キャッチコピーの言語工学および感性工学的解析と自動生成システム構築に関する研究

常勤研究者の部

萩　原　将　文
慶應義塾大学
理工学部情報工学科
教授

1　はじめに

キャッチコピーは短い文字数で注目を引き、端的かつ効果的に対象の特徴を述べるものである。受け手の注意を引き、関心と興味をかきたて、受け手に行動を行わせることを目的とする。本研究では、キャッチコピーの言語工学および感性工学的解析と自動生成システム構築に関する研究を行った。具体的には、通常の文とは異なる性質を持つことが予想されるキャッチコピーの定量的な特徴を自然言語処理の知見を用いて分析を行い、異なったアプローチによって 2 種類のキャッチコピーの生成システムを構築した。

2　複数コーパスを利用したキャッチコピーの特徴分析

キャッチコピーは比較的短い文で、受け手に強いインパクトを与える。従って、通常の文とは異なる様々な特徴を有することが予想される。ここでは、比較的最近の営利目的のキャッチコピーが大量に収録されている文献[1](以下キャッチコピーコーパス) を分析の対象とし、文法的に正しく適切な説明文が用いられているインターネット百科事典 Wikipedia コーパス[2]、様々な言い回しや話題を扱っているブログコーパスである KNB コーパス[3]、文法的には必ずしも厳密でないが自由度の高い会話文コーパス[4] との比較・検討を行った。

2.1 言語工学的分析

2.1.1 他コーパスとの比較による特徴分析

全体の品詞使用割合の分析

図 1 に、解析によって得られたキャッチコピーコーパス全体における品詞の使用割合を示す。全体の品詞使用割合の傾向としては名詞の割合が一番多く、助詞・記号・動詞と続いていることがわかる。キャッチコピーにおいて、名詞が重要であることがわかる。

他コーパスに関しては、会話文コーパスを除く、３種類のコーパスにおいては名詞が占める割合が最も多く、キャッチコピーコーパスに含まれる名詞の割合は百科事典コーパスとブログコーパスの中間程度であることがわかった。

品詞列に着目した分析

キャッチコピーの文法構造に着目した分析を行うために、単語の連なりを扱う N-gram モデルを用いた。表 1 にキャッチコピーコーパスにおいて、高頻度であった順方向の品詞 3gram 上位 10 件の全体に占める割合を示す。また、該当する品詞 N-gram が他コーパスにおいて占める割合を右側に併載する。表よりキャッチコピーは名詞から始まるものが多いことがわかる。他コーパスとの比較では、キャッチコピーコーパスでよく出現した「名詞, 助詞, 名詞」の

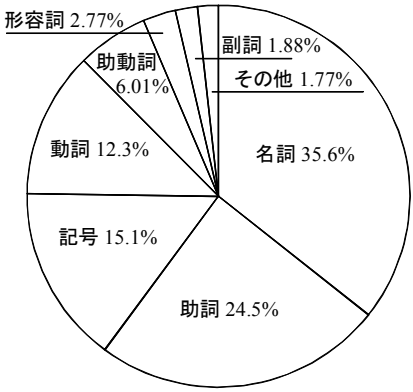

図 1: キャッチコピーコーパスにおける品詞の使用割合

パターンが、百科事典・ブログコーパスにおいても高い割合で見られることがわかる。会話文コーパスについては、特に類似するパターンは見られなかった。順方向と同様に処理を行ったところ、名詞で終わるキャッチコピーが合計で

2,383 個と全体の４割近く占めていることが明らかになった。コーパスとの比較では、一番多かった「記号，名詞」のパターンに関して、百科事典コーパスに類似していることがわかった。

