自動抽出した利用者の視点によるレビュー要約

田窪直人†　鈴木良弥‡
山梨大学工学部コンピュータメディア工学科
t08kf020@yamanashi.ac.jp†　ysuzuki@yamanashi.ac.jp‡

1. はじめに

インターネットを利用した商品の購入が一般的になり，それに伴って大量のレビューを閲覧できる環境になった.『楽天トラベル』[1]の宿泊施設のレビューもその1つである. このレビューは宿泊施設をユーザーが利用した際にサービス，立地，部屋，設備・アメニティ，風呂，食事，そして総合的な7項目についての5段階の評価とその施設についての感想の文章からなる. 2012年1月現在，26,678件の国内宿泊施設が登録されており，それぞれにレビューが存在する. その数は数十件から多いものでは千件以上にものぼる. しかし，利用施設を選択するために全てのレビューを閲覧することは現実的に不可能である. そこで本稿では重要単語(キーワード)に該当する全てのレビューに対してキーワードに関する要約文を作成することを目標としている.

2. 関連研究

重要語を用いた要約文の作成に関する研究として[2]の研究を参考にして行った. この研究ではSupport Vector Machine(SVM)を用いて15の素性による文書中の文の重要文節を決定し，係り受け構造を用いたアルゴリズムを用いて要約を試みている. しかし，この研究ではデータに文語調で書かれた新聞記事を用いており，レビューは口語調で書かれているのでレビュー要約にそのまま利用することは難しい. 本研究ではWeb上での利用を考慮して重要文節を要約文につき1つの単語とし，より結果が見やすい手法を用いた. またレビューについてはアスペクトという評価視点を用いてさらに評点を加えた研究[3]も存在する.

3. 提案手法の概要

システムの流れを図1に示す. なお，ここでのキーワードとはレビュー内に出現する一般名詞を出現回数順に並べた上位10件の単語を指す. これによってその宿泊施設の重要と思われる単語をその施設のレビュー全てから抽出することができる.

ユーザーはシステムに施設番号を入力することでその施設についての特徴を現す複数の単語(キーワードリスト)を取得することができる. さらにユーザーはキーワードリストから1つをキーワードとして選択することで，そのキーワードから情報を検索して要約を行う.

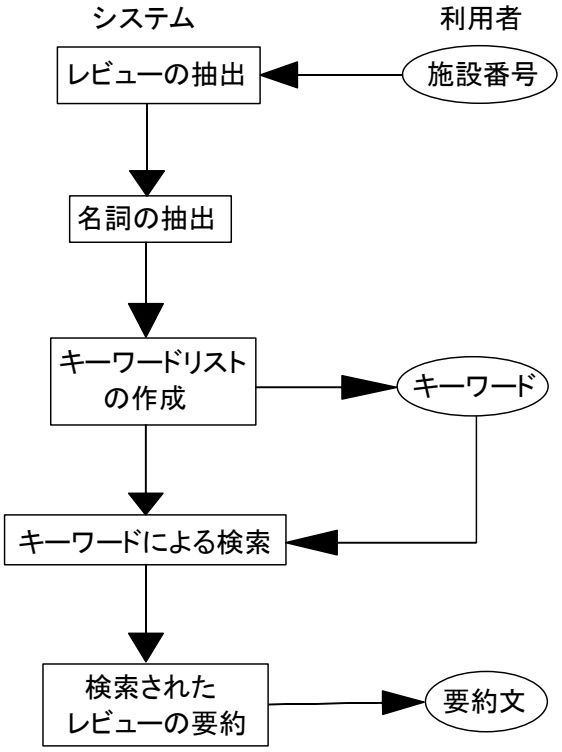

図1. システムの流れ

3.1. ユーザー投稿レビューの抽出

表1. 扱ったデータの形式

項目	説明
施設番号	施設でユニークな数字
ユーザー投稿本文	投稿本文の文字列
投稿番号	投稿でユニークな数字
分類	「感情・情報」「苦情」などの文字列
プランID	プランIDの数字(空欄あり)
プランタイトル	プランタイトル文字列(空欄あり)
部屋種類	「s」「s1」などの文字列
部屋名前	部屋の名前の文字列
施設回答本文	投稿に対する施設側からのコメント(空欄あり)

Copyright(C) 2012 The Association for Natural Language Processing.
All Rights Reserved

要約の手法としてまず表 1 の形式をタブで区切ってまとめられたテキストファイルからユーザー投稿本文を正規表現によって抽出する. 尚, データは 2010 年に楽天技術研究所[4]にて公開された楽天トラベルのデータを使用する.

