

ランキング学習を用いた形容詞に基づく概念語の順序付け

岩成 達哉* 吉永 直樹†‡ 豊田 正史† 喜連川 優†§

* 東京大学大学院 情報理工学系研究科 † 東京大学 生産技術研究所

‡ 情報通信研究機構 § 国立情報学研究所

{nari, ynaga, toyoda, kitsure}@tkl.iis.u-tokyo.ac.jp

1 はじめに

私たちは日々の生活の中で、様々な観点で事物に順序付けを行い、利用するものを選択している。例えば、移動手段を選ぶ場合に、遠くにある目的地により早く着きたいのであれば、我々は常識に基づいて徒歩よりも自転車、自転車よりも車、車よりも飛行機を選ぶ。このように、事物の順序関係に関する知識は、我々が意思決定を行う上で基本となる知識であるが、通常の Wikipedia などの知識ベースには必ずしも含まれておらず、計算機が利用・提示することができない。

仁科ら [4] は、このような概念語（徒歩、車、飛行機）を、共通する性質の強さ（速い）に基づいて順序付けするタスクを提起し、大規模なウェブテキストから、概念語と形容詞の文内共起や係り受けの出現頻度に基づき順序付けを行う手法を提案した。この手法は、ソーシャルメディアユーザの投稿から事物の順序に関する知識が導けるか、という点で科学的観点から興味深いものであるだけでなく、例えば使い勝手の良い携帯電話の機種を調べるなど、実用的観点からも有用である。

本研究では、概念語の順序付けタスクを解くために、従来手法で用いられていた概念語と形容詞の共起や係り受けなどの手がかりだけでなく、比喻表現や比較表現といった、より多様な手がかりをソーシャルメディアテキストから抽出し、これらを教師ありランキング学習により素性として同時に考慮して順序付けを試みる。

実験では、我々の研究室の有する 8 年分の日本語ブログ記事から手がかりを集め、実際に複数の概念語集合と形容詞の組に対して順序付けを行い、提案手法が出力する順序と人手による順序付けで得られた正解との相関係数を求めることで手法の有用性を確かめる。

2 概念語の順序付けタスク

本稿では上記のように、複数の概念語をある形容詞で表される共通の性質の強さで順序付けするタスクを取り上げる。このタスクにおいては、概念語集合が調べたい性質の形容詞と共に与えられ、順序付けされた概念語リストが出力として得られる。例えば、概念語の集合として { イヌ, ゾウ, ネズミ, クジラ } が与えられ、形容詞として「大きい」が与えられた場合は、期待される出力はクジラ > ゾウ > イヌ > ネズミである。

タスクを定義する上で問題となるのは、このような一般的な順序付けの正解をどのように与えるかという点である。本稿では、複数の人手による順序付けに対して、スピアマンの順位相関係数 ρ の平均が最大となる概念語の順序を正解とする。食べ物の美味しさなど、主観的評価を伴う順序付けについては必ずしも絶対的な正解は存在しないが、これによって得られた順序は、人による順序付けの視点を最大公約数的に反映したものととなる。

3 関連研究

仁科ら [4] は、大規模なウェブテキストにおいて、形容詞の否定や対義語を考慮しながら、概念語と形容詞の同一文内共起、係り受けを調べ、それらの値の PMI (Point-wise Mutual Information) を用いて、順序付けを行っている。これらの手がかりはそれぞれ独立に利用されており、情報を組み合わせて使えていない。一方、本研究では、教師ありランキング学習を導入することにより、これらの手がかりを組み合わせることを可能とした。

倉島ら [5] は「良さ」という性質に基いて、具体物を順序付けするタスクに取り組んでいる。この研究では、

比較情報を元に有向グラフを作成し、それぞれのノードの評価値を利用して順序付けを行っている。これは、「良さ」という形容詞についてのみの順序付けになっているため、「大きさ」「美味しさ」など、本研究が扱えるその他の形容詞に基づく順序付けには適用できない。

Yoshinaga と Torisawa [3] は事物の属性/値をウェブの表や箇条書きから獲得する手法を提案しており、数値表現可能な観点については、彼らの手法を概念語の順序付けに用いることが可能であると考えられる。これと関連し、Takamura と Tsujii [1] は、テキストから集めた複数の手がかりを元に、物体の数値に基づく特徴(大きさや長さなど)を推定するタスクを解いている。この手法では、実際の値を探すだけでなく、他の物との比較表現などの手がかりを集め、数値表現可能な特徴に基いて順序付けを行い、大きさを推定する。これらの研究に対し、順序付けという観点でみると、本タスクの方がより一般的なタスクであるといえる。

