感動を与える文の自動取得と分析

端 大輝 †　　村田 真樹 ‡　　徳久 雅人 ‡

† 鳥取大学工学部知能情報工学科

‡ 鳥取大学大学院工学研究科情報エレクトロニクス専攻

{s082042, murata, tokuhisa}@ike.tottori-u.ac.jp

1　はじめに

人は日々感動を追い求めるものであり，感動は人が生きる糧である．また，人間の英知は文章という形で記録，保存され，後世に受け継がれていく．そこで本研究では「感動」と「文」に重きを置き，感動を与える文に関する研究を行う．石岡ら [1] は毎日新聞の社説およびコラムを模範とした採点を行う日本語小論文の自動採点システムを構築したが，感動に関する評価は行われていない．

本研究では具体的には以下のことを行う．

1. 感動を与える文の収集

   まず，感動を与える文を人手で収集する．感動を与えない文も人手で収集する．これらを用いて教師あり機械学習により，ウェブ文書からさらに感動を与える文を収集する．

2. 収集した感動を与える文の分析

   収集した感動を与える文を分析する．感動を与える文で多く使われる単語を収集することで，感動を与える文の言語的特徴を明らかにする．

本研究では上記までを実施したが，今後は教師あり機械学習における素性分析を通じて，感動を与える文で多く使われる単語以外の，感動を与える文の言語的特徴も明らかにしたいと考えている．

将来的には，上記で明らかにする言語的特徴を利用することで，人に感動を与えることのできる文を作成する際に役立つ文作成支援システムを構築できるようになると考える．これは，感動を与える演説をする必要のある政治家の原稿作成や，感動を与えることで採用の可能性を高めたい企画書作成の際にも有用である．対象となる人に感動を与えることで，良い印象を与える可能性が高まる．本研究は，名文を書くという，一見すると文系よりの研究を計算機科学的手法により進めるものである．

本論文の主な主張点を整理すると以下のようになる．

- 本研究は，感動を与える文の自動取得と分析を自然言語処理の技術を使って初めて行ったものである．
- 感動を与える文を自動収集し，収集したデータを分析することで，感動を与える文に多く出現する単語として，「人生」「人々」「幸福」「友情」「青春」「恋愛」などが得られた．これらを用いた文は感動的なものになりやすいと考えられる．この知見は，感動を与える文を作成する際に役立つものと思われる．
- 機械学習により大規模ウェブデータから感動を与える文を適合率 0.4 で抽出できることを示した．

2　感動を与える文の収集

本研究ではまず Google 検索等で，感動を与える文と感動を与えない文を収集する．その後で，これらのデータを学習データとして，教師あり機械学習を利用して，ウェブコーパス [2] からさらに，感動を与える文と感動を与えない文を収集する．

本論文では感動を与える文を正例，感動を与えない文を負例と呼ぶ．

2.1　人手による感動を与える文の収集

Google 検索にて「という言葉に感動した」等の検索ワードで得た文を正例の候補，「という文」等の検索ワードで得た文を負例の候補として取り出す．候補文を人手で判定して正しい正例と負例を得る．

「という言葉に感動した」がつく例文と，「という文」がつく例文を以下に示す．

1. 「という言葉に感動した」がつく例文

   「毎日が未来」という言葉に感動した

2. 「という文」がつく例文

   読み取り専用という文

さらに，Google 検索で「名言集」を検索し，そこから正例の候補を，Yahoo のニュース記事等から負例の候補を集める．これらの候補文も人手で判定し正しい正例と負例を得る．

正例と負例の判断基準としては,「感動した」と書いてあるものと「名言」を正例として，感嘆符等が付いておらず，客観的事実のみを述べているものを負例とした．正例と負例の例を以下に示す．

1. 正例の例

   残りの人生，世のため人のために働きます

2. 負例の例

   フォルダにコピーする

上記の処理を実際に行い，正例 1,018 個，負例 406 個を得た．

2.2 教師あり機械学習を利用した感動を与える文の収集

2.1 節で得た，正例と負例を学習データとした教師あり機械学習を行う．ウェブコーパス [2] を教師あり機械学習で，感動する文かいなかを判定することで，ウェブコーパス [2] から感動する文を収集する．

