意味解析による自由記述アンケートの自動分類システム AQUA

村上　裕人[†]　　谷澤　嘉和[†]　　韓　東力[‡]　　原田　実[‡]

[†,‡] 青山学院大学　理工学部　情報テクノロジー学科

1.　はじめに

アンケートにおいて自由記述型のアンケートは選択型のアンケートに比べ回答者の自由な意見を集約できるという効果がある。しかし、意見の集約は手作業によるものが多く、多大な人的、時間的コストが必要とされる。また、人の判断による分類は客観的基準になりにくいという問題もある。そのため一定の基準を設けて意見集約を行うことが重要と考える。

一方、テキストの自動分類や検索技術では文中の名詞の頻度ベクトル間の内積距離等による類似度計算に基づく方法[1]や文中動詞のガ格、ヲ格等の表層格の対応対の語意類似度に基づく方法[2]が提案されているが、これらは意味解析を行っていないので、語意の信頼性が低く、また、表層格では語の役割が選別できないので、誤った類似性が生じる。また、アンケートの自動分類では表層表現の類似性に着目した統計的手法による賛成、反対、要望・提案、事実等の意図タグへの自動分類の手法[3]があるが、この手法では回答者の「何をどうしたい」や「何をどうすべきか」といった意見毎の分類ができない。

以上の背景を踏まえ、本研究では回答者の意見の効率的な集約と客観的な基準に基づく意見の分類を行うため、語意だけでなく、語間の意味関係を考慮した深層的な理解に基づく自由記述アンケートの自動分類システム AQUA の開発を行った。

2.　基本概念とシステム概要

2.1.　基本概念

名詞は文が表現する話題を容易に推測するための指標になると考え、文に含まれている名詞の出現頻度ベクトル間のコサイン距離で第一次文間類似度を算出しクラスタリングを行う（名詞によるクラスタリング）。

次に述語とそれに係る語句、及びその間（述語と係る語）の意味的関係が回答者の意図を最も反映しており、キーワード（名詞）による分類だけ行うよりも回答者の意図の違いによる分類ができると考え、名詞によるクラスタリングの結果のクラスタ毎に述語とそれに係る語、さらにそれらの間の深層格によって構成される意見フレームを文毎に作成し、それらを用いた第２次文間類似度（述語ベース類似度）を算出する。この類似度を用い、再度クラスタリングを行う（述語によるクラスタリング）。

Automatic classification of Open-Ended Questionnaires based on semantic analysis

Yuto Murakami[†], Yoshikazu Tanizawa[†],Dongli Han[‡] and Minoru Harada[‡]

[†]Department of Integrated Information Technology, Aoyama Gakuin University.

[‡]Department of Integrated Information Technology, Faculty of Science and Engineering, Aoyama Gakuin University.

2.2.　システム概要

本システムの実現に際し、青山学院大学の原田研究室が開発した意味解析システム SAGE[4]を用いアンケート文の意味解析を行い文中の単語の語意を決定し、係り受け関係にある２文節間の格を決定する。意味解析結果は図１に示す frame 述語で表現される。これを本システムの入力とする。

```
frame(1,'平和','ヘイワ','JAM','JAJ','平和','名詞-形容動詞語幹','形容詞-
*','none','none','107fd0',[],[1,1.1]).
frame(2,'的','テキ','JB1','JEA','的','名詞-接尾-形容動詞語幹','接尾辞-形容詞性名詞接
尾辞','none','none','3d0455',[],[1,1.2]).
frame(3,'解決','カイケツ','JSA','JSA','解決','名詞-サ変接続','名詞-サ変名詞
','none','none','3d007a',[[head,3]],[1,1.3]).
                    .
                    .
                    .
frame(9,'めざすべきである','none','JPR','*','めざすべきである
','none','*','none','none','000000',[[consist,5],[consist,6],[consist,7],[consist,8]],[1,2]).
```

