1A2-5

言語的意味と読み手の行動の関係性 : READYFORの運用データを用いた購入行動に関わる本文の影響

Relations between linguistic meanings and reader's behabior : the influence of documents on the purchase action using READYFOR web service

金田 賢哉 *1　堀 浩一 *1
Kenya Kaneda　Koichi Hori

*1東京大学大学院工学系研究科
School of Engineering, University of Tokyo

Most of web documents are evaluated by the readers. For a growth of the media, making an appropriate impression and improving sentences to fit the media is required of editors. As a first step to provide a document edition system for a localized media, I report the result of an analysis about the revisions of documents and the specification of successful cases and not in READYFOR projects witch is cloudfunding web service using Semantic Aggregate Model (SAM).

1. はじめに

インターネットの普及とともに非常に多くの文章が生成されてきた。近年ではソーシャルメディアの提供する評価ボタンなどが幅広く設置され、その内容を読み手に評価させる仕組みが整備され、その数で競ったり順位付けされることも多い。なかでも商用で運用されるメディアでは、広告のクリック数や商品の購入などを収益源としているため、読者層と文章の推敲の方向性が一致しているかという問題や、読み手に合わせた適切な推敲方針を検討することは重要と感得られる。

そこで、本研究では文章を取り扱う上で、特定のメディアの読み手に対してターゲティングをした文章へ推敲を支援することを目的としている。その最初の段階として、クラウドファンディングサービスを行っている READYFOR の文章を用いて、読み手と文章の関係を調べる。

クラウドファンディングサービスとは、インターネット上で金銭やファンを集める仕組みであり、「チケット」と呼ばれるプロジェクト成立の成果物を商品として購入することでプロジェクトの運営資金を調達する。中でも成果物が製品に依存しない社会性の高いプロジェクトのチケットは、金額に対して金銭的には安価でお礼に近い特徴があり、その購入は読み手が本文自体から喚起される共感などの感情が購入行動に影響を与える強い要因であると考えられており、本文が読み手に与える影響の度合いが大きいと考えられる。

また、READYFOR の本文の作成には、プロジェクトの「実行者」に加え、編集専門の「キュレータ」と呼ばれるスタッフと推敲するプロセスがあり、推敲支援モデル [Inui 00] によると、「what-to-say」「how-to-say」「表出/記述」のステップを踏んでいることになるが、表出・記述についてはキュレータがある程度の範囲で一律に担保していると仮定すると、「what-to-say」はプロジェクトを実施したい「実行者」が決める部分であるため、「how-to-say」の編集により読み手に対する伝わり方が変わると考えられる。

過去のプロジェクトの本文について、分析対象となる文章が短いため、複数の本文の平均と分散を比較する手法で分析を行い、推敲の段階で生じている現象を示すとともに、検証方針を示す。

連絡先: 金田賢哉，〒 113-0033 東京都文京区本郷 7 丁目 3-1 東京大学工学部 7 号館 420 号室,(03)5841-1839,kaneda[at]ailab.t.u-tokyo.ac.jp

2. 分析の手法

2.1 意味ベクトルの推定

研究対象となる文章を分析する上で、単語と意味の関係を示す辞書を作成する必要がある。本研究では共起を扱えるトピックモデルとして SAM[Kameya 05] を用いた。これは係り受けなどの関係により存在する共起単語対 w, w' の意味は、潜在クラス c を介して決定されるとするモデルである。

$$P(c, w, w') = P(c)P(w|c)P(w'|c) \tag{1}$$

EM アルゴリズムをこれに適応することで、共起単語対の出現頻度のみで計算を行うことが出来る。なお本研究では $P(c)$, $P(w|c)$ の初期値を乱数で与えてい、抽象クラスの数を $C = 20$ とした。またこれは比喩の研究などで多用されており、文章が読み手に与える影響に注目している本研究でもこの手法が有用ではないかと考えられる。2013 年 1 月の Yahoo ニュースから取得された文章から係り受け関係の頻度データを抽出して計算を行った。