具体的な手順は以下のとおりである．

1. 2.1 節の方法で得た正例 1,018 個，負例 406 個を学習データとする．
2. 学習データを用いて機械学習を行う．学習結果を利用して新しいウェブデータ 1 万文について正例，負例の判定を行う．正例とされた事例を人手でチェックして正しい正例と負例を新たに作成し学習データに追加する．
3. 2 を 10 回繰り返す．

機械学習には認識性能が優れている Support Vecter Machine(SVM) を実装している TinySVM[1]を用いる．素性には名詞，動詞，形容詞，形容動詞，連体詞，副詞，接続詞，感動詞の単語を用いる．素性の取り出しには，形態素解析を行う ChaSen[2]を使用する．

表 1: 追加された正例と負例の個数

	正例	負例
1 回目	111	2108
2 回目	30	164
3 回目	15	137
4 回目	8	119
5 回目	8	111
6 回目	14	90
7 回目	19	96
8 回目	28	57
9 回目	23	65
10 回目	19	59
合計	275	3006

機械学習結果の出力においての正例と負例の判定の基準は以下のとおりである．何らかの感想を得ることのできた文は正例とする（例：「大学に合格した」「明日は映画の公開日だ」）．客観的事実しか書かれていないものは負例とする（例：「大学がある」「プラスチックは燃えないゴミだ」）．

実際に上記のことを行った．10 回の繰り返しで得た正例と負例の個数を表 1 に示す．

最初に収集したものとあわせて，正例と負例は，1,293 個と 3,412 個となった．上記では 10 回しか繰り返さなかったが，さらに繰り返すことでより多くの正例と負例を収集できる．

正例と負例の評価は被験者 a(1 名) が行った．被験者 a(1 名) の評価した正例 20 個と負例 20 個を，被験者 a とは別の被験者 3 名により，評価させた．被験者 3 名の多数決の結果と被験者 a の判定結果を比較すると，カッパ値 0.58 を得た．中程度の一致であった．

3 収集した感動を与える文の分析

前節で収集した 1,293 個の正例と 3,412 個の負例を用いて，感動を与える文の分析を行った．ここでは単語に基づく分析を行った．

正例と負例から単語を取り出し，各単語について正例または負例に出現する頻度，正例に出現する割合をもとめた．正例に出現する割合が 0.8 より大きく，出現頻度が 5 以上である単語を，正例に出現しやすい単語として取り出した．取り出した単語の一部を表 2 に示す．

「人生」「人々」「幸福」「友情」「青春」「恋愛」などの単語が得られた．これらを用いた文は感動的なも

[1] http://chasen.org/~taku/software/TinySVM/

[2] http://chasen-legacy.sourceforge.jp/

表 2: 正例に出現する割合の高い単語の例

単語	正例に出現する割合	正例の頻度	負例の頻度
幸福	1.00	83	0
友情	1.00	29	0
青春	1.00	18	0
悲しみ	1.00	12	0
存在	1.00	10	0
...	...	...	...
我々	0.97	37	1
不幸	0.97	32	1
愛さ	0.96	23	1
恋愛	0.96	44	2
恋	0.95	122	7
孤独	0.94	32	2
この世	0.94	16	1
愛し	0.94	31	2
愛する	0.94	30	2
あらゆる	0.93	14	1
お前	0.93	13	1
瞬間	0.92	11	1
人生	0.91	145	14
未来	0.91	20	2
幸せ	0.91	20	2
喜び	0.91	10	1
女	0.91	115	12
運命	0.90	19	2
死ぬ	0.90	37	4
知る	0.90	9	1
...	...	...	...
人々	0.81	17	4
感動	0.80	8	2

