5K-02

口コミに基づくホテル推薦のための辞書作成法

山本　隼也†　　福原　楓　　田熊　浩二‡　　亀井　清華‡　　藤田　聡‡

広島大学工学部†　　広島大学大学院工学研究科‡

1. はじめに

近年，旅行や仕事などでホテルを探す際に，ホテル予約サイトと呼ばれるものを多くの人が利用している．ホテル予約サイトでは，各ホテルの窓口を統合し，ホテルの仲介代行をインターネット上で行っている．多くのホテルが登録されているサイト上での探索は，ユーザの嗜好や条件に沿ったホテルを見つけ出すことが困難であり，ユーザの利用満足度低下につながっている．

この問題に関して，田熊らは口コミから抽出したユーザの嗜好に基づいてホテルを推薦する手法[1]を提案した．田熊らの手法では，ホテルの口コミに出現する単語をカテゴリ別に分類した単語辞書に基づいて，口コミ投稿者が重視するカテゴリを推定する．この辞書はコーパスのデータから手作業で抽出した 433 個の名詞から構成されており，名詞は 6 つのカテゴリに重複ありで分類されている．

しかし，ここで使用される辞書に登録されている名詞が少ないため，ホテルに対する口コミから投稿者の嗜好を適切に抽出するのは難しい．そこで本研究では，口コミに基づいて自動で単語を追加する辞書作成法を提案する．

2.口コミに基づいた辞書作成法

本手法では口コミコーパスのデータ(2,500,000件)を 1 文単位で使用するため，口コミの文章を句点や改行で区切って分割した 10,703,574 文を用いて新たな辞書を作成する．また，[1]で用いた手作業によって作成された辞書で扱っている単語は全て名詞にのみに限定されているため，追加する単語はすべて名詞のみとする．本稿では[1]の辞書を元辞書と呼ぶ.

2.1. 口コミの文に出現する名詞の抽出と分類

分類に用いたカテゴリは，元辞書に合わせて，(1)サービス(2)立地(3)部屋(4)設備・アメニティ(5)風呂(6)食事の 6 つである．本手法では口コミの文に出現する名詞の抽出，及び，これら６つのカテゴリへの分類を自動で行う．

まず最初に，口コミの文に出現する名詞を抽出するため，形態素解析エンジン『MeCab』[2]を使用する．口コミの文に対して形態素解析を行うことにより，文中に含まれる名詞のみを抽出することが可能である．

次に，抽出した名詞を 6 つのカテゴリへと分類する．抽出した名詞を辞書に登録するためには，その名詞が 6 つのカテゴリのどれに該当するかを判別する必要がある．そこで，既に 6 つのジャンルへと名詞が分類されている元辞書を用いる．本稿では，元辞書に登録済みの単語が現れる文を共起文と呼ぶ．このとき，各カテゴリの登録済み単語との共起文数を，各未登録単語のそのカテゴリへの関連度とする．具体的には，各単語の各カテゴリへの関連度を次のように求める．ある文 A において，ある名詞が元辞書のカテゴリ１に登録されている場合，A はカテゴリ１について言及していると判断する．そして，共起文に含まれるすべての名詞は，言及されていると判断されたカテゴリに関連があるとする．つまりこの場合，A に含まれるすべての名詞がカテゴリ１に関連する．このように各名詞が各カテゴリに関連するとみなされた回数を関連度としてカウントする．もし，異なるカテゴリの名詞が一つの共起文に複数含まれていたり，一つの単語が異なる共起文で異なるカテゴリに関連したりしていても，単純に一つの共起文で各カテゴリへの関連度を１ずつカウントアップする．

そして各共起文において関連度を確認した後，各単語を関連度が最大となるカテゴリへと分類する．また，値が重複する場合は最大となるカテゴリすべてに単語を分類する．この追加する名詞から成る辞書と元辞書を併合させたものを新たな辞書とする．

2.2. 辞書に追加する対象単語

辞書を作成するにあたり，すべての名詞がホテルに関する名詞であるとは限らない．そこで，辞書に追加する対象の名詞を(1)共起文を使用した場合と(2) LDA[3]を使用した場合の 2 パターンを用いて制限する．(1)の場合は，共起文集合中に出現した回数によって制限を設け，閾値以上の回数で出現した名詞のみを辞書に追加する対象とする．また，(2)の場合は，まずトピック数を 30 として LDA を実行する．得られたトピック中から登録済み単語が多く含まれるトピック集合を抽出し，それぞれの名詞において単語出現確率を足し合わせる．そして，共起文と同様に閾値を設け，単語出現確率の総和で降順に並べた名詞一覧から，閾値の個数分だけを辞書に追加する対象とする．

