8N-7　常識判断メカニズムにおける未知語処理

土屋　誠司　　小島　一秀　　渡部　広一　　河岡　司
同志社大学大学院　工学研究科

1. はじめに

人間に優しいコンピュータの開発が望まれている今後，コンピュータにとって「常識的に判断する」能力は，極めて重要な要素となってくる．

本稿は，人間がコンピュータに入力するテキスト情報から，「表現の背景となっている利用者の感情」をコンピュータに推測させる「感情判断」メカニズムの実現を対象としている．あらかじめ感情知識ベースに用意された表現についての感情判断は問題とならないが，感情知識ベースにない未知の語が入力された場合には対処できない．本稿では，このような未知語に対しても適切な感情の推測ができる未知語処理のメカニズムについて報告する．

2. 感情判断メカニズム

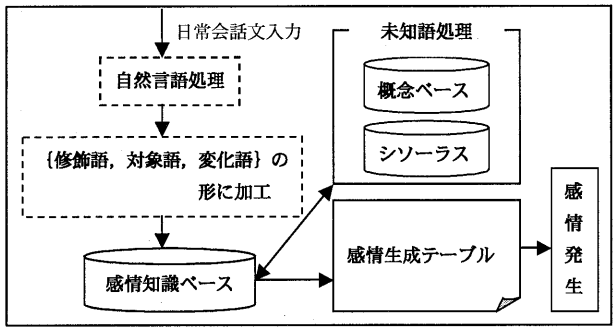

図1．感情判断メカニズムの流れ

感情判断メカニズムとは，入力された文章から，それを発した人の感情を推測するものであり，入力情報としては特に感情発生に強く関係する{修飾語(形容詞・形容動詞)，対象語(名詞)，変化語(動詞)}の 3 要素が，適切な自然言語処理により抽出されたことを前提としている．また，出力として扱う感情は，一般的に基本的であると考えられる{喜び，悲しみ，怒り，落胆，恐れ，恥，安心，後悔，罪悪感}の 9 種類とした．

本メカニズムの流れを図1に示す．修飾語，対象語，変化語は，それぞれ対応する分類語に集約され，感情知識ベースの中に格納されている．修飾語に関しては，日常使用する単語数が比較的少ないため(約 4000 語)，全ての語を用法的に分類し格納している(4 分類)．しかし，対象語と変化語は非常に語数が多く，全てを感情知識ベースに格納することは事実上不可能であるため，それぞれ代表となる語(対象語:99 語，変化語:541 語，以下代表語と呼ぶ)を抽出し，対象語は感情発生の観点で，変化語は動作の観点で分類している(対象語:34 分類，変化語:68 分類)．また，対象語に関しては，34 分類の観点で日本語語彙大系[1]をベースとして，独自に構築した感情シソーラスも感情知識ベースと同時に用いる(約 3 万 7 千語)．尚，発生する感情はこれらの分類を使用した感情生成テーブルにより生成される({(対象語分類＋変化語分類)→感情}:約 2000 通り)．

2. 1. 概念ベース

本実験に用いた概念ベースは，複数の国語辞書などの語義文から自立語を抽出したもので，構成する各概念の構造は，重み情報のない属性語 a_i の集合のみの最も単純な構成とする．

概念 X:$\{a_1, a_2, \cdots, a_i, \cdots\}$

機械精錬が適切に進んだ段階では，概念 X の属性集合$\{a_i\}$は概念 X の意味をそれなりに表現する適切な個数の属性集合(概念集合)となっている[2]．尚，属性語 a_i の語数は最大 30 語としている．

2. 2. 関連度

関連度とは，概念の関連性を定量的に評価するものであり，概念連鎖により概念を n 次属性まで展開したところで，一致する属性の個数を評価することにより算出するものである．本稿では，2 次属性まで展開する方式を用いた[2]．具体的には，2つの概念がもつ各 30 語の 1 次属性の対応をとり，それらの 2 次属性の一致数を基に評価している．

概念 A と概念 B の2つの n 次属性 a_i^n と b_j^n の一致度を $\mathrm{Match}(a_i^n, b_j^n)$，1 次属性数を N_A, N_B とし，関連度を Rel(A, B)とすると，

$$\mathrm{Rel}(A,B)=\frac{\dfrac{\sum_{i=1}^{N_A}\mathrm{Match}(a_i^1,b_{xi}^1)}{N_A}+\dfrac{\sum_{i=1}^{N_B}\mathrm{Match}(a_i^1,b_{xi}^1)}{N_B}}{2}$$

と表すことができる．

2. 3. 未知語

概念ベース，感情シソーラス，感情知識ベース各々の関係は，概念ベースは，語と語の関連性を表したものであり，あらゆる品詞の語を連想することができる．感情シソーラスは，感情判断の観点から名詞の語の意味的分類を行ったものであり，名詞についての親子関係並びに兄弟関係のみ表すことができる．感情知識ベースは，感情発生の観点から人の手で抜粋した語の集合であり，

機械的には収集・分類できない特殊な意味の範囲を表現する.

