E-021

# 常識的大きさ判断システムにおける大きさの曖昧性を表現可能な未知語処理手法
# Unknown Word Processing Considering Ambiguity of Size for Size Judgment System

竹中 愼† 吉村 枝里子‡ 土屋 誠司‡ 渡部 広一‡
Jin Takenaka Eriko Yoshimura Seiji Tsuchiya Hirokazu Watabe

## 1. はじめに

近年，ウェアラブルデバイスの増加や，コンピュータの急速な社会への普及などから，あらゆるユーザにとって使いやすいコンピュータが求められている．そこで，人の意図を適切に汲み取り，人との会話を行うコンピュータの実現の必要性が高まっている．

我々は，日常的に，人が持つ共通の知識，すなわち「常識」を用いて円滑な会話を行っている．そこで，会話システムにおいても，単語や文章に関する「常識」を基準として判断できるシステムが必要不可欠だと考えられる．

「常識」の一つとして大きさに関する常識がある．我々は会話の中で「常識」を用いて日常的に「大きさ判断」を行っている．例えば，「車にテレビを積んでください」という発言をする際に，「車とテレビ」の大きさを判断した上で発言を行う．そこで本稿では，大きさに関する常識を判断するシステムの実現を目標とする．

大きさに関する判断が可能であるシステムとして，入力された 2 語に対し，量の比較を行う既存の量判断システム[1]がある．このシステムは「大きさ」，「長さ」など，全部で 10 つの量を扱い，これらを「観点」と言う．本稿では，10 つの観点のうち，単純に実際の数値で比較が不可能である「大きさ」を扱う大きさ判断システムについて考える．

既存システムには，システムに登録されていない語（未知語）を登録されている語（代表語）に置換する処理において，その大きさが一意に決定されてしまうという問題点が存在している．例えば「国語辞典と六法全書」の場合，我々であれば「六法全書の方が 2cm 大きい」のような厳密な判断はせず，常識的に「ほぼ同じ大きさ」と判断する．しかし既存の量判断システムでは曖昧性を考慮できず厳密な判断を行ってしまい，「六法全書の方が大きい」という回答を行ってしまう．そこで本稿では，シソーラスの利用法を検討し，大きさの幅を表現することで，曖昧性を考慮可能な未知語処理の実現を目指した．

## 2. 関連技術

以下で大きさ判断システムに利用する技術を説明する．

### 2.1 概念ベース

概念ベース[2]は複数の辞書から機械的に生成された知識ベースのことであり，約 9 万の概念から構築されている．概念ベース内で概念は，その意味的特徴を表す「属性」とその属性が概念にとってどれだけ重要かを表す数値である「重み」の対の集合より定義されている．

†同志社大学大学院理工学部研究科
Graduate School of Science and Engineering, Doshisha University
‡同志社大学理工学部
Faculty of Science and Engineering, Doshisha University

### 2.2 関連度計算方式

関連度計算方式[3]とは，ある二つの概念間の関連の強さを定量的に表現する手法である．0〜1 の数値で表され，値が高いほど関連が強い．

### 2.3 NTT シソーラス

NTT シソーラスとは単語の意味や概念を整理し，体系的に表したものである．NTT シソーラスには一般名詞の意味的用法を表す約 2700 個のノードの上位下位関係・全体部分関係が木構造で示され，約 13 万語のリーフが登録されている．ノードは具体的概念のノードと抽象的概念のノードに分類される．また，本稿では NTT シソーラスをシソーラスと呼ぶ．

## 3. 既存の大きさ判断システム

### 3.1 システムの流れ

既存の量判断システムにおける大きさ判断の流れを説明する．大きさを比較する二語を入力し，それぞれの語が量知識ベースに存在する語（代表語）であるかを判断する．代表語でなければ，量知識ベース内に存在しない語（未知語）と判断し，NTT シソーラスと関連度計算を用いて代表語に置換する（未知語処理）．その後，大きさの値を単純比較し，どちらの入力語が大きいか出力する．

### 3.2 量知識ベース

量知識ベースとは約 1400 語からなるデータベースである．量知識ベースの一部を表 1 に示す．表 1 中の表記が代表語である．表 1 に示すように「大きさ」，「長さ」などの様々な値が格納されている．表中の「－1」という値はその代表語にその観点がないことを意味している．

表 1　量知識ベースの一部

| ID | 表記 | 大きさ | 長さ | … |
|---|---|---|---|---|
| 19 | ペン | 10 | 11 | … |
| 69 | 林檎 | 12 | -1 | … |

「大きさ」は全 64 段階で表され，人間に近いほど 1 段の幅が細かくなる．大きさ以外の観点は，例えば「長さ」は「cm」で表されるように，実際の値で表される．

### 3.3 未知語処理

未知語がシソーラス内にリーフとして存在する場合，未知語の直の親ノードを親に持つ代表語から最も未知語との関連度の高い代表語に置き換える．直の親ノード中に代表語が見つからない場合は，代表語が見つかるまで親ノードをたどり，見つかった代表語の中から最も未知語との関連度の高い代表語に置き換える．シソーラス内に存在しない場合，量知識ベース内の全代表語の表記と未知語の関連度を計算し，最も高い語に置き換える．

