

文書の比較に基づく記述対象の評価

理学専攻・情報科学コース 坂梨 優 (指導教員: 小林 一郎)

1 はじめに

近年, ブログや口コミサイトなどの CGM の増加に伴い, 大量の情報や意見があふれている. それらの情報は有益であるが, その数は膨大で全てを読み比べることは不可能であり, 個人が必要な情報を取捨選択していく必要がある. 本研究では, 表層的な特徴に基づく類似文判定による比較と, それに潜在的な特徴に基づく分類を組み合わせることで, 大量文書の比較方法を提案する. 前者では, 同一の事象を説明する複数の文書を比較し, それぞれの情報源の観点を把握しながら, その内容を正確に捉える事ができる手法の構築を目的とする. 後者では, 潜在トピックでレビュー文の分類を行った後に, 同一トピック内で商品ごとに文を分類し, さらに文の持つ特徴に基づいて, 比較する文同士を決定する比較手法を提案する.

2 表層的な特徴に基づく文書比較

2.1 概要

1 つの事象について報告する際, 複数の情報源により異なる観察視点が存在し情報伝達に差異が生じる. 1 つの情報源のみを参考にするとその情報源の影響を強く受けるため, 受け取った情報に偏りが生じる. そこで色々な視点で述べられた情報を比較し, 文書間の差異を見ることを目的とする.

2.2 提案比較手法

文 A, B の類似度として, Jaccard 係数によって求められるそれぞれを構成する単語集合の類似度に WordNet[1] による語の類義度を加えたものを Jaccard+WordNet とし, 式 (1) で定義する.

$$Jaccard + WordNet = \frac{\sum_{a \in A, b \in B} sim(a, b)}{|A \cup B|} \quad (1)$$

2.3 実験

2.3.1 実験仕様

比較対象テキストとして, 朝日新聞と読売新聞の 2010 年 8 月 26 日「民主党代表選」, 2011 年 2 月 3 日「鳥インフルエンザ」, 朝日新聞と日経新聞の 2011 年 2 月 4 日「新燃岳の噴火」の記事を用いる.

上記 3 つのテキストデータセットに対して, Tri-gram, Jaccard 係数, Jaccard+WordNet の 3 つの手法を用いた実験を行った. 類似文判定における正当性の評価については, 提案手法によって得られる結果と人が予め作成した正解データとの比較を行うことにより検証した.

2.3.2 類似文の対応関係抽出処理

文書 D_1 と文書 D_2 における類似文を判定する際, 通常, 閾値による判定が必要となる. 本研究では, 閾値による問題点を解決するため, 文書 D_1 と文書 D_2 の双方から見たクロスチェックをすることにより, 類似文の抽出を行う. クロスチェックによる対応文抽出の手順は以下のとおりである.

step1. 文書 D_1 中のそれぞれの文に対して最も類似する, 文書 D_2 中の 1 文を抽出する.

step2. 文書 D_2 中のそれぞれの文に対して最も類似する, 文書 D_1 の 1 文を抽出する.

step3. 双方ともに順位が 1 位のものをそれぞれの文書の中で真に類似している文として採用し, それ以外のものは削除する.

2.3.3 結果と考察

表 1 に 2011 年 2 月 3 日の鳥インフルエンザの記事に対する Jaccard+WordNet の実験結果を示す.

表 1: Jaccard+WordNet による実験結果

文	朝日新聞	文	読売新聞	類似度	判定	トピック
1	大分県は 2 日, 大分市宮尾の養鶏場で, 採卵鶏 3 8 羽が死に, 鳥インフルエンザの遺伝子検査で感染力の強い高病原性ウイルス (H5 亜型) が確認された と発表された。	1	高病原性鳥インフルエンザ問題で, 農林水産省と大分県は 2 日, 大分市の養鶏場で感染を確認したと発表した。	0.47		発表
2	県は, この養鶏場から半径 10 キロ内を移動制限区域とし, 飼育中の約 8 100 羽の殺処分を始めた。	8	同県はこの養鶏場で飼う約 8 100 羽を殺処分する。	0.38		決定
3	今回の感染は, 国内の養鶏場では今季 1 1 例目となる。	3	全国で 1 1 例目となった。	0.40		件数
4	県によると, 2 日午後 2 時 20 分ごろ, 養鶏場から「鶏がたくさん死んでいる」と連絡があった。	4	同省によると, 同日午後, 食用の卵を出荷する採卵用の養鶏場から「鶏が前日の 2 倍以上死んでいる」と同県に連絡があった。	0.34		通報
5	遺伝子検査で, 死んだ鶏 6 羽中 5 羽, 同じ鶏舎の 5 羽中 4 羽で感染が確認された。	7	遺伝子検査で H5 型のウイルスが検出された。	0.23		確認
6	飼育されている採卵鶏には, 産卵数が少なくなるなどの症状があるという。					症状
7	半径 10 キロ内には臼杵市, 豊後大野市, 津久見市の一部が入るが, 養鶏場は大分, 臼杵両市の 11 力所。					地理
8	計 3 2 万 2 6 10 羽がいる。	5	1 日は 1 2 羽だったが, 2 日は 3 8 羽が死んでいた。	0.36	x	数量
9	内訳は, 採卵鶏が 7 力所で約 2 4 万 1 200 羽, 肉用鶏が 2 力所で約 8 万 1 200 羽, 自家用が 2 力所で約 2 10 羽となっている。					内訳
10	大分県では 2004 年 2 月, 九重町で飼育されていたチャボが高病原性鳥インフルエンザ (H5N1 型) に感染した事例がある。	11	昨年 11 月以降, 鳥根, 宮崎, 鹿児島, 愛知県の計 10 養鶏場で高病原性鳥インフルエンザの感染が確認されている。	0.31	x	病名
		2	この冬の家畜では, 大分県で初めて。			記録
		6	県で死んだ鶏を含む 11 羽を簡易検査したところ, 8 羽で陽性反応。		x	確認
		9	また, 半径 10 キロ圏内の鶏や卵の移動を禁止した。		x	決定
		10	同圏内には, 11 戸の養鶏場があり, 約 3 2 万羽が飼育されているという。		x	数量