図１　AQUA の入力となる frame 述語

本研究のシステム構成を図２に示す。

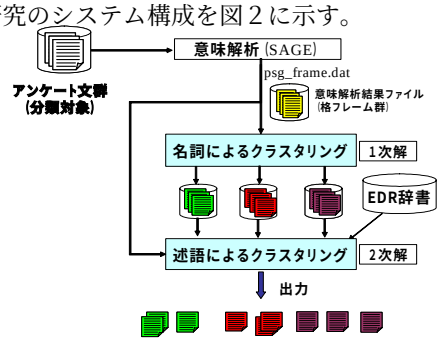

図２　AQUA のデータフロー

3.　名詞によるクラスタリング

各文は出現する全名詞をベクトルの要素としてベクトル空間モデルで表し、各文の各ベクトル要素には TF・IDF 値を得点として与える。このベクトル空間モデルを用い、ベクトル間のコサイン値を類似度とした k-means 法によるクラスタリングを行う。なお、クラスタ数はユーザが指定する。

4.　述語によるクラスタリング

4.1.　意見フレームの作成

アンケート文中の回答者の主たる意見を表現している述語を意見述語と呼び、表１の手がかり語（概念 ID）を元に各行の最右欄の条件を順に評価し見つかった語を意見述語として抽出する。そして、文毎に意見述語とそれに係る語と深層格（係り元情報）からなる格フレーム群を意見フレームとして１つ作成する。

表１　意見述語抽出ルール

意見述語抽出手がかり語(概念ID)	手がかり語のEDR品詞	意見述語と手がかり語の関係
2621b9；べし等 3d029d；｛願望する｝願い望む欲しい 2621ba, 2621c6；断定推量（だろう）	JJD JAX JJD（未然形）	同文節内の主辞となる語
30f87d；望む　※上位概念探索 3ce69b；｛期待する｝物事の実現を待ち望む※上位概念探索 3ce62e；｛要求する｝人に、あることをしてほしいと要求する	JVE JSA	1.object格で係っている語 2.goal格で係っている語 3.purpose格で係っている語
1030aa；｛望ましい｝そうあって欲しいさま	JAJ	a-object格で係っている語
3cf56e,36c396；｛必須だ｝ぜひとも必要なさま、必要だ	JAM	object格で係っている語
30f878；考える　※上位概念探索 3cf498；｛見なす｝あるものを見て、そうであると思うこと 3c1b5e；｛言える｝言うことができる	JVE JSA	1.logical格で係っている語 2.timing格で係っている語
デフォルト		main格となる語

4.2. 否定処理

二重否定を肯定文とする、肯定文同士、否定文同士を比較するという目的から各意見フレームの意見述語や係り元情報と同文節内に否定的な表現（非、不、無、ない等）が含まれる場合、その意見述語や係り元情報に否定フラグを付与する。

次に、意見フレームの組において意見述語に否定フラグが付与されている場合、その否定フラグを使用否定フラグとする。また、否定フラグの付与されている係り元情報に注目し同深層格で結ばれていてかつ語意類似度が高い（0.80 以上）語である場合に否定フラグを使用否定フラグとする。

以上により２つの意見フレーム間で使用否定フラグの数が偶数同士、奇数同士なら述語ベース類似度を 4.3.3.に従って算出し、偶数と奇数となら述語ベース類似度は 0 とする。

4.3. 述語ベース類似度の算出

２つの文間の述語ベース類似度は先に述べた手法で作成された意見フレームを用い、意見述語同士の語意類似度とそれに同格で係る語（同格ペア）の語意類似度、必須格による得点、格一致度から算出する。語意類似度は、意味解析により割り当てられた概念 ID と EDR 概念体系辞書[5]を用い概念の深さと共通上位概念の深さを算出し、下式により算出する。

$$\text{語意類似度} = \frac{2 \times dc}{di + dj}$$

di, dj ： 二つそれぞれの概念の深さ
dc ： 二つの共通上位概念の深さ

4.3.1. 必須格による得点

比較する２文の意見述語の必須格を抽出する。この際、EDR 共起辞書中の用言の各深層格 x の出現割合を統計処理し、その平均値 m_x を求め、ある述語 v の格 x の出現割合 $m_{v,x}$ が m_x 以上であれば、x を v の必須格とする。そして、両方の意見述語で必須格となる係り元には３点、片方の意見述語で必須格となる係り元には２点を与え、同格で係る語の語意類似度の得点とする。

