

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

Automatic classification of Open-Ended Questionnaires based on semantic analysis

谷澤 嘉和(15800047) TANIZAWA Yoshikazu

村上 裕人(15800073) MURAKAMI Yuto

原田研究室 指導教員 原田実教授 韓東力助手

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の類似性に基づく方法^[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）の意味的關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）。

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 frame 述語で表現され、これを本システムの入力とする。

本研究のシステム構成を図1に示す。

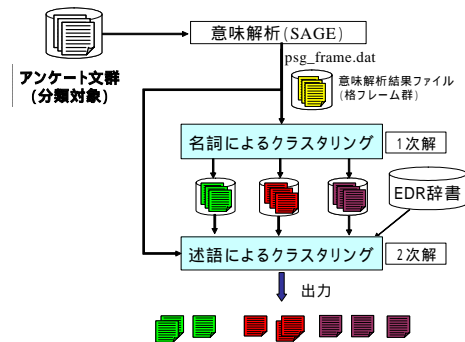


図1 AQUAのデータフロー

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。

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

平成15年度卒業論文要旨

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 価詞	意見述語と手がかり語の関係
2621b9; べし等 3d029d; (願望する)願望む欲しい 2621ba, 2621ob, 断定推量(だるう)	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 3cs89b; (期待する)物事の實現を待ち望む 上位概念探索 3cs62e; (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)なげひとも必要なさま, 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 3ctb5e; (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。

$$\text{語意類似度} = \frac{2 \times dc}{di + dj} \quad \begin{matrix} di, dj : \text{二つそれぞれの概念の深さ} \\ dc : \text{二つの共通上位概念の深さ} \end{matrix}$$

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。

$$\text{格一致度} = \left(\frac{N_{i,j}}{M_i} + \frac{N_{i,j}}{M_j} \right) \times \frac{1}{2} \quad \begin{matrix} N_{i,j} : \text{同格ペアの個数} \\ M_i : \text{述語から出ている格の個数} \end{matrix}$$

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。

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

$Sim_{i,j}$: 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. おわりに

アンケート回答事例に対して評価実験を行った結果、本システムではクラスタ数や閾値などを変えながら最も有効的な分類結果を探す必要がある、述語によるクラスタリングにおいて対象と行為(例:何をどうしたい)を考慮した分類を行っており、その分離性、結合性が最適な分類に影響しているということが分かった。

また、本システムにより客観的な基準を設けて約80%の精度で自由記述アンケートの自動分類を行うことができ、「述語によるクラスタリング」により、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アンケートの分類を行うことができた。

【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]株式会社日本国語辞書研究所: EDR 電子化辞書 (1995).