3 つの手法のうち, 提案手法である Jaccard+WordNet がどの記事においても精度が高いという結果が得られた. 朝日新聞の文 6, 読売新聞の文 2 にて互いに対応する文がないことが示せており, それぞれの記事独自の内容を抽出できた.

3 潜在的意味に基づく文書比較

3.1 概要

ユーザーの視点から口コミ文書を解釈するため, 商品が持つ属性を用いてトピック抽出を行い, そのトピックの下で比較を行う.

表 3 : 制約「乾燥」のトピックに分類された商品ごとの口コミ文

商品 A	商品 B	商品 C	商品 D
<ul style="list-style-type: none"> ●肌が敏感なため、普通の口紅はすぐに痒くなった。皮がべろっと剥けたり；普段は、低刺激なグロスやユリアージュのリップクリーム（無香料）を使っています。 ●ただ冬だし、唇が乾燥しやすいから、リップベース（私は Freeplus のリップ）やリップコート（エテュセの超保湿グロス）を塗らなないと、唇のしわが少し気になってしまいます。 ●リップクリームを下地に付けても厳しい（すぐにリップクリーム塗らなくなる）付けたら皮がむけたり、唇が痒くなったり。 	<ul style="list-style-type: none"> ●他のリップグロスは必ず塗る前にリップクリームを使ってからでないと荒れたり唇がしわしわになったりしたのですがこれは何も塗らずに付けても荒れないむしろリップクリームがわりになってくれるくらい保湿してくれます。 ●私の場合、を塗った瞬間は良いけど、しばらく経つと皮がめくれてきたり、端の方に白い塊が出て来たり...でも、こののは全くそんな事ないんです ●元々唇が乾燥しやすいのですがこのは時間がたつても唇の皮が剥けることもなく潤って本当に助かっています。 	<ul style="list-style-type: none"> ●いつも口紅をつけるときはリップクリームをつけてからじゃないと荒れてしまうのですが、これはリップクリームをつけなくても荒れません。 ●それにすぐ唇が弱くて、夕方になるとリップクリームだけになってしまいう私でも、一日中リップなしで使えるくらい潤ってました。 ●口紅は皮剥けするし、縦皺ができてしまうのに、これは塗りやすく時間もたつてもぶっくらうるうです。 	<ul style="list-style-type: none"> ●ちなみに今のところは、朝や乾燥が気になってきたときに唇をオイルマッサージ マークスアンドウェブのリップでマメに保湿、で、ケアをして、今のところトラブルなく使用できています。 ●すぐに唇にかゆみが...気にせず使ったら皮がむけまくりましたーそれ以来、唇がかなり過敏になってしまい、よく皮がむけるようになってしまいました。 ●時間がたつと少し唇がひりひりしてしまうのが気になりますが、我慢できないほどではないので、唇のケアをしっかりして、この口紅を使いたいと思います。

