されるドメイン (ne.jp) を選択し，その中での出現頻度を使用している．具体的には，難易度を求める式の重みを，人手により作成した訓練データで学習させる．そうすることで，ドメインでの出現頻度から難易度を求める式を決定し，その難易度が人手による難易度と相関があることを示した．

3. 文書校正について

3.1 問題設定

本研究では文書校正を言い換えの一種であると捉える．言い換えとはある一つの表現に注目し，その表現を他の表現に変換することである．そして，表現としては 1 単語の場合や複数の単語の場合など様々である．本研究ではこの中の日本語 1 単語を対象とした問題を取り扱う．また，その言い換える対象とする 1 単語を**対象語**，同じ意味をもつ単語の集合を**意味クラスタ**呼ぶ．そして，ある対象語に注目したときに，その属する意味クラスタ内からその文書に最も適切な単語を選択するという問題を解く．

表 1: 内容語への情報付加

単語	難易度					文体			
	1	2	3	4	5	常体	敬体	口語体	堅い文体
増える	0.8717	0.0698	0.0256	0.0183	0.0144	0.8224	0.0451	0.1323	0
増加	0.7173	0.1037	0.0122	0.1212	0.0453	0.8578	0.0991	0.0430	0
嵩む	0.6705	0.1516	0.0117	0.1338	0.0321	0.8674	0.0844	0.0372	0.0108

3.2 意味クラスタの作成

対象語を言い換えるためには，その単語がどんな意味クラスタに含まれるかを定義する必要がある．そこで，本研究では日本語 WordNet [2] を用いて，意味クラスタにまとめる．

日本語 WordNet では単語を類義関係のセット (synset) でグループ化している．そこで本研究では synset を意味クラスタとする．

3.3 日本語機能表現辞書つつじ

本節では日本語機能表現辞書つつじ [3] について説明する．

日本語の単語は主に内容的な意味を表す “内容語” 以外に，助詞や助動詞といった，主に文の構成に関わるもの “機能語” にわけることができる．また，1 単語でなく複数の単語で構成されたものも含めた場合にそれぞれ “内容表現”，“機能表現” と呼ばれる．

つつじは松吉らによって作成された機能表現の辞書であり，各機能表現に対してその難易度や文体，否定表現であるかなどの情報が付与されている．

そして，各文書はそのジャンルに応じて使用される単語に特徴があり，その情報が文書校正に有用であることが報告されている [4]．そこで我々は各文書ジャンルで特定の難易度，文体の単語が使用されやすいという特徴を文書ジャンルの特徴として用いる．

4. 提案手法

4.1 内容語への情報付加

機能語の情報に注目すると各文書ジャンルには特定の難易度，文体の単語が使用されやすい特徴が存在する．そこで，我々は同様に内容語に対しても難易度等の情報を利用すれば，ジャンルに適した表現ができると考えた．しかし，内容語に関してはそのような辞書はなく，また機能表現よりも非常に多い内容語に対し難易度を定義し付加するのはとても高コストである．そこで，日本語機能表現辞書つつじにより付与された機能語の情報を基にすることで内容語に難易度等の情報を付与できると考えた．

そのために本研究では 3 つのステップにより内容語と共起しやすい機能語から内容語に難易度等の情報を付与する手法を提案する．

4.1.1 内容語と共起する機能語の取得

最初のステップではコーパスから内容語と機能語の共起頻度を調べる．今回は対象と成り得る日本語 WordNet 全内容語に対しそれぞれ Google で検索を行い，上位 10 件のスニペットの集合を内容語の検索結果文とする．そして共起頻度は，クエリとした内容語とその検索結果文に含まれる機能語の共起回数とする．