表 1：各コーパスの順方向品詞 3gram の頻度割合

品詞*N*-gram	キャッチコピー(%)	百科事典(%)	ブログ(%)	会話文(%)
名詞, 助詞, 名詞	21.6	17.9	18.0	2.32
名詞, 助詞, 動詞	12.1	2.12	6.43	0.95
名詞, 名詞, 助詞	8.71	18.8	8.61	2.02
名詞, 助詞, 記号	5.00	3.04	2.21	1.50
名詞, 助動詞, 名詞	3.94	0.508	1.04	0.223
形容詞, 名詞, 助詞	2.74	0.254	0.584	0.056
名詞, 助詞, 助詞	2.63	3.12	2.95	0.409
名詞, 名詞, 名詞	2.47	15.5	3.73	0.799
名詞, 助詞, 形容詞	2.37	0.333	1.37	0.297
連体詞, 名詞, 助詞	2.24	2.91	3.15	0.910

次に、キャッチコピー特有の品詞列の定量的な分析を行った。具体的には、キャッチコピーコーパスと他コーパスとの違いを明確にするためにキャッチコピーコーパスと他コーパスの品詞 *N*-gram の頻度をそれぞれの文の数で割って正規化した上で、キャッチコピーコーパスの正規化された値を他コーパスの正規化された値で割っている。これを頻度比 S_f として今後用いる。

順方向において各コーパスと比較を行った際の頻度比 S_f を表 2 にまとめる。まず、表の左側の百科事典との比較では、名詞が支配的なのではなく、動詞・形容詞から始まるところがキャッチコピーの特徴であることがわかる。表中央のブログコーパスと比較では、「動詞，助詞，記号」の品詞列の組み合わせが最も頻度比において高かった。従って、キャッチコピーはブログと異なり、端的で歯切れのよい構造を持つ傾向が強いことがわかる。表の右側の会話文コーパスとの比較では左側・中央と比べると、右側は全般的に上位の値が高かった。特に「形容詞，名詞，助詞」の構造を頻繁に用いられることが会話文との大きな違いであると言える。

さらに、キャッチコピーの文頭から 3gram 分の品詞列が全体としてどのような傾向を持っているのか、表 1 及び頻度スコア式を基に計算を行った。その結果、文の前半において、キャッチコピーの品詞列は百科事典及びブログのもの

に近いことがわかった。

また、表の左側の百科事典との比較では、名詞終わりの体言止めが多いことがわかった。

以上より、キャッチコピーにおける前半の品詞列は、「名詞, 助詞, 名詞」などのパターンが多く、百科事典やブログの構造に似ていると言える。さらに、後半の品詞列は名詞で終わることが多く、体言止めが多用されている傾向があることがわかった。

表 2: 順方向品詞 3gram 頻度比

キャッチコピーVS. 百科事典		キャッチコピーVS. ブログ		キャッチコピーVS. 会話文	
品詞*N*-gram	頻度比Sf	品詞*N*-gram	頻度比Sf	品詞*N*-gram	頻度比Sf
動詞, 助詞, 動詞	39.4	動詞, 助動詞, 記号	14.0	形容詞, 名詞, 助詞	49.1
動詞, 助動詞, 助詞	26.5	名詞, 記号, 動詞	11.0	動詞, 名詞, 助詞	33.9
形容詞, 助詞, 記号	24.7	名詞, 動詞, 記号	10.4	名詞, 助動詞, 名詞	17.7
動詞, 助詞, 助詞	22.0	動詞, 形容詞, 名詞	7.92	副詞, 助詞, 名詞	15.4
動詞, 記号, 動詞	17.6	動詞, 名詞, 助詞	6.75	名詞, 動詞, 名詞	15.0

