2V-4

検索サイトを用いた自由記述式アンケートの特徴語抽出法

星野　詞文†　　吉村　宏樹‡　　岡　誠‡　　森　博彦‡
東京都市大学大学院†　　東京都市大学‡

1.　研究背景

企業が製品に関して行うアンケートの中に自由記述式アンケートがある。これは回答者の意見が直接的に記述されるため,自由記述式アンケートが分類可能となれば次製品考案の足掛かりになる。分類法は多岐にわたるが,どの分類法にも特徴語は不可欠な要素となる。また製品に関する語は企業独自の語や複合名詞が存在するため,これらの語が特徴語となるような抽出法が必要となる。

2.　研究目的

本研究では製品に関する自由記述式アンケート 1 文において,その文を端的に表す特徴語の抽出を目的とする。

3.　既存研究

自由記述形式での特徴語抽出を行うにあたり複合名詞同定が必要となる。峠〔1〕によれば,特徴語になる語は構成形態素数も様々であるとしている。そこで複合名詞同定に検索サイトを使用する手法を提案した。しかし構成形態素が多いほど同定され難いという問題がある。

また沢井〔2〕によれば,検索ヒット数をそのまま使用することは不適切であるとし,複合名詞を構成する形態素の AND 検索ヒット数内で,その複合名詞が存在する割合から複合名詞を同定する手法を提案している。しかし,この手法でも構成形態素の多い場合は同定され難い。

4.　提案手法

既存研究を受けて,本研究では以下 2 項目を提案する

4.1.　連接係数を用いた割合計算法

沢井〔2〕の用いた割合計算では検索ヒット数のみで同定した。しかし名詞にはその種類によって前部や後部と結合されるべきものが存在する。これは必ず結合されるものではないがその可能性は大きいため,品詞情報を用いた名詞の連接確率による新たな割合計算法を提案する。

4.2.　割合を用いた特徴語抽出手法

従来の特徴語抽出には tf*idf を用いる手法がある。これは文書内単語出現頻度(tf)と他文書出現頻度(idf)を用いるものだが,idf は正確な入力を要するため精度上昇が困難となる。そこで新たに idf の変則法を提案する。

5.　複合名詞同定処理

特徴語を抽出する際に形態素解析を行うが,解析器は内蔵辞書に沿って形態素毎に分割する。このままでは企業独自の語や複合名詞が分割されてしまい,特徴語抽出に支障をきたす可能性がある。これを回避するため提案手法を用いた複合名詞同定処理を行う。

5.1.　システム

初めに自由記述式アンケートに対して形態素解析を行う。形態素解析には MeCab〔3〕を使用した。複合名詞候補は「名詞-(一般,サ変接続,固有名詞,接尾,数,接頭詞,形容動詞語幹,副詞可能)」が連続して出現する部分とする。これらに対して複合名詞同定を行う。検索サイトは「Google」〔4〕とした。表記ゆれを回避するために候補をダブルクォーテーションで囲んでいる。割合計算は検索対象 Q の検索ヒット数を HIT(q),複合名詞候補を W,W に含まれる形態素を $\{\omega_1,\omega_2,\ldots\omega_n\}$,連接検索ヒット数をHIT_CO(W)とすると

$$\mathrm{HIT\_CO(W)} = \mathrm{HIT\_CO("}\omega_1\omega_2\ldots\omega_n\mathrm{")} \quad \cdots(1)$$

また AND 検索ヒット数をHIT_AND(W)とすると

$$\mathrm{HIT\_AND(W)} = \mathrm{HIT\_AND("}\omega_1\mathrm{"and"}\omega_2\mathrm{"and\ \ldots\ and"}\omega_n\mathrm{")} \quad \cdots(2)$$

さらに割合を$\mathrm{HIT_PAR(W)}$とすると

$$\mathrm{HIT\_PAR(W)} = \frac{\mathrm{HIT\_CO(W)}}{\mathrm{HIT\_AND(W)}} \quad \cdots(3)$$

と式を立てることができる。ここで提案手法で

Feature Extraction Method for the questionnaire free text search
†Tokyo City University Graduate school
‡Tokyo City University