4.1.2 機能語の重要度計算

このステップでは，各機能語の重要度を求める．機能語の重要度は，多くの難易度の文書で出現している場合は低く，特定の難易度の文書でしか出現しない場合は高くなることが期待される．そのため，重要度はより多くの種類の内容語と共起しているほど低くなると言える．そこで，各内容語の検索結果から各機能語に対し，以下の式で重要度を求める：

$$F_w = \log \frac{|V|}{\sum_{c \in V} \delta(w \in D_c)} \tag{1}$$

この式において，F_w が機能語 w の重要度，V が日本語 WordNet に含まれる単語集合，D_c は内容語 c に対する検索結果，$\delta(w \in D_c)$ は検索結果 D_c に機能語 w が含まれたら 1，それ以外なら 0 となる関数である．

4.1.3 内容語への情報付加

最後に検索結果文に含まれる機能語をつつじの情報を基に分類，集計し，内容語毎の付加情報とする．そのために，機能語の難易度と文体に注目して内容語毎に以下の処理を行う．

1. 内容語の検索結果に含まれる機能語毎に難易度 5 段階，文体 4 段階のそれぞれ 1 つに変換する．
2. 機能語 w が検索結果に出現した毎に，重要度 F_w を w の難易度，文体と内容語が共起した回数として加算する．
3. 全ての機能語に対して 2 の処理を行い，各難易度，文体で合計をとる．
4. 難易度と文体それぞれで合計値が 1 となるように正規化する．

表 1 は本手法による “増える” とその意味クラスタの語への情報付加の結果の例を示す．表 1 により同じ意味クラスタであってもその共起する機能語の難易度等に異なりがあることがわかる．例えば，“増える” は一般によく使われいるようだが，周囲の難易度が高いと “増加” が使われやすくなること，“嵩む” は低い難易度のものとはあまり共起しないことなどがわかる．

4.2 文書校正

本研究の文書校正では対象語の意味クラスタ内で最も適切な単語を提示することを目的とする．そのため，対象語及び意味クラスタ内の他の単語へ言い換えた場合でそれぞれで事例を作り，対象語の事例が，それ以外の事例よりも適した事例として，その分類器を学習する．今回は，分類器の一種である Ranking-Support Vector Machine（Ranking-SVM）[5] を用いる．

5. 実験

4 節の提案手法による内容語への情報付加を基にする素性を，既存の手法で使用される素性と組み合わせてその性能を向上できることを示すため，以下の実験を行う．また，情報付加の手法として千田らの手法を基にした場合とも比較する．

表 2: 文書校正に用いる素性の一覧

素性	値	種類
対象語 unigram	0,1	文脈依存
対象語を含む単語 bigram	0,1	文脈依存
対象語がコーパス内で出現した割合	0〜1	出現頻度
対象語の各難易度との共起割合	0〜1	提案手法
対象語の各文体との共起割合	0〜1	提案手法
対象語との共起割合で難易度 j は大きい方からｎ番目	0,1	提案手法
対象語との共起割合で文体 j は大きい方からｎ番目	0,1	提案手法
ドメイン内での対象語の出現回数/意味クラスタの出現回数	0〜1	千田
ドメイン内での対象語の出現回数/内容語の全出現回数	0〜1	千田

表 3: グループ分け

グループ	説明	問題比率
all	全ての単語	1
high	訓練データに１０回以上出現した単語	0.7872
low	訓練データに１回以上１０回未満出現した単語	0.1460
zero	訓練データに出現しなかった単語	0.0668

5.1 データ

本研究では，言語処理学会第 15 回年次大会の論文 50 本を使用し，それらを 5 等分し，4 つを訓練データ，1 つをテストデータとして実験する 5 分割交差検定を行う．

5.2 実験設定

実験では，コーパス内の文書を MeCab [8] により形態素解析を行い，その結果中の日本語 WordNet に含まれる全ての単語を対象語とする．コーパスには 5.1 節の論文のデータを用いた 5 分割交差検定を行う．また，Ranking-SVM の実装として SVM$^{\text{rank}}$ [*1] を用いる．