2.2 解析手法

SAM により生成された各単語の潜在クラスへの確率分布を用いて、READYFOR の各プロジェクトにある本文より抽出された共起単語対について解析を行う。潜在クラスの条件確立を求める。

$$P(c_j|w_i) = \frac{P(c_j)P(w_i|c_j)}{\sum_k P(c_k)P(w_i|c_k)} \tag{2}$$

また、２単語の類似度を次のように定義する。

$$sim(w_1, w_2) = \frac{P(c|w_1) \cdot P(c|w_2)}{|P(c|w_1)||P(c|w_2)|} \tag{3}$$

解析対象も共起単語対であるため、共起単語対のベクトル $V_{w,w'}$ を定義する上で、被修飾語 w 側に重みを付けて修飾語 w' と平均をとることとし、重みの付け方として、潜在クラス

の条件付き確率から被修飾語と最も類似度が大きくなる L 個の単語を選び、それを含む平均を取る。本研究では $L = 8$ とする。

$$V_{w,w'}(c) = \frac{1}{L+2}\left(\sum^{L} P(c|w_l) + P(c|w) + P(c|w')\right) \quad (4)$$

また、本研究では解析対象となる各プロジェクトの文章の傾向を調べるために、そのプロジェクトに内在する共起単語対から、文章全体の確率分布を定義する。プロジェクト本文にある全体のベクトル V を、共起単語対の総数 W を用いて次のように定義する。

$$V(c) = \sum^{W} V_{w,w'}(c) \quad (5)$$

なお、解析対象となる文章の長さが短いため、係り受け関係の文法上の種類の区別を行っていない。

3. 本文の現状分析

READYFOR には、文章を書き資金を調達しプロジェクトを実施したいと思っている「実行者」が実施する「プロジェクト」を公開するまでのプロセスとして、実行者が「プロポーザル」と呼ばれる簡単な企画書を提出の上、READYFOR のスタッフによりサービスに対して適切で無いような企画でないことを審査により確認を行う。その上で、最初に実行者が作成する本文となる「第一稿」をもとに、「キュレータ」と呼ばれるスタッフと本文の推敲作業を進め、最終的に公開する「最終稿」を作成する。公開されるプロジェクトは公開前に期限と金額を決め、期限の内にこれを達成した場合には「成立」しなかった場合には「不成立」のプロジェクトとなる。

3.1 推敲前後の違い

この第一稿と最終稿の比較を 2013 年 9 月末までに期限が終了する成立、不成立それぞれの新しい方から 10 プロジェクトを対象に行った。Figure 1 はこれらの傾向を知るために全てのプロジェクトの第一稿と最終稿をそれぞれプロジェクト本文のベクトル V の平均・分散を計算したものである。

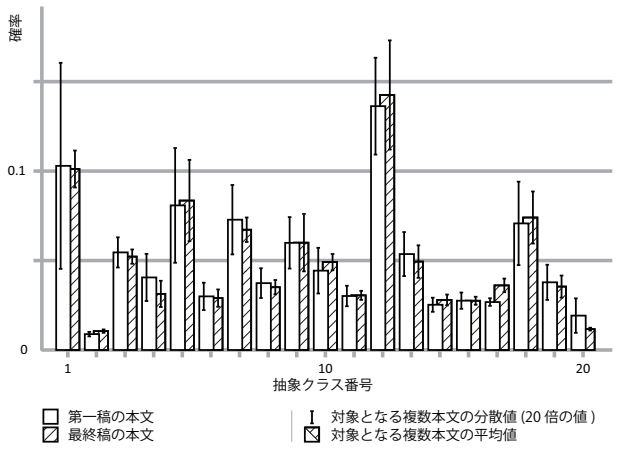

Figure 1: 第一稿と最終稿の比較：平均と分散

この図式は横軸に SAM で用いる抽象クラスをとるため、同じ凡例の和は 1 となるほか、解析を行った単語列の、ある抽象クラスの番号について確率の大きなものを選び集めるとその抽象クラスの中心にある意味合いが分かる場合がある。

Figure 1 より、第一稿と最終稿の平均値は大差が無いことが見て取れる。ことから READYFOR において審査はメディアの特徴にそった内容のものを比較的適切に収集しているのではないかと考えられる。