3.2. キーワードのための名詞の抽出

その後文章を日本語係り受け解析システム CaboCha[5]によって図 2 のように構文解析して一般名詞を抽出する. また, “*”の後ろにある数字はそれぞれ文節番号, 係り先番号+係り種別, 主辞/機械語位置, 係りやすさのスコアになっている. 本稿では文節番号と係り先番号の係り受け情報を用いることによって後の要約文作成を行う.

（例文）フロントやスタッフの対応はとてもよく、
* 0 1D 0/1 1.05936835
フロント　フロント　フロント　名詞-一般　O
や　ヤ　や　助詞-並立助詞　O
* 1 2D 0/1 1.86707095
スタッフ　スタッフ　スタッフ　名詞-一般　O
の　ノ　の　助詞-連体化　O
* 2 4D 0/1 1.66298873
対応　タイオウ　対応　名詞-サ変接続　O
は　ハ　は　助詞-係助詞　O
* 3 4D 0/0 0.00000000
とても トテモ とても 副詞-助詞類接続　O
* 4 -1O 0/0 0.00000000
よく　ヨク　よい　形容詞-自立　形容詞・アウオ段　連用テ接続　O
、　、　、　記号-読点　O

図 2. CaboCha の解析結果

3.3. キーワードリストの作成

前項の結果より名詞を抽出したものを表 2 のように出現回数順に並べる. ここで生成された名詞はその施設における特徴を表すキーワードとなり, その中の 1 つを利用者が選択することによって利用者に興味のある項目を検索することができる. 施設番号 416 に登場した名詞を出現回数でソートしたものを表 2 に示す. キーワードにはその宿泊施設の特色を表す単語を選択する必要があるので『部屋』,『フロント』,『ホテル』など全ての施設に頻出する単語を削除する. 今回は全ての施設における出現回数の多い名詞の中から上位 10 件に該当する単語を削除する. 本稿では『冷蔵庫』をキーワードとして要約手法の説明を行う.

表 2. 施設番号 416 と全体の施設での出現回数の多い名詞

	416の名詞	出現回数	全体の名詞	出現回数
1	部屋	14	部屋	149151
2	フロント	14	ホテル	109057
3	ホテル	13	プラン	82442
4	プラン	8	風呂	54842
5	料金	6	フロント	44523
6	冷蔵庫	6	駅	38665
7	駅	6	立地	38505
8	シングル	5	お部屋	32656
9	タクシー	5	値段	27607
10	繁華	5	ルーム	26629

3.4. レビューのキーワードによる検索

前項で作成したキーワードリストから利用者はキーワードを 1 つ選択してユーザー投稿レビューよりパターンマッチングを用いた検索を行う[6]. これによって, キーワードを含むレビューのみに絞り込むことで, その宿泊施設について全てのレビューを用いる場合よりその施設についての特徴が含まれるレビューを抽出し, 要約文を作成することができる.

3.5. 検索されたレビューの要約

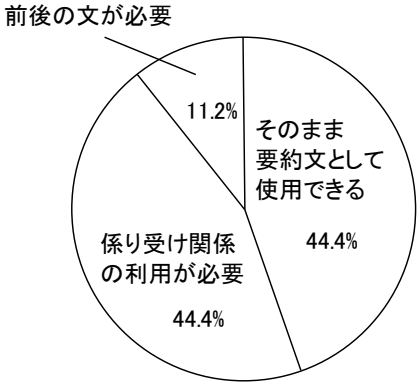

図 3.『冷蔵庫』について要約方法の結果

この項では選択されたキーワードを含む文章より CaboCha の係り受け関係を用いて簡易的な要約文を作成する. レビューは口語調で, 文法的に間違った文章も含まれるため, CaboCha での構文解析が正常に行われない場合がある. よって後の要約で係り受け関係を用いるなどの工夫が必要である.『冷蔵庫』というキーワードを全ての宿泊施設のレビューについて人手で要約の方法を調べたところ図 3 のような結果になった. レビューの文章がそのまま要約文として使

