概念ベースを用いた「要するに」の推定

1H－1

笠原 要†　松澤 和光†　金杉 友子‡
†NTT(株) コミュニケーション科学研究所 ‡NTT アドバンステクノロジー(株)

1　はじめに

国語辞書から単語の連想関係に関する知識ベース(連想ベース)を構築し、これを用いて文章を要約する単語(「要するに」)を推定する方式を提案する。

文章を要約する研究の一つとして、国語辞書から説明文を要約する単語、すなわち上位語を獲得する研究が行なわれている[1]。そこでは説明文の構造をモデル化し、構文解析等の手段を用いて上位語を獲得する。この場合、説明文中に上位語が含まれていない場合には獲得できず、さらに、説明文の汎用的な構造のモデル化ができないことが問題である。しかし人間は、どのような場合でも文章の「要するに」を獲得できる。これは、「拾い読み」、すなわち、文章の論理構造を重視せずに重要な単語を拾い読みし、さらにその単語から連想することで獲得していると考えられる。

本稿では、人間の行なうような「拾い読み」により文章から「要するに」を表す単語を推定する方式を提案する。まず、常識に関する２つの知識ベースの構築に関して説明し、次に辞書中の見出し語の「要するに」を表す上位語を獲得する方法と、一般的な説明文の「要するに」を表す主旨語を獲得する方法を提案する。最後に、上位語の獲得に関する予備実験の結果について説明する。

2　概念ベースと連想ベースの構築

我々は、人間の常識的な推論の計算機での実現を目指し、「アバウト推論」と名付けた研究を進めている[2]。そのためには、言葉の意味、すなわち概念に関する常識的な知識が必須であり、辞書から４万の日常語の概念に関する知識ベース(「概念ベース」)を構築し、多様な観点に応じた変化する概念の類似性判別を実現した[3]。また、同様に辞書を用いて４万語に関する連想知識ベース(「連想ベース」)を構築し、クロスワードパズルの自動解答に応用した[4]。

概念ベースでは、個々の概念を、特徴を表す単語である属性とその重みの対の集合で表現している。辞書の説明文中の自立語の出現頻度に基づいて１つの概念について平均して44属性を自動獲得した。また、辞書中の見出し語に対する説明語は、見出し語の概念を構成する特徴であり連想のキーと考えられる。そこで、辞書中の説明語に対する見出し語を連想語、説明語の出現頻度を連想語の重みとした連想ベースを辞書から機械的に抽出した。、１つの単語にたいして平均して16語の連想語が獲得されている。表1に概念ベースと連想ベース中の「愛」に対する内容を示す。

表1:「愛」の概念知識と連想知識(一部)

概念ベース		連想ベース	
属性名	重み	属性名	重み
愛	5.7	愛	5.0
気持ち	9.6	純愛	4.0
心	9.4	愛情	3.0
愛する	1.1	花言葉	2.0
恋愛	8.8	哲学	2.0
愛情	7.1	児	2.0
思う	4.5	子	2.0
渇き	2.9	求愛	2.0
大切	2.6	恩寵	2.0
対す	1.3	覚え	2.0
表す	9.3	嘉する	1.0
付く	8.3	櫓	1.0
異性	8.2	賞でる	1.0
名詞	8.1	目出度い	1.0
人類	7.5	参る	1.0
子	5.8	二筋道	1.0

3　単語の上位語の獲得

単語の説明文から説明語を抽出し、連想ベースを用いて説明語を連想し、上位語を獲得する。その方式を図1に示す。説明文を分かち書きし、自立語を抽出し説明語とする。そして、説明語のそれぞれについて、連想ベースを参照し、連想語とその出現頻度を獲得する。最後に同じ連想語の出現頻度はまとめ、最も頻度の高い連想語を上位語とする。その際に、見出し語と同じ単語は連想語から除く。また、見出し語が概念ベース中の概念として含まれる場合には、概念ベースの属性について連想ベースを参照し、連想された単語の回数のみから単純に上位語を獲得することもできる。

4　文章の主旨語の獲得

文章の主旨を表わす単語（主旨語）文章を構成する単語（構成語）の連想により生成する。主旨語は、必ずしも構成語とは限らない。方法は、前記の上位語獲得法と同じく構成語について連想ベースを参照し、連想語の出現頻度を総合する方法がある。また、構成語について、その特徴語（属性）を概念ベースを参照して獲得し、属性について連想ベースを参照して連想語を獲得し、出現頻度に基づいて主旨語を獲得する方法もある。

5　実験

概念ベース[5]の概念の属性から連想される単語が、概念の見出し語の上位語になっているかについて実験を行なった。

表2に、上位語の獲得結果の一部と、比較のため対応する単語の角川類語辞典[6]での上位語を挙げる。類似

A Method for Inferring Summarized Words from a Word and a text
Kasahara. K.†, Matsuzawa. K.†, and Kanasugi. T.‡
†NTT Communication Science Laboratories ‡NTT Advanced Technology
1-2356, Take, Yokosuka-shi, Kanagawa 238-03 Japan
kaname@nttkb.ntt.jp