Generating a dictionary for hotel recommendation based on reviews
† Faculty of Engineering, Hiroshima University
‡Graduate School of Engineering, Hiroshima University

3. 評価

3.1. 辞書の単語数

共起文と LDA を用いて名詞を制限した辞書の比較を行うため，単語数が 2000，1500，1000 になるよう考慮して閾値を定めた辞書を作成した．単語数を 1500 となるよう閾値を定めた場合のカテゴリ別単語数を表 1 に示す．表で示す各カテゴリの数字は 2.1 章と対応している．

表 1. 各辞書の単語数

カテゴリ	(1)	(2)	(3)	(4)	(5)	(6)	合計
元辞書	131	60	125	91	73	77	557
共起文	434	132	659	21	93	197	1536
LDA	503	144	614	16	70	156	1503

3.2. 辞書の精度

本手法で作成した辞書を評価するにあたり，対象とする名詞において 3 人にアンケートを行った．アンケートでは，それぞれの名詞が使用されているロコミの文を参考にして，名詞が文中で 6 つのカテゴリのうち，どのカテゴリの内容を意味するものとして使用されやすいかを選択した．また，6 つのジャンルのいずれにもあてはまらない場合があるため，選択肢として分類不可を設けた．辞書の評価には，3 人の回答したアンケートの和集合，つまりいずれか 1 人でも選択したカテゴリを名詞に対する正解とした．このとき，和集合をとって 6 カテゴリに解答が分散した場合，様々なカテゴリの内容を示す名詞であると考え，分類不可とした．

アンケートを用いた辞書の評価を表 2 に示す．精度は，辞書に追加した単語において正解カテゴリに分類された単語の割合を示している．精度は共起文および LDA を用いた辞書の各単語数に関わらず，同等の精度を示している．

しかし，ここでの評価には分類不可が多く含まれており，それらは評価対象外としている．そこで，辞書作成後に改めて各単語の評価を行った．すると，3 割の名詞が分類不可ではなく有用な名詞であることがわかった．辞書全体に対する分類不可の割合は，共起文を用いた辞書に比べて LDA を用いた辞書の割合が高いが，有用な名詞の割合は LDA を用いた辞書よりも共起文を用いた辞書の割合が高くなっている．各辞書に対する割合を表 3 に示し，分類不可を再評価した後の最終的な辞書の評価を表 4 に示す．

表 2. 各辞書の精度評価

	共起	LDA	共起	LDA	共起	LDA
単語数	2000		1500		1000	
精度	82.90	83.83	82.72	82.97	82.88	82.82

表 3. 分類不可および有用な名詞の割合（%）

	共起	LDA	共起	LDA	共起	LDA
単語数	2000		1500		1000	
分類不可の割合	35.5	38.4	34.2	38.1	39.6	41.9
有用な割合	35.0	33.2	34.3	32.3	36.3	34.8

表 4. 各辞書の最終精度評価

	共起	LDA	共起	LDA	共起	LDA
単語数	2000		1500		1000	
精度	92.78	93.24	92.75	93.19	92.96	92.58

3.3. 分類できた文の数

楽天トラベルの実際のロコミ 103,206 件を使用し，元辞書および作成した辞書によって分類できたロコミ文の数を測定した．辞書の作成と同様に 1 文単位でのデータを扱うためロコミを分割した 360,016 文を対象とした．元辞書で分類できた文が全体の 55%であった．本手法で作成した登録単語数 1500 の辞書で分類できた文は，いずれの辞書の場合も全体の 93%となり，それぞれの辞書において，元辞書よりも多くの分類ができるようになった．

4. おわりに

本研究では，ロコミに基づいたホテル推薦のための辞書作成法の提案を行った．そして，アンケートに基づいて，辞書の精度および辞書を適用した推薦手法の精度を示した．これにより，LDA は，ロコミのように短い文章の中に複数のトピックが入っているような場合には，単純に共起回数を見たのと同程度の性能を示すものと考えられる．

謝辞

本研究では，楽天株式会社が提供し，国立情報学研究所が配布している「楽天公開データ」を利用させていただきました．ここに記して謝意を表します．

参考文献

[1] 田熊 浩二，福原 楓，亀井 清華，藤田 聡．ロコミを用いた嗜好抽出に基づくホテル推薦手法．情報処理学会全国大会 2016.

[2] MeCab: Yet Another Part-of-Speech and Morphological Analyzer. http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

[3] David M. Blei, Andrew Y. Ng, Michael I. Jordan, “Latent Dirichlet Allocation”, The Journal of Machine Learning Research, pp. 993-1022, 2003.