Copyright(C) 2012 The Association for Natural Language Processing.
All Rights Reserved

用できる文が 4,570 文中 2,031 文と 5 割に満たなかった. そして残りの 2,500 件あまりの 8 割ほどが文頭に来る接続詞やキーワードの出現位置などの理由によって要約文にする際に係り受け関係を用いた要約が必要であった. さらに残りの 2 割ほどはキーワードを含む文章のみでは要約文が作成できず前後の文を用いて要約文を作成する必要がある.

3.5.1 先祖子孫関係を用いた要約

まず, 係り受け関係を用いた要約を始めるにあたって適当な文を例にして[2]のアルゴリズムを基にして図 4 のようなキーワードの先祖子孫関係の文節を要約文とした. この要約ではキーワードが主語でなく目的語である場合, 特に主語がキーワードに直接係らない場合は元の文から要約文にする際に主語が無くなってしまうので意味が通じない文になってしまう. また, この手法の仕様上最後の文節が必ず要約文に入るのでこの方法での要約はほとんどの文に冗長部分が付け足されてしまう.

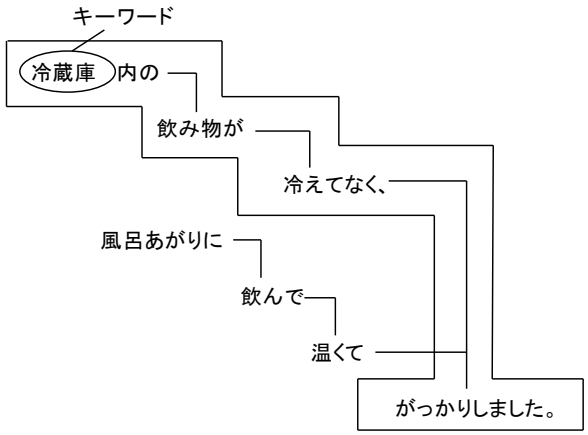

図 4. 先祖子孫関係を用いた係り受け関係の例

3.5.2 親子関係を用いた要約

この項では CaboCha による係り受け関係を用いて親子関係についての要約を行い, キーワードが主語で無い文についても要約を行う. 図 5 のようにキーワードに係るまでの文節とその親子関係が途切れるまでの文節を要約文とし, 文末を終止形にすることで要約を行う. しかし, この要約ではキーワードが主語に直接係らない場合は要約文中に主語が出現せず文が成り立たないことが多い. 例えば『冷蔵庫の中にサービス品が入っていた。』という文があると, この手法では『冷蔵庫の中に』という要約文になってしまう. なのでキーワードが主語のときとそれ以外のときで場合分けをする必要がある. 主語の判別はキーワードの後ろに係助詞の”は”もしくは格助詞の”が”が存在する場合にその語を主語と判断する.

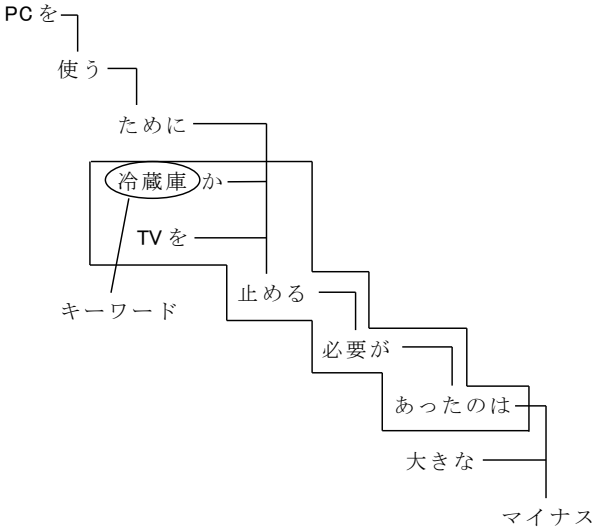

図 5. 親子関係を用いた係り受け関係の例

3.5.2.1 キーワードが主語に係る場合

この項ではまず, キーワードを含む文節の直前と直後の項に着目して, その項とキーワードが親子関係にあった場合その文節を要約文に追加する. そしてさらに追加した文節の直近の文節について親子関係があるかを調べる. これを繰り返して親子関係が成立しなくなるまで行うことで要約文を作成する. これによって最小の文節数で要約文を作成することができる.