また，今回分類に用いる素性を表 2 に示す．素性は既存の研究で非常によく使用される前後の文脈依存の素性，訓練データでの出現頻度による素性，提案手法による付加情報による素性，千田らの手法を基にした付加情報の素性の 4 つに分類される．提案手法の素性は 4 節の結果から作成する．千田らの手法を基にした素性は内容語の難易度の決定に ac.jp と ne.jp ドメインでの出現回数を利用しているため，その出現回数から素性を作成した．

5.3 評価について

今回の評価の方法として Ranking-SVM によって分類された値が大きい物ほど適切だとする．そして，正解率と Mean Reciprocal Rank で評価を行う．それぞれ以下の 2 つの式で表現される．正解率は元の論文の対象語を復元できた割合を示す．MRR は元の対象語をどれだけ他の意味クラスタの語より上位と判定できたかを示す：

$$\text{正解率} = \frac{1}{Q}\sum_{i=1}^{Q}\delta(rank_i = 1), \tag{2}$$

$$\delta(rank_i = 1) = \begin{cases} 1, & \text{if } rank_i = 1 \\ 0, & \text{if } rank_i > 1 \end{cases}$$

$$MRR = \frac{1}{Q}\sum_{i=1}^{Q}\frac{1}{rank_i}. \tag{3}$$

正解率，MRR の式において，Q はテストデータの問題となった全対象語の数，$rank_i$ は問題 i における元の対象語の順位，$\delta(rank_i = 1)$ は問題 i において元の単語が 1 位であれば 1，それ以外であれば 0 を返す値である．

そして，評価は全問題を対象としたグループ ALL の評価と対象語の訓練データでの出現回数毎に表 3 のようにグループに分けて正解率，MRR を計算して行う．

5.4 比較手法

文書校正における baseline として，意味クラスタ内から訓練データでもっとも出現回数の多い単語を選択するモデルとした．また，出現回数が同じ回数の場合は判別出来ていないとする．また，文脈と出現頻度と千田らの手法を元にした素性を使用した場合と，提案手法と千田らの素性の両方を使用した場合とも比較する．さらに，4 節での Google 検索はドメインを指定せずに行ったが，千田らのようにドメインを指定した検索結果を使用した場合の提案手法とも比較する．

6. 実験結果と考察

実験結果の表 4 のグループ ALL の比較により，Baseline や出現頻度+文脈のみの場合や千田らの手法を元にした付加情報の素性を加えた場合に比べ，提案手法による付加情報を用いた場合が良い性能であることがわかる．また，出現頻度+文脈+提案手法と，出現頻度+文脈+提案手法 (ドメイン指定) を比較すると ALL，high，zero グループの精度が上昇しているため，よりノイズの少ない文書群を用いることが有効であることがわかった．しかし，千田らのような他の付加情報の素性と組み合わせた場合にその精度の向上は見られなかった．

また，文書校正を目的とした場合，対象文書内の語が全て正しいという保証がないため，テストデータを誤りを含ませたエラーデータで実験も行った．

エラーデータは論文毎に，以下の処理を行い作成する．

*1 http://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html

表 4: 実験結果

素性種類	正解率				MRR			
	ALL	high	low	zero	ALL	high	low	zero
Baseline (出現頻度)	0.5335	0.6245	0.2868	0.0000	0.6840	0.7718	0.4850	0.0832
文脈+出現頻度	0.5826	0.6816	0.2983	0.0000	0.7163	0.8055	0.4970	0.1073
文脈+出現頻度+提案手法	0.5954	0.6939	0.3207	0.0351	0.7209	0.8083	0.5091	0.1529
文脈+出現頻度+千田	0.5930	0.6932	0.3043	0.0431	0.7189	0.8081	0.5005	0.1442
文脈+出現頻度+提案手法+千田	0.5939	0.6932	0.3120	0.0381	0.7205	0.8083	0.5045	0.1572
文脈+出現頻度+提案手法 (ドメイン指定)	0.5967	0.6986	0.3033	0.0371	0.7210	0.8098	0.5024	0.1512