表 4 : Jaccard+WordNet+口コミ感謝件数の重みでの比較

商品 A で気になった文：素の唇の色がやや赤いので、コンシーラーで色を消すと程よく発色。											
N _o	商品 B	加重	件数	N _o	商品 C	加重	件数	N _o	商品 D	加重	件数
9	●色がクリアなタイプということもあり若干地の色が透けるので、地の色が濃い方はあまり発色しないかもですね。	0.35	3	17	●こちらのピンクは、唇の色に馴染む自然なお色。	0.30	1	19	●コンシーラーのように唇の色を消してしまうと思いつ、唇だけ浮いてしまうということはないです。	0.50	0
4	●見た目はだいたい濃い色のライラックという感じですが、このグロス、ホントいい意味で裏切ってくれる発色をします。	0.35	8	8	●元々の唇が赤いので程よいピンクになりました	0.25	0	15	●もとも唇の赤い私にとっては良く発色するので派手に感じます。	0.44	7
10	●もとも唇の色に近くなり（ほんのり赤い感じ）ぼけてりときれない唇になります。	0.32	1	5	●艶は出ますが、色がもとの唇の色がでます。	0.23	0	9	●わたしの唇では、濃いめピンクベージュで、肌色がきれいに見えます。	0.42	12

3.1.1 ディリクレ分布を用いた LDA

LDA[2] を利用し、制約を組み込むことで潜在的なトピックの分類を行うために、ディリクレ分布 [3] を用いる。ディリクレ分布とはディリクレ分布を階層化したものであり、これにより、LDA により同じトピックに入る単語を制御することが可能となる。制約知識は、消費者が商品を選ぶ視点に基づき人手により用意し構成する。与える制約が本文中にない場合、日本語 WordNet により制約単語と文書中の単語との類似度を測り、用意した制約単語との類似度が最も高い語を本文中から探し、置き換えて制約単語とする。

3.1.2 口コミ感謝件数を考慮した判定

口コミがどれだけ参考になったかを表す口コミ感謝件数を、2.2 節に示す表層的な類似度を測る Jaccard+WordNet に重みとして加える。口コミ感謝件数の重みは式 (2) と定義する。

$$\text{口コミ感謝件数の重み} = \frac{\text{各口コミの口コミ感謝件数}}{\text{口コミ感謝件数の最大値}} \quad (2)$$

3.2 実験

3.2.1 実験仕様

対象とするレビュー文書に株式会社アイスタイルの化粧品クチコミサイト@cosme の 2010 年 2 月 1 日から 2011 年 1 月 31 日までのレビュー文書を用いる。期間内の上位 20 位以内にランキングされた商品のうち、口紅・グロス・リップライナーのカテゴリに属する 4 商品を比較する。文数は 24037(文書数: 2800)。3.2.2 節に示す 6 つの制約によるトピック加え、潜在的なトピックも抽出できるよう、トピック数を 10 とする。

3.2.2 制約知識

今回与えた制約は表 2 に示す 6 つのグループとなる。

表 2: 制約知識で構成したグループ

トピックのグループ	トピックを構成する単語
色	色, 発色, 肌
ツヤ感	ツヤ, 潤い
乾燥	荒れる, 乾燥, 剥ける, 皮
持ち	持ち, 時間, 食事, 落ちる
ラメ・パール感	ラメ, パール
香り	香り, 匂い

3.3 結果と考察

トピックごとの単語の確率分布から、予め用意した制約知識で同じグループとした語彙がトピックの構成に反映されたことが確認できた。

表 3 に、制約「乾燥」の単語を含むトピックに分類された口コミ文を商品ごとに示す。表 3 を見ると、商品 B, C では保湿力があり、唇が荒れることなく潤い、一方商品 A, D では唇が荒れやすくなるという内容の文がみられた。また「色」グループを構成する単語が表れたトピックでは、商品 A は色が薄く B はほんのり色づき、商品 C は肌の色に馴染む、そして商品 D は発色がよいという内容が示された。ディリクレ分布を用いた LDA により、ある決まった視点でのおおまかなトピック分類ができたと言える。

表 4 には Jaccard+WordNet による類似度に口コミ感謝件数の重みを加えたもので比較したものを示す。表 4 では、口コミを参考になったとする件数を重みとして類似度に加えることで、ある商品の 1 文に対する他の商品での表層的な類似だけでなく、多くの人々が参考になった口コミを優先的に提示した。類似度みの比較と比べ肯定的な意見ばかりでなく、参考になる意見がより提示されるようになった。

4 おわりに

本研究では、レビュー文の決まった視点での分類を行うため、ディリクレ分布を用いた LDA によって、制約知識を組み込んだトピック抽出を行った後、シソーラスを考慮した表層的類似度や、口コミ感謝件数の重みを利用して商品間の比較を行った。

各商品の特徴を考慮して商品間の比較するためには、多数派の意見であるかどうかを踏まえる必要がある。今後は商品ごとに語の共起情報を捉え、各商品の特徴を考慮した情報の提示を行いたい。

参考文献

- [1] <http://nlpwww.nict.go.jp/wn-ja/>
- [2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. : Latent Dirichlet Allocation, *Journal of Machine Learning Research*, 3:993-1002(2003)
- [3] Minka, T. P.: The Dirichlet-tree distribution (Technical Report) <http://research.microsoft.com/en-us/um/people/minka/papers/dirichlet/minka-dirtree.pdf>, 1999.