4.3.2. 格一致度と同格ペアの決定

比較する２つの意見フレームにおいて同格で係る語が多対多で存在する場合、係り元の類似度の合計が最大になる組み合わせを同格ペアとする。これを元に、格一致度を下式により算出する。

$$\text{格一致度} = \left(\frac{N_{i,j}}{M_i} + \frac{N_{i,j}}{M_j}\right) \times \frac{1}{2}$$

$N_{i,j}$:同格ペアの個数
M_i :述語から出ている格の個数

4.3.3. 述語ベース類似度

本節で述べてきた値を用い、下式により述語ベース類似度を算出する。この類似度を元に名詞によるクラスタリングで得た同クラスタ内のすべての文の組み合わせに対して算出し、階層型クラスタリングを行う。なお閾値はユーザが指定する。

$$Sim_{i,j} = \left(p_{i,j} + C_{i,j} \times \frac{\sum(\alpha_x \times w_x)}{\text{必須格得点の合計}} \right) \times \frac{1}{2} \quad (0 < Sim_{i,j} \leq 1)$$

$Sim_{i,j}$: 文 i と文 j の述語ベース類似度
$p_{i,j}$: 意見述語の語意類似度
w_x : x 格が必須格の時の得点
α_x : x 格で係る係り元の語意類似度
$C_{i,j}$: 意見フレーム i, j の格一致度

5. 実験とまとめ

複数の事例に対して評価実験を行った。例えば、「イラク攻撃に関する全国会議員公開アンケート」の「今後、対イラク攻撃に関係し、アメリカから協力(自衛隊派遣、資金負担など)要請があった場合、日本はどう対応すべきだと思われますか。」という質問文に対する回答３５文の分類について述べる。
(http://www.eeeweb.com/~research/top.htm)

ただし、回答が２文に渡る場合、別の回答として扱った。また、名詞によるクラスタリングではクラスタ数を２とし、述語によるクラスタリングでは閾値を 0.50 とし２つのクラスタの一番近い要素間の類似度をクラスタ間の類似度とした。

本システムによる分類により、回答文群は１４のクラスタに分類された。図３はその分類結果の一部である。（各クラスタは点線で区切られている）

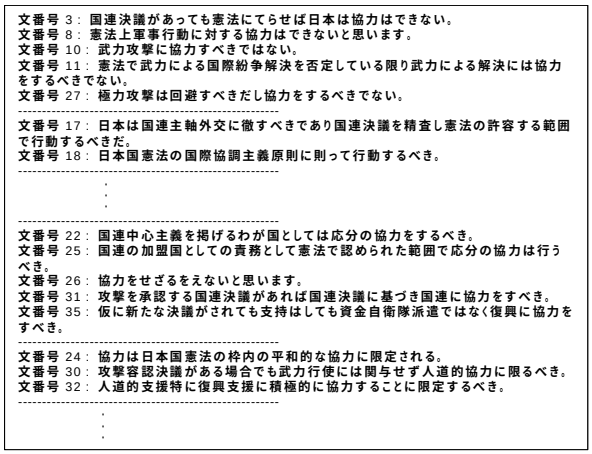

図３　実験による分類結果（一部）

いずれの実験においても、同一クラスタ内の文は対象においても、対象間の関係においても非常によく類似しており、またクラスタ間ではそれらに差があった。この結果から本システムにより語意と語間の意味的関係を考慮した客観的基準による自由記述アンケートの分類を行うことができたと言える。

【参考文献】

[1]徳永健伸 “ 情報検索と言語処理 ” 東京大学出版会 (1998)

[2]立石・峯・雨宮 “ 係り受け構造や語の意味情報を利用した日本語テキスト検索システム ” 言語処理学会 第 5 回年次大会発表論文集 p317-p320 (1999)

[3]乾・村田・内元・井佐原 “ 表層表現に着目した自由回答アンケートの意図に基づく自動分類 ” 自然言語処理 vol.10 No.2 p19-p42 (2003)

[4][原田 2002b] 原田実, 田淵和幸, 大野博之, "日本語意味解析システム SAGE の高速化・高精度化とコーパスによる精度評価", 情報処理学会論文誌, Vol.43, No.9, pp.2894-2902, (2002.9)

[5]㈱日本語電子化辞書研究所：EDR 電子化辞書 (1995).