表 3: 選択されたキャッチコピーの評価

キャッチコピー	5: 面白い	1: つまらない	0: 意味不明	平均
姫の秘めごと。	5	1	0	3.87
助手席を退屈させたくない人へ。	4	4	0	3.07
ステキに、ハートフルな毎日。	0	2	0	2.93
ようこそ日本の、ファーストクラスへ。	0	1	0	2.87
わくわくするから夢がふくらむ、未来が広がる	1	2	0	2.80
目を奪う「夏視線」へ。	1	1	2	2.73
ソファは座るだけのもの?子供たちの答えはもちろん「NO!」です。	1	1	2	2.60
乗る、買う、話すがひとつに。	0	3	0	2.60
ドレッシーを気軽に着る。	0	0	1	2.57
カワイの革新性は、カワイにしか超えられない。	0	2	2	2.53
湯煙おおう、世界有数の温泉の町へ。	2	2	1	2.53
このパフォーマンスを、全ての「音楽」に捧げる。	1	3	1	2.47
創造力から、総造力へ。	0	4	1	2.47
給湯室ブレイクは、「刺激がウマイ!」	0	4	2	2.33
宇宙を旅したウオッチを、地球で手にする贅沢。	1	3	1	2.27
「振って使う」が新しい。	1	3	1	2.27
試そう、合格力の高めかた。	0	0	1	2.27
コンビネーション肌	1	1	4	2.13
ラタンのチェアに身をゆだねて、素直な自分に戻る。	0	1	2	2.07
パソコンするなら、やっぱりノート。	1	5	1	1.93

2.1.2 キャッチコピーコーパスにおける特徴

キャッチコピーコーパスそのものに含まれる単語、係り受け関係、分野ごとの名詞について分析を行った。名詞については「人」、「あなた」、「私」、「心」

などの人を意識したメッセージ性が強い単語が高頻度で現れていることがわかった。一方、固有名詞では地名が多く見られた。

また、構文解析器 CaboCha [5] を用いてキャッチコピーコーパスにおける係り受け関係を分析したところ、係る側として「この」、「その」、「もっと」のような強調する役割を果たすと考えられる指示語・副詞、「あなたの」のような関係性を強調するもの、「新しい」、「いい」、「美しい」などのポジティブなもの、「夏」や「大人の」などの特定のイメージを喚起する単語が高頻度で現れていることがわかった。係られる側では、「ある」，「なる」などの係る側を現実にする単語が多く現れていることが明らかになった。

2.2 感性工学的観点からの分析

ここでは、キャッチコピーの感性工学的観点からの分析を行う。具体的には、ユーザとって「面白い」、あるいは「印象に残る」キャッチコピーの特徴について評価実験の結果を基に分析を行う。

2.2.1 実験結果

表 3 に 20～50 代の男女により評価されたキャッチコピーを示す。表からわかるように、「姫の秘めごと。」の評価値が特に高いことがわかる。ここで、構造と単語について分析を行う。構造については、「名詞，助詞，動詞，名詞，記号」と分けられ、順方向 3gram では二番目に、逆方向 2gram では一番目に多い品詞の組み合わせである。従って、奇抜な転置等は行っていないことがわかる。一方、使われている品詞として、「姫」と「秘め」の２つの名詞が用いられている。「姫」については、キャッチコピーコーパスにおいて３回、「秘め」については 10 回出現している。親しみやすい単語であり、組み合わせが斬新であるものが評価されやすいと言える。

評価が低かったものとの比較より、キャッチコピーは名詞の組み合わせが非常に重要であることが示唆された。

3 提案するキャッチコピー自動生成システム

本システムでは、テーマおよびキーワードを指定することにより、それぞれのテーマとキーワードに特有なキャッチフレーズの生成を試みる。図 2 に提案システム全体の流れを示す。

3.1　分野特有の関連語取得(I.)

テーマ特有でキーワードを考慮したキャッチコピーを生成するために、テーマおよびキーワードが入力される。ここで、図3 のようにキャッチコピーにおける単語の階層性を考慮する。図は、i）のわかりやすさ、ii）のキャッチコピーらしさ、iii）対象分野らしさ、iv）キーワードへの近さがピラミッド構造であることを示している。

3.2　キャッチコピー候補生成(II.)

