4S-7

検索サイトを利用した自由記述式アンケートの特徴名詞抽出

星野 詞文† 岡 誠‡ 吉村 宏樹‡

†東京都市大学大学院工学研究科 ‡東京都市大学

1 はじめに

企業において商品に対する消費者の意見を次の商品に反映させるためにアンケートは必要不可欠なものとなっている。このアンケートの回答方式は「選択式アンケート」と「自由記述式アンケート」の２つに分類される。中でも自由記述式アンケートは消費者の意見が具体的に記述されているため、選択式アンケートよりも重要視される場合が多い。この自由記述式アンケートを解析する手法の１つに P/N 分類がある。P/N 分類は自由記述式アンケートを満足意見(Positive)と不満足意見(Negative)に分類するもので、満足意見は商品の販売促進案として、また不満足意見は商品改善案として、それぞれに重要な意味を持つ。しかし自由記述式アンケートは消費者の回答形式に明確な決まりが存在しないため、解析が難点といえる。この自由記述式アンケートの解析に関して峠[1]は大規模テキスト中の意見・評判情報の抽出を行う際に、ドメインを最も端的に表す１語であるクエリーに関連するドメイン特徴語の抽出を行っている。これはドメインによって評価の対象となる表現が異なるという観点から、ドメインを決定するクエリーに対して関連度の高い語彙を特徴語とする抽出方法である。抽出精度は約75%と比較的良好な方法であるといえるが、特徴語と評価表現語彙の対応はなされていない。

2 提案

本研究では特徴名詞と評価表現語彙との関係に着目し、構文解析を用いて複合名詞同定処理を行うことで特徴名詞とその評価を表現する語彙の抽出に関連を持たせる手法を提案する。評価表現語彙とは特徴名詞を修飾する語のことである。

P/N Classification of Free descriptive survey using Search Site

†Norifumi HOSHINO ‡Makoto OKA ‡Hiroki YOSHIMURA

‡System Information Engineering,Grad. Sch. Of Engineering, Tokyo City Univ. ‡Tokyo City Univ.

3 システム内容

図１は自由記述式アンケートの特徴名詞を抽出するまでの流れを示した図である。以下にシステムの流れを説明する。

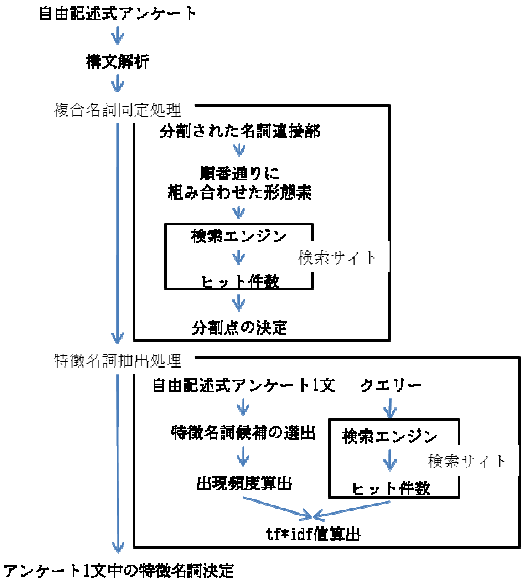

図１：システムの流れ

3.1 構文解析

自由記述式アンケート全体に構文解析を行う。構文解析には CaboCha[2]を使用した。

3.2 複合名詞同定処理

形態素解析によって分割されすぎてしまった名詞を再接続することで適切な特徴名詞を抽出するための補助となる処理である。連接した名詞を順番通りに組み合わせたものが候補となり、検索サイトを用いて検索ヒット件数を調べる。これ以降、検索サイトを使用する場合は全て Google[3]の検索エンジンを使用する。ここで「温度調節機能」を例にとると「温度/調節/機能」となる。これを順番に組み合わせていき(温度調節機能)，(温度調節)，(調節機能)，(温度)，(調節)，(機能)の6語を検索語として検索ヒット件数と最長一致法を用いて複合名詞を同定する。表１に検索結果を記載する。また最長一致法とは構成形態素数の多い順に閾値と比較し、閾値を超