また、推敲の課程でいくつかの抽象クラスにおいて分散が小さくなっており、その他のクラスについても同等か微増程度であることがわかる。キュレータの行う推敲の役割の 1 つが、文章の品質をそろえる役割を持っていると考えられるとともに、その平均値はもととなるニュース文章の特徴に対する READYFOR のメディアとしての文章の特徴を示していると言える。新しいプロジェクトの文章の推敲を行う上の指標として、その本文のベクトルをこの図式に重ね合わせ、確率分布がある程度類似するよう編集することで一定の効果が期待できると考えられる。 n

3.2 成立・不成立プロジェクトの違い

次に成立、不成立のプロジェクトの本文の最終稿の比較を同様の図式にて行う。

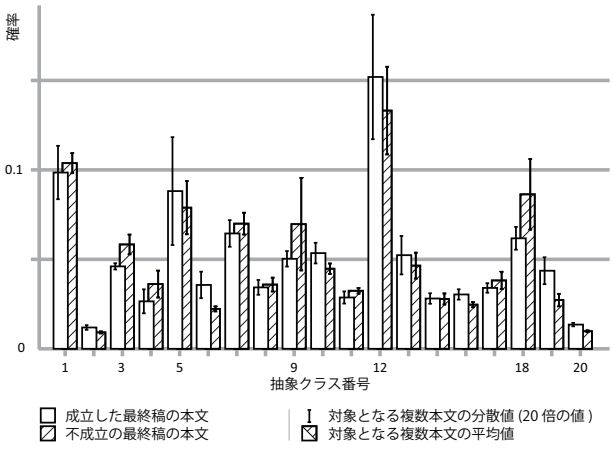

Figure 2: 成立、不成立の最終稿比較：平均と分散

Figure 2 によると、平均や分散についてはいくつかの要素で相違が見られる。抽象クラス番号で 3,9 や 18 のように、平均値に大きな隔たりがありながら成立した最終稿の本文のの分散値が不成立のそれに比べて著しく小さい場合には、その抽象クラスの読み手に対する効果について適切量があるのではないかと考えられる。そのため、新しいプロジェクトの文章を推敲する場合には、その抽象クラス番号の値を可能な範囲で成立した最終稿の本文の平均値に近づけることを指標とすることができると考えられる。一方で、抽象クラスの番号が 5 や 12 に示されるように平均値が大きく異なっても、成立した最終稿の本文の分散が大きい場合には、その要素は成立不成立に大きな影響を与えない可能性が高いと考えられるが、確率分布は計算的に求められているため、読み手が受ける感覚として異なる要素が混在する場合もあるため、内容に注意しながら取り扱うことが望ましいと考えられる。

4. 検証方法

ユーザの行動について効果の程度を検証するには AB テストが有用である。ウェブにおける AB テストは、あらかじめ検証したいページや構成・図表について、2 種類またはそれ以上のものを用意し、訪問者に対してそれぞれのものの出現が等確率になるように振り分け、目的の達成数を比較することで、それぞれの効果を集計する手法である。本研究では、本文のページを 2 種類用意し、購入ページへの遷移をしたユーザ数をそれぞれ調べることになる。現在はこれを検証中であるが、ここでは具体的な例を次に示す。

例としてプロジェクトの ID=1080 の文章を用いる。最終稿になっている、またはそれに近いものについてを、本文 A と呼び、前述の分析を行いプロジェクト本文にある全体のベクトルを求める。この例で取得されたベクトルは、Figure 3 の「プロジェクト本文 A(修正前)」の凡例に示す。