図4 にキャッチコピーの候補生成の流れを示す。図に示すように、キーワードと関連語がまず最初に入力され、それに基づいて文が抽出される。もし数量が閾値以下なら以前に生成した文の一部および関連語を含む文を再抽出し、閾値を超えるまで再生成を行う。

3.3　キャッチコピー候補選択(III.)

キャッチコピーの選択を行う際、以下の 3 つの指標を用いる。

a)　χ^2値(関連度) によるスコア

キャッチコピー候補において、関連語が含まれていた場合にその値をその関連度を加算することで、それぞれの候補の関連度スコアを算出する。

b)-i キャッチコピーコーパスにおける文構造スコア

キャッチコピーにおいて、より出現頻度が大きいほど、キャッチコピーに相応しい文構造とみなすことができ、これを考慮した指標である。

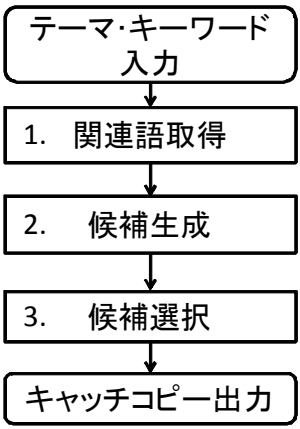

図 2: システム全体の流れ

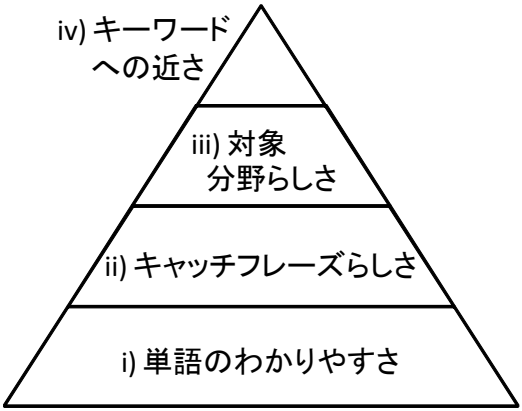

図 3: キャッチコピーにおける単語の階層性

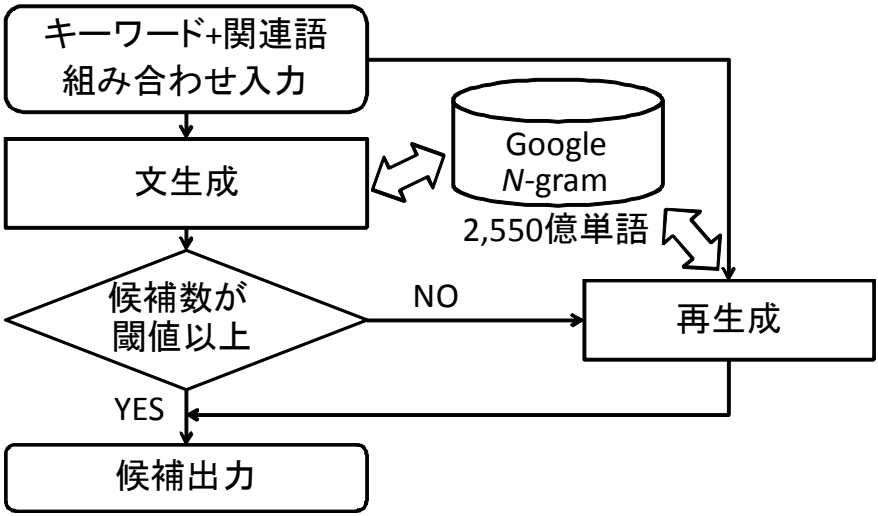

図 4: 候補生成の流れ

b)-ii **複数コーパスを用いた文構造スコア**

高品質なキャッチコピー選択のためキャッチコピーコーパス特有に含まれている品詞列を優先的に抽出する。よって、ある品詞列が存在しているときの、キャッチコピーコーパスにどれくらいの確率で含まれているかを表す指標といえる。

c) 斬新な名詞の組み合わせスコア[6]

