E-038

N 語連想を用いた単一文の概念化
Single Sentence Conceptualization by N Words Association Method

三瀬 慶久† 渡部 広一† 河岡 司†
Yoshihisa Mise Hirokazu Watabe Tsukasa Kawaoka

1. はじめに

人間と会話する機械としてのコンピュータには会話の基本となる十分な国語知識が求められる．本稿では，国語知識の一環として，ある言葉の意味を表現する単一文からその文の意味に該当する語を推測して導く，単一文を概念化するシステムの構築方法について提案している．入力した文章を形態素解析して得られた自立語の集合を基に，答の候補となる単語を抽出し，それらを答候補語とする．そして概念ベースや関連度計算，シソーラス等を用いて答候補語から単語を絞っていき，最終的に一つの答語を導く．本提案システムは単一文を単語に変換することによる文章圧縮や，単語の意味は知っているが，単語自体が思いつかない場合などに活用できる．

2. 使用技術

2.1 概念ベース

概念ベース[1]とは，ある単語（概念）と，その意味的特徴を表す属性と重みの集合で構成されたものである．ある概念A に対して，その語の i番目の属性をa_i ，重みをw_i ，概念 Aの属性数をN個とすると，概念A は以下の式(1)のように表される．

$$A = \{(a_1, w_1), (a_2, w_2), \ldots, (a_i, w_i), \ldots, (a_N, w_N)\} \quad (1)$$

属性 a_iを概念A の一次属性と呼ぶ．一次属性 a_iを一つの概念と見なせば，a_i からさらにその一次属性を導くことができ， a_iの属性 a_{ij}を概念A の二次属性という．これを展開していくと，一つの概念A はn 次属性まで持つことができる．概念ベースの例を図 1 に示す．

概念	属性, 重み
雪	(雪, 0.61) , (白い, 0.30) , …
白い	(雪, 0.16) , (白地, 0.14) , …
…	…

図 1　概念ベース

2.2 関連度計算

関連度計算[2]とは，概念と概念の関連の強さを定量的に評価するもので 0 と 1 の間の実数値で表す．値が大きいほど，概念間の関連が強い．

2.3 シソーラス

シソーラス[3]とは，一般名詞の意味的用法を表す約 2700 の意味属性(ノード)の上位下位関係，全体部分関係を木構造で示したものであり，約 13 万語(リーフ)が登録されている．

†同志社大学大学院工学研究科
Graduate School of Engineering, Doshisha University

2.4 関係辞書

関係辞書とは，同義語，類義語，上位語，反対語の関係にある語を登録したデータベースである．

3. 単一文の概念化手法

本システムは，言葉の意味を表した文章（意味文章）を入力するとその意味文章に見合った単語を出力する．例えば，「円で割合を表したグラフ」と入力すると，システムが「円グラフ」と出力する．

システムの流れを図 2 に示す．

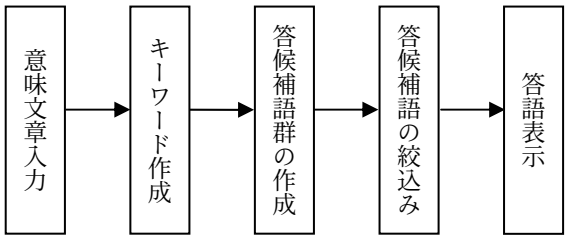

図 2　システムの流れ

3.1 キーワード作成

意味文章を形態素解析して獲得した自立語より，品詞が動詞，名詞，形容詞，副詞の語を抽出する．あらかじめ人手で作成した不要語リストより不要となる語（「こと」や「する」のように頻繁に出現し，あまり意味をなさない語）を削除し，残った語をキーワードとする．否定語（「ない」や「ぬ」）が存在すれば，否定語の前の単語を関係辞書を用いて反対語に置き換える否定語処理を行う．概念ベースに存在しない複合語（「探し求める」や「走り出す」など）がキーワードに含まれればその語を分解して，概念ベースに存在する語に置き換える．

3.2 答候補語群作成

概念ベースよりキーワードを属性として持つ概念全てを答候補語とする．

3.3 答候補語の絞込み

3.2 で作成された答候補語に対して絞込みを行い，最終的には 1 つの語に絞る．流れとしては，「属性による絞込み」，「関連度計算による絞込み」を順に行う．そして，意味文章の末尾にある語（末尾語）が答語を導く際に有効な単語であれば（「人」や「乗り物」など），末尾語処理を行い，そうでなければ「関連度計算による絞込み」を用いて 1 つの語まで絞込む．

3.3.1 属性による絞込み

答候補語の属性に出現する語とキーワードを比較することにより，答候補語の絞込みを行う．

図 3 は属性による絞込みの例である．キーワードは，意味文章「自分の家に帰ること」より生成された「自分，家，帰る」である．答候補語の属性とキーワードとの一致する語の回数を調べ，一致回数が最大になる答候補語のみを残し，他の答候補語を削除する．

答候補語	答候補語の属性					一致回数
面会	面談	接見	拒否	自分	・	1
帰宅	帰る	家	自分	所属	・	3
家路	帰途	帰る	家	自分	・	3
門衛	門守	家	番人	自分	・	2
・・	・	・	・	・	・	・・・