のになりやすいと考えられる．

4 感動を与える文の自動抽出性能

本研究の技術は，感動を与える文を自動抽出することに役立つ．本節では，感動を与える文を自動抽出する性能を評価する．

評価結果を表 3 に示す．評価データはウェブコーパスの新たな 1 万文とし，各手法で正例とした事例からランダムに抽出した 100 個の事例を人手で評価し (ベースラインのみ 200 個の事例を人手で評価)，その結果から近似的に適合率，再現率，F 値を算出した．ベースラインは，すべてを正例と判断する手法であり，この手法で検出した正例の個数から，再現率の分母を推定している．

「ML x 回目」は，2.2 節の機械学習に基づく方法で x 回目の正例と負例の追加をした後の学習データを用いた場合である．パターン 1 は，3 節の分析において頻度が 5 以上でありかつ正例に出現する割合が 0.8 以上であった単語を一つでも含む文をすべて正例として抜き出す方法である．パターン 2 は,「感動」という語を一つでも含む文をすべて正例として抜き出す方法である．

表 3: 種々の手法の抽出性能

手法	適合率	再現率	F 値
ML 0 回目	0.06	0.25	0.10
ML 1 回目	0.26	0.08	0.12
ML 2 回目	0.29	0.07	0.11
ML 5 回目	0.31	0.05	0.09
ML 10 回目	0.40	0.05	0.09
ベースライン	0.07	1.00	0.12
パターン 1	0.11	0.08	0.09
パターン 2	1.00	0.002	0.003

10 回正例と負例の追加をした後の機械学習では適合率が 0.40 が得られている．

5 関連研究

本研究と同様に文を評価するものとして，石岡らのもの [1] がある．本研究は，文が感動を与えるものであるかを評価するものであるのに対し，石岡らは日本語小論文の自動採点を行うシステムを構築した．石岡らは毎日新聞の社説およびコラム (余録) を学習し，これを模範とし採点を行う日本語小論文の自動採点システムを構築した．そのシステムでは，文章の形式的な側面である「修辞」と，アイディアの理路整然とした表現の程度を示す「論理構成」と，トピックに関連した語彙が用いられているかを示す「内容」の 3 つの観点から小論文を評価する．

分析に機械学習を利用した研究としては以下がある．

村田ら [3] は情報の重要度を決める要因を明らかにし，その知見に基き情報の重要度を自動推定するシステムの構築を目指し，新聞記事やアンケートデータを用いた機械学習を利用した重要度に関する研究を行った．この研究により 1 つの記事のみを用意して重要かどうかを判別するよりも，2 つの記事を用意してどちらがより重要であるかを判別する方が簡単であると報告している．

村田ら [4] は情報の重要度を自動推定するシステムの構築を目指し，ユーザ毎に異なる情報の重要度について調査する研究を行った．どのような情報を重要と考えるかは個々人により異なるものであり，ユーザごとの興味をアンケートより抽出しその結果を利用してユーザ毎に異なる情報の重要度についての調査を行った．この研究によりユーザ個人の興味情報がそのユーザの重要な記事の判断と相関があると報告している．

馬場ら [5] は小説テキストを「ジャンル」と「登場人物」により分類する手法を提案した．SVM を用いて登場人物とその特徴を抽出し，それらからジャンル推定を行う．ジャンルによって有効な特徴量が異なること，学習データとテストデータの語の頻度分布の差異で分類精度が変化すると報告している．

本研究は読者が感動するかどうか，すなわち読者の感情を推定しているものととらえることができる．感情に関わる研究としては以下がある．

松本ら [6] は会話文から感情を推定するためには，文に含まれる語句の感情的意味と文の表す事象の意味内容を読み取る必要があるとし，語の意味属性と感情生起事象文型パターンに基いた感情推定アルゴリズムを提案した．この研究は，本研究での正例負例の判定基準作成で参考にした．