本稿でいう「未知語」とは,感情知識ベースと感情シソーラスに格納されていないが概念ベースには存在する語のことを指す.前にも述べた通り,概念ベースには約18万語,感情シソーラスは約3万7千語,感情知識ベースには約99語収録されているので,約14万語が「未知語」ということになる.

3. 未知語処理

前述したように,対象語と変化語は非常に語数が多く,全てを感情知識ベースに格納することは事実上不可能であるため,各々代表となる語を抽出し分類している.

そこで,対象語として未知語が入力された場合,変化語として入力された動詞から,文法的に接続可能な名詞の代表語を自動的に既存の文法辞書を引くことにより選択する.この文法辞書には,接続可能な動詞と名詞の組が掲載されており,名詞については感情シソーラスのノード名で記載されている.

そして,入力された未知語が,選択されたどの代表語と意味的に近いかを関連度を用いて判断し,その未知語をある代表語に置き換えて処理を行う.尚,関連度の数値は,絶対評価を行うことが不可能であるため,自動的に選択された代表語を母集団として,関連度を偏差値化して評価する.評価の基準は「偏差値97以上の数値を得たものが意味的に近い」と実験による結果を基に設定した.未知語処理のイメージを図2に示す.

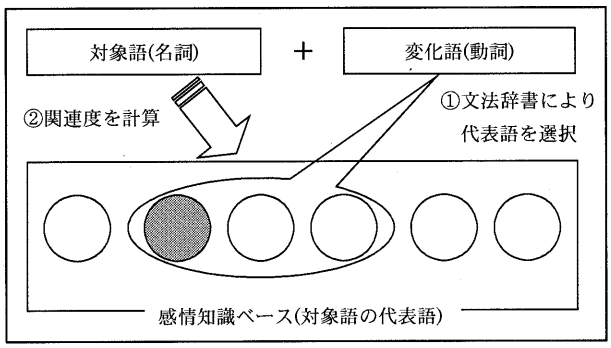

図2. 未知語処理のイメージ図

また,変化語についても未知語処理をする必要性はあるが,名詞と動詞ではその語が表す意味的な範囲が異なるため,名詞と同様の処理ではうまく扱うことはできない.よって本稿では未知語処理としては名詞のみを扱うものとする.

4. 実験と結果

実験データとしてはアンケートで集めた1078文(有効回答数441文)について行った.結果を表1,表2と図3に示す.

表2の処理結果は,一般的に判断すると間違った分類に処理されたとはいえないが,感情発生の観点からは誤りである.これらの未知語は,その語が持つ意味の範囲が広く,感情発生の観点で連想される(関連のある)語が複数になり,又それら連想される語の分類はそれぞれ異なる.そして,それらの語が代表語として登録されているため,うまく扱うことが出来ない.

結果として,すでに報告している方法[3]よりも今回提案した新方式の方が,約10%正解率が向上している.内訳としては,感情知識ベースと感情シソーラスでカバーできたものは全体の約57%であり,未知語処理により処理できたものが,約25%である.入力文全体の約4分の1が未知語処理により扱えるようになったことからこの処理方法は有効であると考えられる.

表1. 正解結果例(カッコ内は代表語の分類名)

未知語	処理結果
熊	猛獣(危ない)
同級生	知人(大切な)
日曜日	休息(待ち遠しい)

表2. 誤りの結果例(カッコ内は代表語の分類名)

未知語	処理結果	期待する解
手術	病気(不幸な)	救助(心強い)
渋滞	乗り物(便利な)	優柔不断(苛立たしい)
卒業	勉強(苦しい)	祝い(めでたい)

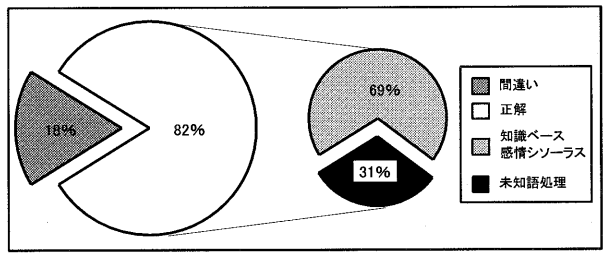

図3. 新方式による未知語処理結果

5. おわりに

感情知識ベースにない未知の語が入力された場合の対処法である未知語処理のメカニズムについて新しい方式を提案し,その有効性を実験により確認した.

本研究は文部省からの補助を受けた同志社大学の学術フロンティア研究プロジェクトにおける研究の一環として行った.

参考文献

[1]NTTコミュニケーション科学研究所監修,“日本語語彙大系”,岩波書店(1997)

[2]入江毅 渡部広一 河岡司 松澤和光:“知的判断メカニズムのための概念間類似度評価モデル”,電子情報処理信学会 信学技報 AI98-75(1999)

[3]土屋誠司 馬場秀樹 渡部広一 河岡司:“入力文から感情を判断するシステムにおける未知語の処理”,電子情報通信学会 信学技報 AI99-107(2000)