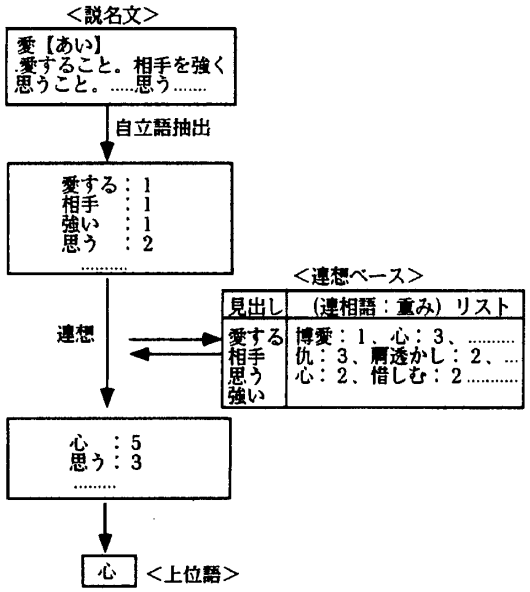

図 1: 上位語獲得の方式

語辞典の上位語は、分類の見地から見ては適切であるが、「要するに」という観点からは不適切なものもある(愛犬)。一方、提案手法では、ほぼ上位語が獲得されており、その上、「要するに」をよく反映した結果となっている。

表 2: 上位語の獲得結果

単語	上位語 (提案手法)	上位語 ([6])
亜	次	副
愛	心	愛憎
愛育	育てる	養成
合縁奇縁	縁	縁
哀感	感じ	悲嘆
哀願	願う	愛
愛犬	犬	愛
愛妻	妻	愛、妻
哀史	物語	説話
妹	女	姉・妹
嫌がる	思う	憎

次に、「愛」の上位語を決定する際に、上位語を連想する重みを 3 種類の手法で計算した場合の結果を表 3に示す。手法 1 では、概念ベースと連想ベースの重み情報を用いずに連想語のの参照回数のみから重みを計算している。また、手法 2 では、概念ベースの重みの情報のみを用い、手法 3 では、両方の重み情報の積を用いている。この例では、手法 3 において、適切な上位語の重みがもっとも大きくなり、さらに、「愛」の類似語が含まれていないので、良い方式といえる。今後、評価手法の検討も含めて評価を行なう予定である。

表 3: 「愛」の上位語

手法 1		手法 2		手法 3	
上位語	重み	上位語	重み	上位語	重み
子	25	子	64	心	415
心	25	心	59	気	401
人	24	人	59	御す	377
愛情	20	児	58	子	236
御する	20	愛情	52	為る	223
目	20	目	51	様だ	194
児	19	頭	48	児	180
頭	18	気	48	方	179
鬼	18	手	48	吾輩	177
手	18	吾輩	47	等	174
情け	18	御する	46	目	173
掛ける	17	情け	46	言う	170
気	17	参る	46	思う	170
吾輩	17	愛する	45	頭	162

6 おわりに

本稿では、単語の上位語や文章の主旨語等「要するに」を表わす単語を辞書から自動構築した連想ベースと概念ベースを用いて獲得する方法について提案した。今後は、提案方式の定量評価による方式の有効性の確認を行う予定である。また、共通の上位語を持つ単語を分類する事によるシソーラスの作成も検討している。

参考文献

[1] 鶴丸弘昭, 竹下克典, 伊丹克企, 柳川俊英, 吉田将: 国語辞典情報を用いたシソーラスの作成について, 情報処理学会自然言語処理研究会, Vol. 83-16, pp. 121–128 (1991).

[2] 松澤和光, 石川勉, 河岡司: アバウト推論とその類似性判別機構, AI 学会研究会資料, Vol. SIG-J-9401, pp. 103–110 (1994).

[3] Kasahara, K., Matsuzawa, K. and Ishikawa, T.: Refinement Method for a Large-Scale Knowledge Base of Words, *Working Papers of the Third Symposium on Logical Formalizations of Commonsense Reasoning*, pp. 73–82 (1996).

[4] 松澤和光, 金杉友子, 石川勉: 概念ベースによる類似性判別の「クロスワードパズル」への応用, *11th Fuzzy System Symposium*, Vol. 1, pp. 319–320 (1995).

[5] 笠原要, 藤本和則, 松澤和光, 石川勉: 精錬に基づく概念ベース構成法, 信学技報, Vol. DE95-7, pp. 49–56 (1995).

[6] 大野晋, 浜西正人: 類語国語辞典, 角川書店, 4 edition (1990).