4 提案手法

本研究では、我々が概念語に対して持っている印象は、明示的あるいは暗黙的に我々が書く文章に影響していると仮定し、ソーシャルメディアのテキストを解析することにより、4.1 節に示す様々な手がかりを集める。その後、集めた手がかりを教師あり学習に用い、形容詞に基づく概念語の順序を導く(4.2 節)。

4.1 順序付けに用いる手がかり

本手法では、仁科らが提案した同一文内共起、係り受けに加えて、比喩表現と比較表現を含めた4つの手がかりを順序付けに利用する^{*1}。最初の3つは、暗黙的に性質の強さを示すが、4つ目の手がかりは、明示的に2つの概念語間の順序関係を示すものである。

概念語と形容詞の同一文内共起 概念語がある性質を持つのであれば、その性質を表す形容詞は、概念語と頻繁に共起すると考えられる(例: ゾウ はなんて 大きい んだ)。

概念語と形容詞の係り受け 同一文内共起は常にその概念語と形容詞の関係を示すとは限らないため(例: アリ は ゾウ に踏まれないくらい 小さい)、概念語が形容詞に直接係る場合は、係り受けが共起よりも有効な手がかりになると考えられる。

^{*1} 実験では、日本語のテキストを用いたが、ここで示す手がかりは言語に非依存なものであり、他言語に拡張が可能である。

直喩表現 概念語はその性質の強さが際立って強い場合、直喩として用いられることがある(例: それは 雪 のように 白い)。実験では、このような直喩表現のパターンを列挙し、それらにマッチするものを数え上げた。

比較表現 2つの概念語間の比較表現が現れた場合、それは順序付けに関する大きな手がかりとなりうる(例: ゾウ は イヌ よりも 大きい)。実験では、直喩表現と同様に、比較表現のパターンを列挙し、それらにマッチするものを数え上げた。

本研究では、与えられた形容詞の対義語を1つ選び、与えられた形容詞と同様に上記の手がかりを集めることで、性質に関する負の極性の情報も集める。さらに、それらの形容詞に否定が付属する場合は、それぞれ反対の極性の手がかりとして数え上げている(例: 「イヌは大きくない」は「イヌは小さい」と同様と考える)^{*2}。

4.2 多様な手がかりを用いた教師ありランキング学習による概念語の順序付け

本研究では、4.1 節の多様な手がかりを順序付けに利用するために、教師ありランキング学習法の1つであるランキング SVM^{*3} を利用した。ランキング SVM では、素性(手がかり)の値を元に、概念語集合のそれぞれの2つの要素の組について、正解と比較した時の順序関係の誤差が最小となるように学習を行う。

本稿では、概念語と形容詞の組を変えた場合にも、正しく学習ができるようにそれぞれの手がかりの値を正規化し、素性の値としている。4.1 節で取り上げた手がかりのうち、最初の3つについては、与えられた概念語と形容詞、与えられた概念語と形容詞の対義語のそれぞれの PMI を式(1)で計算する。

$$\text{PMI}(x, y) = \log \frac{p(x, y)}{p(x)p(y)} \quad (1)$$

その後、仁科らに倣い、式(2)で表される Turney ら [2] の評価極性の分類によって、正と負の両方の極性の

^{*2} 直喩表現においては、付属する否定が、その性質の否定を表すわけではないため、否定が付属するかどうかは調べていない(例: 「それは雪みたいに白くない」は雪が白いことを表す)。

^{*3} https://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html

手がかりを 1 つにまとめた値を素性とする*4.

$$\begin{aligned} & \text{PMI}_{\text{pos}} - \text{PMI}_{\text{neg}} \\ &= \text{PMI}(\text{概念語, 肯定 (形容詞 or 対義語の否定)}) (2) \\ & - \text{PMI}(\text{概念語, 否定 (対義語 or 形容詞の否定)}) \end{aligned}$$

4 つ目の手がかりは, 2 つの概念語の間の比較表現である. この手がかりは, ある概念語が, 与えられた他のすべての概念語のうち何個に対して, コーパス内で 1 度でも優っていると記述されたかを素性の値とする. また, 反対の極性の素性として, 与えられた他のすべての概念語のうち, 何個に対して 1 度でも劣っていると記述されたかを同様に利用する. これらの値は, 与えられた概念語の数によって正規化する.

5 評価

5.1 データ

解析するソーシャルメディアテキストとして, 本稿では, 我々が収集した 200 万記事に及び 2005 年から 2013 年の日本語のブログ記事を利用した. ブログ記事は, 20 億程度の文で, 100 万人を超えるユーザによって書かれたものである.

実験のための概念語の集合は, 仁科ら [4] のデータセットにしたがって選んだ. 形容詞は, テキストに含まれる頻度の高いものから「ない」「すごい」などの, 性質の程度を表さないものを除いて選ばれている. 概念語は, それぞれの形容詞に高頻度で係る概念語を 1 つ取り出し, 日本語 WordNet *5 と NLTK (Natural Language Toolkit) *6 の WordNet モジュールを利用することで最大で 10 個までの概念語を集めている.