斬新な名詞の組み合わせスコアにおいては、相互情報量(PMI) を用いてキャッチコピーらしさを考慮する。西原ら[6] は相互情報量を研究タイトルの評価指標に用いている。提案システムにおいても、相互情報量を斬新な名詞の組み合わせスコアとして取り入れている。

3.4 実験

3.4.1 実験の概要

生成されたキャッチコピーに関して主観評価実験を行い、質の検証を行った。具体的には、「システム生成のキャッチコピー(提案システム)」と「キャッチコピーコーパスからランダムに選択したもの(プロ)」、「Google N-gram コーパスに対して２つのキーワード単語を入力し得られた複数の文からランダムに抽出したもの(ベースライン)」について比較を行なった。なお、被験者は 20 代の学生 13 名である。

評価項目は、適切な文か(3 段階評価)、テーマに関して適しているか(3 段階評価)、キーワードに対して適当な文か(3 段階評価)、キャッチコピーとしての質(5 段階評価) である。

3.4.2 生成例

提案システムが生成したキャッチコピーを以下に示す。

テーマ: 食品

キーワード: ジュース, かき氷

夏のかき氷は懐かしい味。ジュースとヨーグルトの優しいハーモニー♪

テーマ: 交通・レジャー

キーワード: 温泉, リラックス

温泉で心安らぐ森の休日リゾートの風を感じてリラックス

テーマ:メディア

キーワード: 映画, 感動

涙と笑いの青春映画。最高の感動がここにある

3.4.3 実験結果と考察

表 4 に食品、交通・レジャー、メディアに関する評価値の平均を示す。表よ

り、システム生成のキャッチコピーはベースラインよりも文として成立し、テーマに関しても有意水準 5%でベースラインを上回っていた。キーワードについては、プロのものよりも分かりやすく、総合評価においてはベースラインよりも有意水準 5%でベースラインを上回っていることが示された。

表 5 にシステム生成、ベースライン、プロ作成のものに関する評価値を示す。表に示すように、評価値 4 の「やや良い」に関してシステム生成のものはプロ作成のものとベースラインの中間程度となっている。評価値 5 の「良い」に関しては、ベースラインに対して大幅に上回っている。

表 4: 食品、交通・レジャー、メディアの主観評価実験結果

	文の適切さ	テーマ	キーワード	総合評価
提案システム	2.55	2.59	2.51	2.94
ベースライン	1.89	2.49	2.61	2.27
プロ	2.91	2.56	1.91	3.88

表 5: 評価値の割合

	評価値４ (%)	評価値５ (%)
提案システム	22.3	7.05
ベースライン	12.4	1.67
プロ	36.7	31.3

4 キャッチコピーの特性を考慮した自動生成システム

ここでは、特に単語の関係性を考慮に入れたキャッチコピー生成システムについて説明する。

本システムは、Web から抽出した知識を用い、得られた知識ベースとして用いることでキャッチコピーを生成する。これは、Web 上の知識を用いることにより、How to say の言い回しを取得することを主眼としている。

図 5 に提案システムの流れを示す。まず始めに、キャッチコピーにおいて述べる対象(Ex. 景色) とその性質(Ex.美しい) が入力される。その後、Web を用いて対象および性質に関する情報の取得を行い、知識ベースを構築する。次に、知識およびキャッチコピーコーパスの分析によって得られた特徴を用いてキャッチコピーの候補が生成される。