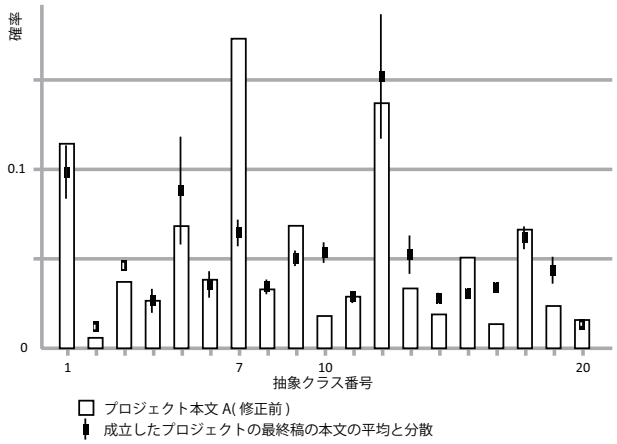

Figure 3: ID=1080 プロジェクトと成立プロジェクトの比較

Figure 3 はプロジェクト本文 A に、Figure 2 に示す成立したプロジェクトの平均と分散の値を重ねたものである。これによると、このプロジェクトは抽象クラス番号の 7 の項目が成功したプロジェクトに比べて著しく大きな値を持つことがわかり、これにより他の確率分布が全体的に小さく見積もられているのではないかと考えられる。また、抽象クラス番号の 7 は成立したプロジェクトの最終稿の分散が成立しないプロジェクトのそれよりも小さいため、前述の考察より、調整することが望ましいと考えられる。

そこで、この抽象クラスに多くの確率分布が集中する本文内の共起単語対を調べると、「○○がない」「支援の現状」など比較的ネガティブな印象につながる表現が多いことが分かり、これらが特に集まっている部分を中心に、文章の本筋が変わらないよう注意をしながら how-to-say について再校正をし、修正後のプロジェクトの本文を作成する。これを、プロジェクト本文 B と呼ぶ。

このプロジェクト本文 B をプロジェクト本文 A と同様にプロジェクト本文のベクトルを求め、その 2 つを比較したものを Figure 4 に示す。

この図表では図中の矢印に示すように、意図に沿って抽象クラスの 7 が推敲によって減少していることを確認出来るとともに、その他の確率分布が全体的に増していることがわかる。一方で結果として望ましく無い変化も確認され、完全に一致するような推敲を行うことは難しいが、効果を検証したい抽象クラスの番号群に沿うように、必要に応じて推敲のプロセスを重ねることで、検証に用いるプロジェクト本文 B を作成することが出来る。

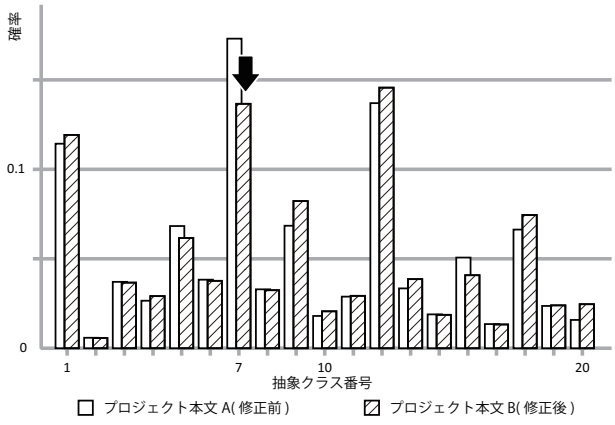

Figure 4: 本文の修正前後の確率分布

この結果については残念ながら本稿執筆中には間に合っていない。

5. まとめ

メディアに専属の編集業務においては、編集者が無意識のうちに文章の特徴のばらつきを押さえる推敲をしていること、またそれにより、SAM を用いた分析でメディアの特性が分かるような文章の確率分布をとらえることができる。また、文章内容の選別に関しても、最終稿から見られるメディアの特徴に合わせた取捨選択を行っていることが分かった。

成立・不成立プロジェクトには文章上の特徴に一定の相違が見られることが分かった。

References

[Inui 00] 乾 裕子, 岡田 直之 : 長い文は常にわかりにくいか? : わかりにくさの要因とその依存関係, 情報処理学会研究報告, 2000.

[Kameya 05] Kameya, Y. and Sato, T.: Computation of probabilistic relationship between concepts and their attributes using a statistical analysis of Japanese corpora., Proceedings of Symposium on Large-scale Knowledge Resources (LKR2005), 2005.

[Terai 09] 寺井あすか, 中川正宣 : 特徴間の相互作用を持つ比喩理解の計算モデル:日本語コーパスの統計解析 を用いて, 日本認知科学会, 2009.