## 4. 提案システム

### 4.1 提案未知語処理

2 語を 1 セットとする計 100 の評価セットを作成した．うち 50 セットは新聞から目視で具体物を収集し，残りの 50 セットは学生 50 人の実験データから収集し作製した．

既存システムにおいて 100 セットの大小判断を行い，その結果を三人に示し，「○:常識的」「×:非常識的」の 2 択でアンケート形式の評価実験を行った．過半数が○の場合を○，過半数が×の場合を×として集計している．

×と判断された 59 セットのうち，大きさの幅を考慮する必要があるものが 17 セット存在した．例えば，「国語辞典と六法全書」の場合，大きさの値が異なるので大小を答えてしまうが，本来は大きさの幅を考慮して「なんとも言えない」という回答を行うべきである．そこで，シソーラスのノードにサイズを割り振り(4.1.1 節後述)，未知語のノードのサイズを幅のある大きさとする未知語処理を提案する．

#### 4.1.1 シソーラスノードサイズ

シソーラスのノードそれぞれに最小の大きさと，最大の大きさをそのノードのサイズとして割り振る．あるノードの自分以下のノードに属する，リーフとしてシソーラスに存在する代表語の大きさすべての中で最大の値と最小の値をそのノードのサイズに設定する．最下位のノードからサイズを設定して行き，最終的に最上位ノードのサイズを設定する．

#### 4.1.2 未知語処理のフロー

未知語が入力されると，始めにシソーラスのノードを特定する．まずシソーラスにノードとして存在するかどうかを判断し，存在しなかった場合には，代表語と全ノード名とを関連度計算を行い，関連度が最も高いノードに特定する．

シソーラスに存在した場合，さらに，そのノードが抽象的概念のノードであるか判断し， 抽象的概念であれば入力語は比較不可と出力する．ノードが具体的概念のノードである場合，そのノードに特定する．

上記で特定したノードの名と未知語の関連度が閾値以上の場合，ノードに与えた最小サイズと最大サイズを大きさの幅として取得する．また，閾値より小さい場合には，既存システムの未知語処理により，代表語と置換する．この場合，置換された代表語の大きさを取得する．

閾値に関しては，基準概念 X と，その概念 X に対して関連が非常に強い概念 A，概念 A ほどではないが関連があると思われる概念 B を 500 セット作成し，概念 X と概念 A との関連度の平均値，概念 B との関連度の平均値の 2 つの値の平均値(0.19)を使用した．2 語の関連度がこの閾値以上であれば，その 2 語は関係があると考えられる．

### 4.2 判断

我々は厳密な大きさ判断を行わず，曖昧性を考慮して比較を行う．これに習い，提案未知語処理では，大きさの幅を考慮した判断を行う． 入力語①の最小の大きさが入力語②の最大の大きさ以上である場合は「①の方が②よりも大きい」，逆の場合は「①の方が②よりも小さい」という判断を行う．それらでない場合には「なんとも言えない」という判断を行う．

### 4.3 評価結果

4.1 節と同様の方法で新たに作成した 100 セットに対し，既存システムと提案システムによる大小判断を行った．また，3 人に対してアンケートを行い，その結果を集計した．この 100 セットのうち，76 セットは 2 語のどちらか，または両方に未知語を含む．

表 2　評価結果

| 手法 | ○ | × |
|---|---|---|
| 既存システム | 48 | 52 |
| 提案システム | 57 | 43 |

表 2 より，○は 9%上昇したので，提案手法の有効性が示されたといえる．

## 5. 考察

提案手法は既存の量判断システムよりも 9%精度が向上した．しかし，入力語がシソーラス内に存在せず，かつ概念ベースに存在しなかった場合，関連度計算が行えないため，未知語の大きさを代表語，もしくはノードから導き出すことは不可能である．この問題を解決するためには，WEB の利用が考えられる．新語や複合語，固有名詞等であっても，日常的に我々が会話に用いる概念である以上 WEB 上に必ず存在する．

さらに，提案未知語処理において，関連度計算を用いてシソーラスノードの特定と，代表語への置換を行っているが，関連度計算は意味の近さを表すものであるため，大きさの近さを考慮した未知語処理とは言えない．本来は大きさの近いシソーラスノードの特定，代表語との置換を行う必要がある．

## 6. まとめ

本稿では，量判断システムにおいて，「大きさ」に注目し，大きさに関する常識を判断可能なシステムの実現を目的とした．4 章で述べた提案手法を実装することで，9%の精度向上に成功し大きさの幅を考慮する有意性を示した．しかし，新語等への対応や，大きさの近さを考慮できる未知語処理が必要であると考えられる．

**謝辞**

本研究の一部は，科学研究費補助金（若手研究（B）24700215）の補助を受けて行った．

**参考文献**

[1] 佐藤祐介, 渡部広一, 河岡　司，“常識的量判断システムの構築-量に関する相対的評価の拡張-”，情報科学技術フォーラム FIT2007，pp.283-286，2007.

[2]奥村紀之，土屋誠司，渡部広一，河岡司，“概念間の関連度計算のための大規模概念ベースの構築”，自然言語処理，Vol.14, No.5，pp.41–64，2007.

[3] 渡部広一, 奥村紀之, 河岡　司，“概念の意味属性と共起情報を用いた関連度計算方式”，自然言語処理，Vol.13，No.1, pp.53-74，2006.