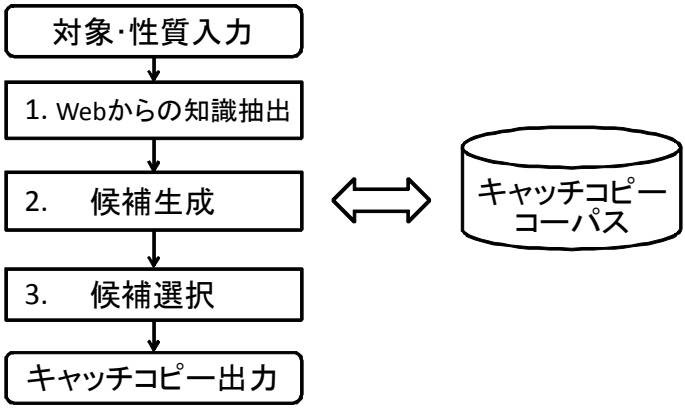

図 5: システム全体の流れ

4.1　実験

提案システムに対象と性質の単語対を入力することで出力を得る。以下に、システム生成の例を示す。

入力：　景色，美しい　　「何もしない贅沢、もある」
入力：ケーキ，おいしい　　「どこでも気分はピクニック！」

4.2　考察

提案法により、「対象」「性質」からより魅力的に対象を伝えるためのキャッチコピーが生成できる可能性が示唆された。一方、現段階では膨大量の候補が生成され、より高精度の選択方法の開発が課題となっている。

5　おわりに

本報告では、キャッチコピーの言語工学および感性工学的解析と自動生成システム構築に関する研究について述べた。

本研究ではまず、統計的処理に基づく言語工学的及びユーザによる感性評価の２つの観点からキャッチコピーの分析を行った。具体的には、一つ目の言語工学的分析については、統計的分析を行うため 6,466 個の膨大な数のキャッチコピーを含むキャッチコピーコーパス及び､比較対象として百科事典コーパス、ブログコーパス、会話文コーパスを利用した。キャッチコピーは順方向に関して「名詞、助詞、名詞」「名詞、助詞、動詞」のような品詞列が多いことがわかっ

た。これに関して、他コーパスと比較により、キャッチコピーの順方向の品詞列は百科事典やブログに近いことが明らかになった。

感性工学的分析については、ユーザの評価実験によって、文法において奇抜な構造をとっているケースはあまり評価されなかった。また、高評価のキャッチコピーにはユーザにとって印象に残った単語が含まれているケースが多かった。

次にこれまでの定量的分析により明らかになった文構造及び単語特徴の知見を利用し、キャッチコピーの自動生成システムの構築を行った。評価実験により、本システムはテーマとキーワードに対して、適切なキャッチコピーを生成可能であることが示された。

さらに、より効果的なキャッチコピーを生成するために異なったアプローチをめざした。これは、キャッチコピーの特徴である言いたいことを直接的に表現しない方法である。具体的には、直接的な表現を間接的な表現に変更するために、「対象と性質」がセットになったクエリを用いて Web へのアクセスを行う。Web から得られた単語を用いて、それらの単語を類似するキャッチコピーの構造に当てはめることにより、よりバリエーションの多い変化に富んだキャッチコピー生成への可能性が示唆された。

参考文献

[1] 久野寧子, “カタログチラシキャッチコピー大百科, ” ピエ・ブックス, 2008.

[2] “Wikipedia 日本語アーカイブ,” http://dumps.wikimedia.org/jawiki/.

[3] 京都大学大学院情報学研究科黒橋研究室,“KNB コーパス(Kyoto-University and NTT Blog コーパス),”http://nlp.kuee.kyoto-u.ac.jp/kuntt/, 2009.

[4] 北九州市立大学国際環境工学部情報メディア工学科上村研究室, “インタビュー形式による日本語会話データベース,”http://www.env.kitakyu-u.ac.jp/corpus/texts/index.html, 1996.

[5] T.Kudo and Y.Matsumoto, “Japanese dependency analysis using cascaded chunking,″ In *Proc. of Natural Language Learning*, pp. 63-69, 2002.

[6] 西原陽子, 砂山渡, 谷内田正彦, “聴講者の興味をひく研究発表タイトルの作成支援, ”言語処理学会年次大会発表論文集, pp. 448-451. 言語処理学会, 2007.