正解データの作成では, 4 人の被験者にそれらの概念語の集合と形容詞の組を提示し, 順序付けを行ってもらった後, 考えられる概念語のすべての順序の中で, 4 つの人手による順序付けとのスピアマンの順位相関係数 ρ の平均が最も大きくなる順序付けを正解とした. 得られた正解は表 1 に示した.

評価においては, 表 1 に示した 8 個の概念語の集合と形容詞の組から 1 つをテストデータとし, 他のデータを教師データとする, 8-fold 交差検定を行い, 提案手法から得られた順序付けと正解データとのスピアマン

の順位相関係数 ρ を計算した. また, ランキング SVM には線形カーネルを用い, 各テストにおけるランキング SVM のハイパーパラメータ C は教師データの中で 7-fold 交差検定を行い, 最適な値を求めた.

5.2 結果

結果は, 表 2 のとおりである. ここでは, ベースラインとして, 仁科ら [4] の手法において, スピアマンの順位相関係数の平均が最も良くなる対義語や形容詞の否定を考慮した係り受けを手がかりとする手法を用いている. 結果として, 食べ物 (美味しい) 以外において, 提案手法はベースラインを上回る結果となっている.

食べ物 (美味しい) の相関係数が低くなったことについて, 人との比較においても他の項目に比べて, 相関係数の平均が低くなっていることがわかる. したがって, ソーシャルメディアテキスト中でもユーザの意見が割れ, その結果正しく学習ができていないことが 1 つの要因として考えられる. 表 1 の食べ物 (美味しい) に関する順序付けを見ると, 日常的に食べられるものが提案手法では上位になっていることが伺えるため, 共起による情報が強く影響し, ベースラインよりも劣る結果となったと推察できる.

また, 他の順序付けに関しては, 表 1 の食べ物 (安い) の順序付けにおいて, 人手では「高い」と順序付けられた「寿司」が, 提案した手法では一番上に来る結果になっていたが, 「自分と夫は夕方セールでちょっと安くなった普通のお寿司を買って食べました」のように限定的な条件で「安い」とよく共起することが原因であると考えられる.

6 おわりに

本研究では, 形容詞 (例: 大きい) が表す性質を元に, 与えられた概念語の集合 (例: ソウ, イヌ, ネズミ) を順序付けるタスクに取り組んだ. 提案手法では, 大量のソーシャルメディアテキストを用いて, そこに現れる概念語と形容詞の同一文内共起や係り受け, 比較表現などを数えあげて手がかりとし, ランキング SVM による教師あり学習を用いることで順序付けを行った. また, 評価では, 人手によって作成した正解データと比較をすることで, 手法の有用性を確かめた. 結果として, 人手によって作成したもので, 人の間でもばらつきが少ないものに対しては, ベースラインを上回る結果が得られ, 手法の有用性が確かめられた.

*4 PMI の計算においては, 対数をとる際に, 頻度に 1 を加算するスムージングによって 0 頻度のものにも対応している.