徳久ら [7] は話者の感情を適切に表現する応答を生成するために，発話の意味する感情を感情極性より粒度の細かい感情クラスで推定する必要があるとし，文の意味する感情を推定する手法を提案した．ウェブ上のテキストから感情が生起する要因となる事態を獲得し，それを用いて感情を推定する．あらかじめ感情極性を推定することで感情推定精度が向上すること，複数の類似事例を用いることで感情推定精度が向上することを報告している．本研究と同じくウェブ上からテキストを入手し判定している．

石川ら [8] はリビング環境において ”深い感動 ” を喚起させることを目的とした「場」の実現方法を検討し，その結果を報告した．また，リビング環境において ”深い感動 ” を喚起させる「場」の一つの実現方法を提案した.「体感させること」「フレームレス」「音源に含まれる音が身体に直に伝わること」の重要性を報告している．この研究は，本研究での正例負例の判定基準作成で参考にした．

大出ら [9] は音楽を聴取するという行為においても，喚起される感動の種類に違いがあることを実験的に示すことを目的とし，既存の研究において分類した感動語を評価語とした感動評価尺度を作成した．音楽を聴いてどういった良さを感じたのかを調べる目的においては感情評価尺度は有効であると報告している．この感情評価尺度の箇所を，本研究での正例負例の判定基準作成で参考にした．

6　おわりに

本研究では感動を与える文の作成支援のために，感動を与える文の収集とそれらの分析を行った．分析の結果，感動を与える文に多く出現する単語として,「人生」「人々」「幸福」「友情」「青春」「恋愛」などが得られた．これらを用いた文は感動的なものになりやすいと考えられる．この知見は，感動を与える文を作成する際に役立つものと思われる．

本研究では感動を与える文を機械学習で抽出することも行った．現状で大規模ウェブデータから適合率 0.40 で感動を与える文を取り出すことができることがわかった．

今後は，本研究の技術を利用してより多くの感動を与える文を収集したいと考えている．さらに，感動を与える文の分析として，構文パターン，修辞的表現などの単語以外の分析も行いたいと考えている．

参考文献

[1] 石岡恒憲, 鷲坂由紀子, 二村英幸. Jess: 日本語小論文の自動採点システム. *http://coca.rd.dnc.ac.jp/jess/*, 2004.

[2] Daisuke Kawahara and Sadao Kurohashi. Case frame compilation from the web using high-performance computing. *In Proceedings of the 5th International Conference on Language Resources and Evaluation*, pp. 1344–1347, 2006.

[3] 村田真樹, 西村涼, 金丸敏幸, 土井晃一, 松岡雅裕, 井佐原均. 情報の重要度を決める要因の抽出・分析と重要度の自動推定. 言語処理学会第 14 回年次大会, pp. 907–910, 2008.

[4] 村田真樹, 西村涼, 金丸敏幸, 土井晃一, 鳥澤健太郎. ユーザ個人の興味の影響を考慮した情報の重要度を決める要因の抽出・分析. 言語処理学会第 15 回年次大会, pp. 554–557, 2009.

[5] 馬場こづえ, 藤井敦, 石川徹也. 小説テキストを対象としたジャンル推定と人物抽出. 言語処理学会第 11 回年次大会, pp. 574–577, 2005.

[6] 松本和幸, 任福継, 黒岩眞吾. 語の意味情報を考慮した感情推定アルゴリズム. 言語処理学会第 11 回年次大会, pp. 145–148, 2005.

[7] 徳久良子, 乾健太郎. Web から獲得した感情生起要因コーパスに基く感情推定. 情報処理学会論文誌, Vol. 50, No. 4, pp. 1365–1374, 2009.

[8] 石川智治, 宮原誠. リビング環境において “深い感動” を喚起させる「場」の実現方法の検討. 芸術科学会論文誌, Vol. 2, No. 3, pp. 91–93, 2003.

[9] 大出訓史, 今井篤, 安藤彰男, 谷口高士. 音楽聴取における “感動” の評価要因 — 感動の種類と音楽の感情価の関係. 情報処理学会論文誌, Vol. 50, No. 3, pp. 1111–1121, 2009.