ある連接係数を用いる。連接係数とは候補内の品詞情報から名詞の連接確率を用いたものであり、今回は正解ラベルから名詞の連接確率を計算した。この連接係数と(3)式を用いた(4)式で候補の値を定める。

$$\text{連接係数} \times \text{HIT\_PAR}(W) \quad \cdots (4)$$

最長一致法により(4)式の値が閾値を越えた時点で複合名詞を同定する。閾値は沢井〔2〕と同じく0.001とした。

5.2. 検証

複合名詞同定処理について「冷蔵庫の製品に関するアンケート」499文を用いて検証を行う。これは企業が実際に行ったアンケートである。検証のための正解ラベル作成は候補に対して人手で複合名詞として同定したものを使用した。表2は正解ラベルに対する処理結果の正解率であり,提案手法と既存手法の精度を比較したものである。

表1：複合名詞同定処理正解率

	2語	3語	4語
連接係数を用いた割合	83.13%	65.00%	75.00%
割合	80.54%	63.26%	42.25%
ヒット件数	82.49%	68.39%	51.90%

表2から,構成形態素が多くなるに連れて割合とヒット件数は正解率が下がっている。しかし提案手法ではどの構成でも正解率が大きく下がることはなかった

5.3. 考察

今回使用したアンケートにおいて,回答者は企業の付けた名称を曖昧に記述する傾向が多く見られた。これは回答者にとって名称を重要視していないためであると考えられる。また照応詞を用いた記述も見られた。提案手法では検索サイトを利用しているため,曖昧な記述や照応詞が含まれている場合では意図しない結果となってしまう。

6. 特徴語抽出処理

複合名詞同定処理結果から1文中の特徴語を抽出する。この特徴語の条件は複合名詞候補条件と同じである。

6.1. システム

1文中の条件を満たす名詞のtfと検索ヒット数の割合を用いて特徴語を決定する。ここで割合計算法は以下(5)式のようになる。

$$\text{HIT\_CHAR}(W) = \frac{\text{HIT\_AND}(\text{"MQ"and "W"})}{\text{HIT\_CO}(W)} \quad \cdots (5)$$

(5)式の分子内MQはメインクエリーの略であり,アンケートを端的に表す語となる。今回使用するアンケートからMQを「冷蔵庫」と設定する。(5)式は特徴語候補を含むページの中で冷蔵庫が同時に出現するページはどれだけ存在するかを表している。この(5)式とtfを用いた(6)式にて特徴語候補の値を決定する。

$$\text{tf} \times \text{HIT\_CHAR}(W) \quad \cdots (6)$$

6.2. 検証

特徴語抽出処理について,「冷蔵庫の製品に関するアンケート」499文に対して複合名詞同定処理を行った場合と行わなかった場合で特徴語の判定率を検証する。表2に各判定率を掲載する。

表2：特徴語抽出精度

	判定率
複合名詞同定処理あり	74.8%
複合名詞同定処理なし	65.5%

表2から複合名詞同定処理を行った場合の方が特徴語抽出精度を向上させることができた。

6.3. 考察

本研究にて,意図した特徴語が得られなかった原因に以下のことが考えられる。提案手法では名詞のみを特徴語候補として挙げていたが,企業は覚えやすい名称を設定するため,文節をまたがっているような付け方が存在した。提案手法では名詞以外の品詞を候補としていないため,本来は複合名詞となるべき品詞も候補から外れてしまっており,これが判定率低下を招いていると考えられる。

今後は解析器内蔵の辞書に特徴語となるべき語を追加し,再実験を行う予定である。

参考文献

〔1〕 峠泰成：“大規模テキストからの意見・評判情報の抽出手法”,長岡技術科学大学大学院修士論文（2006）
〔2〕 沢井康孝：“Web検索を用いた複合名詞同定”,言語処理学会,第14回年次大会発表論文集,pp.205-208（2008）
〔3〕 MeCab：形態素解析ツール「MeCab」ver.0.98,http://mecab.sourceforge.net/
〔4〕 Google：http://www.google.co.jp/