えた組み合わせが出現した時点で形態素の分割点を決定する方法である。今回は閾値を1,000,000とした。

表1：各形態素を組み合わせた検索結果

	検索内容	検索件数
3形態素	温度調節機能	552,000
2形態素	温度調節	705,000
	調節機能	5,201,000
1形態素	温度	61,500,000
	調節	15,500,000
	機能	199,000,000

表1の例では「調節機能」が閾値を越えているため、「温度/調節機能」がこの複合名詞の分割点となる。

3.3　特徴名詞抽出処理

ドメインにより特徴名詞となるものが異なるため、ドメインに基づくクエリーとの関連度が高いものを特徴名詞とする。特徴名詞は特定のドメインでは頻繁に出現するが、他のドメインでは出現頻度が低くなるという性質を持つので、「自由記述式アンケート内出現頻度」と「クエリーと特徴名詞候補とのAND検索」の結果を用いてtf*idf値を関連度として算出する。クエリーは「冷蔵庫」と設定した。特徴名詞候補は(名詞-一般),(名詞-サ変接続),(名詞-固有名詞),(名詞-複合名詞),(未知語)の品詞であり(自由記述式アンケート全体で出現頻度3%)かつ(記号,数字列,ひらがな列を除く)というフィルタリングを行うことで得られる。以下に例を挙げる。

例)チルドルームの温度調節機能が使い易い。

表2：1文中の特徴名詞候補のtf*idf値

検索候補	tf*idf値
(冷蔵庫,チルドルーム)	32.46
(冷蔵庫,温度)	70.77
(冷蔵庫,調節機能)	264.03

よってこの例では「調節機能」が特徴名詞となる。

4　結果

特徴名詞抽出処理の精度から検証を行う。検証方法は自由記述式アンケートから人手で特徴名詞を決定し正解データとしたものと抽出結果の比較により正誤判定を行った。本研究での特徴名詞の抽出精度は74.28%となった。ここに峠[1]が行った抽出結果との比較を表1に記載する。表1を見てわかるようにほぼ峠[1]の場合と変わらない精度となった。

表3：抽出精度の比較

	本研究の抽出精度	峠[2]の抽出精度	
ドメイン名	**冷蔵庫**	携帯電話	デジタルカメラ
精度	**74.28%**	71%	76%

5　考察

本研究ではP/N分類のために特徴名詞を抽出したが、峠[1]の場合は意見文抽出のために特徴語抽出を行っている。このとき検索ヒット件数をそのまま使用していたため他のドメインでは出現頻度が低いという特徴語の性質に対応することが難しいと考えられる。よって本研究では自由記述式アンケート内出現頻度を用いることでこの性質に対応している。

また特徴名詞抽出処理において正解データと抽出結果が一致しなかった場合が存在したが、これは特徴名詞候補として挙げた5つの品詞のどれも含んでいないものがほとんどであった。P/N分類を行うには特徴名詞が存在しなければならないため、特徴名詞が存在しない場合の分類方法を検討する必要がある。

6　おわりに

今後は獲得した特徴名詞を用いて評価表現語彙抽出処理とベイジアンフィルタリングを用いてP/N分類を行う。評価表現語彙抽出処理は(名詞-サ変接続),(名詞-ナイ形容動詞語幹),(名詞-形容動詞語幹),(動詞-自立),(形容詞-自立),(副詞-一般)を評価表現語彙候補として特徴名詞の前後で係り受け関係にある語彙を抽出する手法である。P/N分類後は再現率・適合率・F値を算出し、精度の検証を行う予定である。

参考文献

[1]峠 泰成,山本 和英（2006）：“大規模テキストからの意見・評判情報の抽出手法”,平成18年度長岡技術科学大学大学院修士論文
[2]工藤 拓,松本 裕治(2002):”チャンキングの段階適用による日本語係り受け解析”,平成14年度情報処理学会論文誌,No.6
[3]Google:”http://www.google.co.jp/”,2011年1月14日現在