表 5: 誤り率を考慮した実験

素性種類	誤り率							
	0		25		50		100	
	正解率	MRR	正解率	MRR	正解率	MRR	正解率	MRR
Baseline	0.5335	0.6840	0.5335	0.6840	0.5335	0.6840	0.5335	0.6840
文脈+出現頻度	0.5826	0.7163	0.5523	0.6891	0.5040	0.6630	0.4894	0.6508
文脈+出現頻度+提案手法	0.5954	0.7209	0.5787	0.7120	0.5638	0.7071	0.4992	0.6599

1. 論文内に出現した単語のクラスタを列挙する．ここでの内容語のクラスタは “synset”，機能語のクラスタはつつじにおいて “意味クラス・左右接続が同じもの” である．

2. 列挙したクラスタのうち，一定の割合を選択しクラスタ毎に 1 語を選択する．

3. 論文内の選択されたクラスタに含まれる語は選択された語に全て置き換える．

この置き換えられるクラスタの割合を誤り率といい，それを変化させた場合の結果を表 5 に示す．

これにより，文脈の素性を用いた場合に，前後文脈に誤りが含まれているほど，その精度が下がっている事がわかる．しかし，文脈のみの場合に比べると，付加情報を用いた素性を加えた場合のほうがより精度の減少を抑えていることがわかる．

7. 結論と今後の課題

本研究では特定のジャンルに適した校正を行うにはそこで使用される難易度等の情報が有用であるとし，対象文書の校正の実験を行った．その際，機能語が日本語機能表現辞書つつじによって付与された情報を基に統一的な校正を行えることに注目し，内容語にも共起する機能語から同様の情報を付与する手法を提案し，校正に利用した．その結果，既存の手法のベースとなる n-gram の素性に組み合わせた際にその精度を向上させることが出来た．また，既存の情報付加手法よりも文書校正に適した情報で有ることを示すことができた．

今後は，ノイズの少ないドメインの文書から情報付加した場合に性能の向上が見られたことから，より多くの文書から情報付加を目指したい．さらに，他の情報付加手法とも上手く組み合わせて性能を向上させることについても考えていきたい．

また，本研究では複数の語を対象とした言い換えや内容語と機能語を複合させた言い換えについては対応出来ていない．そのため，それらの意味クラスタをまとめることが出来れば，より実用的な校正ができると考えている．そして，最終的には実データに対して校正を適応しその効果を検証していきたい．

参考文献

[1] Yasuko Senda, Yasusi Sinohara and Manabu Okumura. Automatic Terminology Intelligibility Estimation for Readership-Oriented Technical Writing. In *Proceedings of LREC 2006*, pp.1506–1509, 2006.

[2] Francis Bond, Hitoshi Isahara, Kyoko Kanzaki, and Kiyotaka Uchimoto. Boot-strapping a WordNet using multiple existing WordNets. In *Proceedings of the 6th International Language Resources and Evaluation (LREC2008)*, pp.2420–2423, 2008.

[3] 松吉俊，佐藤理史，宇津呂武仁．日本語機能表現辞書の編纂．自然言語処理，Vol.15，Nol.2，pp.75–99，2007.

[4] 松吉俊，佐藤理史．文体と難易度を制御可能な日本語機能表現の言い換え．自然言語処理，Vol.15，Nol.2，pp.75–99,2008.

[5] Torsten Joachims. Training Linear SVMs in Linear Time. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 217–226, 2006.

[6] Xiaohua Liu, Bo Han, Kuan Li, Stephan Hyeonjun Stiller, Ming Zhou. SRL-based Verb Selection for ESL. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp.1068–1076,2010.

[7] Liang-Chih Yu, Hsiu-Min Shih, Yu-Ling Lai, Jui-Feng Yeh, Chung-Hsien Wu. Discriminative Training for Near-Synonym Substitution. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pp. 1254–1262,2010.

[8] Taku Kudo. MeCab: Yet Another Part-of-Speech and Morphological Analyzer. http://mecab.sourceforge.net/ .