3.5.2.2 キーワードが主語に係らない場合

キーワードが主語に係らない場合は前述のように主語の無い文になってしまうので, ここではキーワードの 1 つ後の文節について同様の作業を行う. これによってキーワードが目的語のときの主語や述語へ係り受け関係がある場合に文が成立する.
レビューによっては主語が元々無い場合もあるが, この手法では主語の有無に関わらず要約を行うことができるので主語が文中に存在しない場合でも同様に要約文を作成することができる.

4.実験結果

1 つの施設について要約を行った結果, 3.5.1 項が 11 件中 2 件, 3.5.2.1 項が 30 件中 27 件であった. 3.5.2.2 項については元々件数が少なく 200 件中 10 件しか存在しなく, 10 件中 7 件成功した. キーワードが主語のものは 3.5.1 項の手法では『冷蔵庫があるのは助かります。』や『流石に冷蔵庫が無いときついです。』など, 中には要約が成功するものも存在するが, 『冷蔵庫はあるけど全然冷えてなくて使えなかった。』

Copyright(C) 2012 The Association for Natural Language Processing.
All Rights Reserved

という文は『冷蔵庫は使えなかった。』という要約文になって理由が冗長部分と判断されて省略されてしまう.これに対して 3.5.2.1 項の手法では正常に要約することができる.また, この手法で『冷蔵庫内に髪の毛が落ちていたので、気になる人は無理かもしれません。』という文について要約を行うと『冷蔵庫内に髪の毛が落ちていた。』という文になった.3.5.2.2 項については主語が無い文にも適用することができたが, 『冷蔵庫にビールを入れておいたら、いざ飲もうとしたとき凍っておりました。』という文章に関しては『冷蔵庫にビールを入れておいた。』という文になってしまった.

5. まとめ

これまでの研究で特定の宿泊施設についてのレビューからその施設に関連のある名詞をキーワードとして抽出し, そのキーワードについての要約文を作成した. 本稿ではユーザー投稿レビューから名詞を抽出してその出現回数を用いた重要単語のリストの作成を行った. その中の 1 つをキーワードとして検索を行うことによって利用者の必要とするレビューを抽出して要約文を作成することができた.

今後の課題としてまず, レビューの宿泊施設についての特徴を持たない実用性の無いキーワードを排除する必要がある. 本システムでは利用者がキーワードを選択してそのキーワードについて要約を行うのでその宿泊施設に特徴のあるキーワードである必要がある. そのために出現回数が多くかつ特徴的なキーワードを選定しなければならない.

そして, 本稿のシステムでは宿泊施設についてのレビューについての要約文を作成してきたが, このシステムを利用して Web 上にある一般的な商品やその他のサービスなどさまざまな分野のレビューについてキーワードに関する要約文を作成するシステムを製作する予定である.

6. 参考文献

[1] 楽天トラベル,http://travel.rakuten.co.jp/

[2] 塚田大介, 内海彰:Support Vector Machine を用いた文書の重要文節抽出,人工知能学会論誌,21 巻 4 号 B,pp.330-339,2006

[3] 唯野良介, 嶋田和孝, 遠藤勉:意見の重要度と客観的補足情報を考慮したレビュー要約, 言語処理学会第 17 年次大会, pp.204-207,2011

[4] 楽天技術研究所,http://rit.rakuten.co.jp/

[5] 日本語係り受け解析システム CaboCha http://code.google.com/p/cabocha/

[6] 関根聡:テキストからの情報抽出,情報処理学会誌,40 巻 4 号,1999

Copyright(C) 2012 The Association for Natural Language Processing.
All Rights Reserved