帰宅，家路，・・・・・・・

図 3　属性による絞込み

答候補語の属性とキーワードを比較する際，表記情報が一致する場合のみカウントするだけでは，同義語や類義語のような意味的にほとんど近い語には対応できない．そこで各キーワードと非常に関連の深い語を概念ベースと関連度計算を用いて獲得する．それらの語もキーワードとみなし，答候補語の属性とキーワードを比較する．

3.3.2　関連度計算による絞込み

答候補語と各キーワードとの関連度の平均値を求め，平均値の高い上位 5 件の答候補語を残す．答候補語と各キーワードの関連度計算を行う際，一律に関連度計算するだけでは，キーワードごとの重要性が考慮されない．キーワードの中には意味文章に深く関連する語も存在するので，一番最後に出現するキーワードを重要キーワードとする．そして，答候補語と各キーワードとの関連度計算をする際，重要キーワードのみ重みを 2 倍に付与して関連度計算を行う．

図 4 は関連度計算による絞込みの例である．重要キーワードは「帰る」となっているので，答候補語と「帰る」との関連度の値を 2 倍にする．

答候補語	キーワード			平均値
	自分	家	**帰る**	
朝帰り	0.01	0.01	**0.05*2**	0.04
帰宅	0.19	0.81	**0.20*2**	0.46
帰休	0.05	0.70	**0.08*2**	0.30
・・	・・	・・	・・	・・

帰宅，帰休，・・・・・・・

図 4　関連度計算による絞込み

3.3.3 末尾語処理による絞込み

意味文章の末尾語は，キーワードの中でも意味文章を表した語（元語）と特に関連が深い場合が多い．シソーラス上において，元語と末尾語は親子関係，あるいは兄弟関係にある場合がよく見受けられるので，この関係を満たす語のみを答候補語に残す．

図 5 は末尾語処理による絞込みの例である．意味文章「春・夏・秋・冬の四つの季節」に対して，末尾語は「季節」である．この時点で残っている答候補語が「四季，時候，季刊，四時」とすると，シソーラス上において「季節」と兄弟，あるいは親子関係を満たす「四季」や「時候」を答候補語として残す．

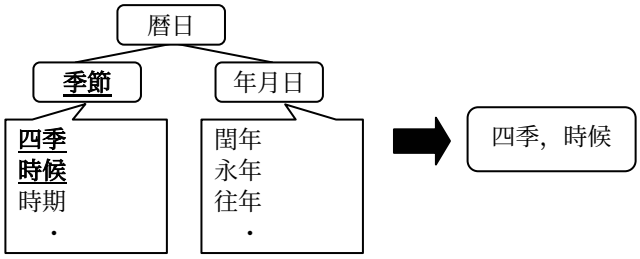

図 5　末尾語処理による絞込みの例

4. 評価

テストデータとして国語辞典[4]に掲載されている名詞の意味文章 400 文に対して評価を行う（図 6）．出力された語が，元語と一致していれば○，異なっていても意味文章から推測されるものが出力されたら△とした．

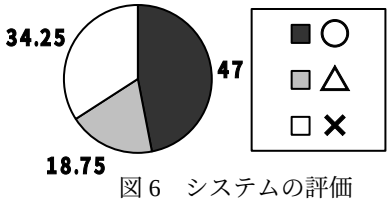

図 6　システムの評価

○と△の合計を正解語とすると，正解語の割合は 65.75%であった．

5. 考察

本システムの評価として，65.75%の正解語が得られた．答候補語群が作成された段階で，評価が○，あるいは△となる正解語は含まれている．しかし，答候補語の絞込みの過程で，多くの正解語が削除されてしまう．

その大きな要因としては，入力された意味文章をキーワードに分解していることである．例えば，「弟」の意味文章である「年下の男の兄弟」をシステムに入力すると，「兄」が出力される．これは，意味文章「年下の男の兄弟」に対しては間違っているが，キーワード「年下，男，兄弟」との関連を考えると，妥当であると言える．意味文章をキーワード化することにより，文章の持つ意味が曖昧になってしまい，意味の似た語の分別ができない．

6. おわりに

本研究では，ある言葉の意味を表した文章から，元の語を推測して導く手法を提案した．これにより，コンピュータが意味を説明した文章を，適宜に同じ意味の概念語に置き換え表現することが可能となった．

国語の知識は非常に広範囲に渡っており，今後より一層の知識拡大が望まれる．

参考文献

[1]小島一秀，渡部広一，河岡司，“連想システムのための概念ベース構成法－語間の論理的関係を用いた属性拡張”，自然言語処理，Vol.11，No.3，pp.21-38，2004.
[2]渡部広一，河岡司，“常識的判断のための概念間の関連度評価モデル”，自然言語処理，Vol.8，No.2，pp.39-54，2001
[3]NTT コミュニケーション科学研究所，“日本語語彙体系”，岩波書店，1997
[4]金田一京助，“例解学習国語辞典　第八版”，小学館，2004