*5 <http://nlpwww.nict.go.jp/wn-ja/>

*6 <http://nltk.org/>

表1 形容詞と概念語集合の組み合わせ．概念語は各手法の出力順に並んでいる．

形容詞	対義語	手法	概念語集合 (出力順．左ほど程度が強い．)
大きい	小さい	正解	クジラ, キリン, ソウ, クマ, ウシ, ウマ, イヌ, サル, ネコ, ネズミ
		ベースライン	ゾウ, ウシ, クジラ, ウマ, キリン, クマ, イヌ, ネコ, ネズミ, サル
		提案手法	ゾウ, クマ, クジラ, キリン, ウシ, イヌ, ウマ, ネコ, サル, ネズミ
安い	高い	正解	ハンバーガー, パン, 焼きそば, チャーハン, カレー, パスタ, ピザ, ステーキ, 寿司
		ベースライン	焼きそば, ステーキ, 寿司, チャーハン, パスタ, ハンバーガー, パン, カレー, ピザ
		提案手法	寿司, ハンバーガー, パン, カレー, チャーハン, 焼きそば, ステーキ, ピザ, パスタ
可愛い	恐い	正解	ネコ, ウサギ, ハムスター, リス, イヌ, ヒツジ, カメ, ウマ, トカゲ, サル
		ベースライン	ハムスター, リス, カメ, ウサギ, ネコ, ヒツジ, サル, トカゲ, イヌ, ウマ
		提案手法	ウサギ, ハムスター, リス, ネコ, サル, ヒツジ, カメ, イヌ, トカゲ, ウマ
美味しい	不味い	正解	寿司, ステーキ, ピザ, カレー, ハンバーガー, 焼きそば, パスタ, チャーハン, パン
		ベースライン	ピザ, パスタ, パン, ステーキ, 寿司, カレー, チャーハン, 焼きそば, ハンバーガー
		提案手法	パン, パスタ, カレー, ピザ, ステーキ, 焼きそば, 寿司, チャーハン, ハンバーガー
速い	遅い	正解	飛行機, 新幹線, ヘリコプター, 電車, 自動車, 船, スクーター, 自転車
		ベースライン	新幹線, 船, 自動車, ヘリコプター, 自転車, 飛行機, スクーター, 電車
		提案手法	新幹線, 飛行機, 自動車, 船, 自転車, 電車, ヘリコプター, スクーター
重い	軽い	正解	タンス, ベッド, ソファ, 机, じゅうたん, 椅子, カーテン, 電気スタンド
		ベースライン	じゅうたん, カーテン, 机, タンス, ベッド, 椅子, ソファ, 電気スタンド
		提案手法	カーテン, タンス, ベッド, じゅうたん, 机, ソファ, 電気スタンド, 椅子
重い	軽い	正解	冷蔵庫, 洗濯機, 電子レンジ, ストープ, 掃除機, ホットプレート, トースター, コーヒーメーカー, アイロン
		ベースライン	洗濯機, 冷蔵庫, アイロン, 掃除機, ストープ, 電子レンジ, ホットプレート, コーヒーメーカー, トースター
		提案手法	冷蔵庫, 洗濯機, 電子レンジ, アイロン, ストープ, コーヒーメーカー, 掃除機, ホットプレート, トースター
甘い	酸っぱい	正解	マンゴー, モモ, イチゴ, ミカン, ナシ, リンゴ, パイナップル, カキ, グレープフルーツ, レモン
		ベースライン	カキ, マンゴー, モモ, ナシ, パイナップル, リンゴ, イチゴ, ミカン, グレープフルーツ, レモン
		提案手法	マンゴー, パイナップル, リンゴ, イチゴ, モモ, ナシ, ミカン, カキ, グレープフルーツ, レモン

表2 実験結果．正解データと4つの人手による順序付けとの相関係数 ρ の平均を参考に示した．

概念語 (形容詞)	ρ の平均	ベースライン	提案手法
動物 (大きい)	0.975	0.758	0.879
食べ物 (安い)	0.933	-0.167	0.267
動物 (可愛い)	0.719	0.552	0.697
食べ物 (美味しい)	0.653	0.083	-0.267
乗り物 (速い)	0.939	0.143	0.524
家具 (重い)	0.949	0.095	0.357
家電 (重い)	0.942	0.550	0.650
果物 (甘い)	0.854	0.467	0.673
平均	0.882	0.310	0.472

今後の課題として、クラウドソーシングなどを用いて、より大規模な形容詞と概念語集合の組を用意し、それらに基づいた評価を行う点が挙げられる、また、本手法では、概念語とある形容詞（あるいはその対義語）の関係のみを調べるため、その共起頻度が十分でない。これを解決するためには、与えられた形容詞に似た意味の言葉や関連する言葉（例：「大きい」は「重い」との関連が強い）についても共起を数え上げるなど、より多くの手がかりを集めることが重要だと言える。これは、概念語についても同様であり、概念語の下位語を検索の対象に含めるなどの工夫ができると考えられる。

既存研究では複数の手がかりを独立に用いることしかできなかったが、本研究によって、様々な手がかりを組み合わせるようになった。これにより、多様な手がかりと手がかり間の関係を手法に加える事ができるようになるため、手法の改善がより容易になったといえる。

参考文献

- [1] H. Takamura and J. Tsujii. Estimating Numerical Attributes by Bringing Together Fragmentary Clues. In *Proc. HLT-NAACL*, pp. 1305–1310, 2015.
- [2] P. D. Turney. Thumbs Up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proc. ACL*, pp. 417–424, 2002.
- [3] N. Yoshinaga and K. Torisawa. Open-domain attribute-value acquisition from semi-structured texts. In *Proc. ISWC Workshop, OntoLex*, pp. 55–66, 2007.
- [4] 仁科俊晴, 鍛冶伸裕, 吉永直樹, 豊田正史. 対義形容詞対との相互情報量を利用した概念語の順序付け. *IPSJ SIG NL*, Vol. 2013, No. 8, pp. 1–7, nov 2013.
- [5] 倉島健, 別所克人, 戸田浩之, 内山俊郎, 片岡良治, 奥雅博. 比較評価情報に基づくランキング手法. *DBSJ*, Vol. 6, No. 1, pp. 1–4, 2007.