ユニバーサル概念ベースの構想

1R−5

笠原 要† 松澤 和光† 石川 勉‡

†NTT コミュニケーション科学研究所 ‡拓殖大学工学部情報工学科

1 はじめに

我々は、人間が不完全な知識の下で行う概括的な判断を計算機上で実現する「アバウト推論」の研究を行なっており [1]、人間の常識に匹敵する大規模で不完全な知識を前提とした、新たな推論方式を検討している。常識のモデルについては、これまでにAIで様々検討されたが、その有効性の検証のほとんどは、小規模なToy modelレベルにとどまっている。そこで我々は、常識の基本をなす言葉の意味(概念)に関する大規模な知識ベース(概念ベース)を辞書から自動構築する方式を検討し、日常語4万を含む概念ベースについて、類似性判別において評価を行なった [2]。

概念ベースの検証において明らかとなった問題点は、ユーザとの入出力において、4万語の日常語では不十分であることである。例えば、固有名詞や専門語、流行語などは含まれていないが、これらの一部は、常識として必要である。また、ユーザの入力は、表記あるいは読みで行ない、表記揺らぎを考慮して概念を指定する方針を取っている。しかしユーザは実際に、概念ベースに含まれていない表記やひらがな混じり、、適切ではない表記で入力する場合が多いことも明らかとなった。さらに、既に獲得された4万語の概念の品質も十分ではなく、質的向上が必要である。

そこで、上記の問題点を解決する手段として、ユニバーサル概念ベースの構想を提案する。基本的アイデアは、概念ベースに含まれていない単語であっても、既に獲得されている概念を用いて近似的に表現・合成することにある。また同時に、既に獲得されている概念の質の向上も、ユーザからの入力を通して同時に実現する。

2 概念ベースの構築と類似性判別方式

これまでに検討を行なった、概念ベースの構築手法と、概念の類似性判別方式について、説明する。概念間の関係は、固定的ではなく、状況や文脈に応じて多様に変化する。例えば「馬」の概念は、動物の話しをしていれば「豚」に似ているが、乗りものの話しでは「豚」よりも「自動車」に似ている。類似性を判別する際には、こうした状況に応じた変化を考慮する必要がある。そこで、状況を表す単語を「観点」として指定し、その下で概念間の類似度を計算する方式の提案を行なった。

まず、各概念が特徴を表す属性と重みの対で表された知識ベース(「概念ベース」)を、国語辞書の語義文中の自立語とその出現頻度から獲得する。例えば、「馬」の概念は、(四つ足, 1), (家畜, 3), (運搬, 1), ・・・の様に表される。次に、属性を既存のシソーラスの分類名に変換し、分類数次元の意味空間での概念の比較を可能とする。そして、観点に指定した概念の属性に応じて一部属性の重みをを強調し、概念同士の距離に基づいて類似度を計算する。また、複数の国語辞書を用いて4万語の日常語について平均44属性を保有する概念ベースを自動構築し、判別方式の有効性を明らかにした。図1は、類似性判別の一例である。

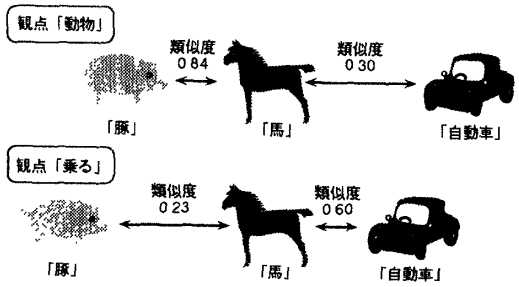

図 1: 類似性判別結果の一例

3 ユニバーサル概念ベース

前節で説明した概念ベースを元にして、より多数の語彙と柔軟な入力に対応するための枠組みが、汎用的概念ベース、すなわちユニバーサル概念ベースである。拡張の方針として、これまでに獲得された概念を用いて、未知語の概念の表現・合成を行ない、次の段階として、概念知識の種類の拡張を行う。これは、人間が未知語を、知っている語の組み合わせ等で近似的に理解する方法を模したものである。我々は、人間の言語生活の基幹となる日常語の概念の獲得を試みてきたので、これだけでも、かなりの拡張が期待できる。また、概念知識の種類の拡張では、専門語や固有名詞などの知識の性格に応じた表現・獲得を目指し、日常語の概念を獲得する方法論を単純に適用はしない。

また我々は、概念とはその体系(語彙)のなかでこそ良く定義されると考えている。そこで、上記のように拡張されたユニバーサル概念ベースにおいて、日常語の概念を再定義し、その品質・精度を高める方法論を検討する。このようなユニバーサル概念ベースの構想を図2に図示する。ユニバーサル概念ベースは、以下のような手順で段階的に構築を行う。

3.1 概念マッピング【既知語の概念で代用】

概念ベースにおいて、概念の表記や読みの記号情報は辞書より自動獲得しており、複数辞書の参照により、1つの概念に対して複数の表記情報を保有している(平均1.5表記/概念)。しかし、ユーザの入力に対する概念の指定は完全一致に基づいているので、人間ならば容易に連想・指定できる文字列であっても、概念を指定できない場合がある。例えば、概念ベースに「事柄(ことがら)」の表記で登録されているが概念は、「事がら」というユー

Formulation of Universal Concept Base of a word

Kasahara. K.†, Matuzawa. K.†, and Ishikawa. T.‡

†NTT Communication Science Laboratories ‡Takushoku University

‡1-1, Hikarinooka Yokosuka-shi, Kanagawa 239 Japan

kaname@cslab.kecl.ntt.co.jp

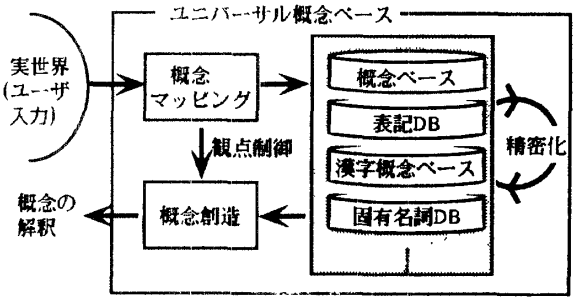

図 2: ユニバーサル概念ベースの全体図

ザ入力で指定できない。これは人間が、複数の種類の記号情報を用いて柔軟に概念を推定しているが、そのような方法論が概念ベースに備わっていないためである。また、人間ならば誤りを容易に訂正できるような表記であっても、現状の概念ベースでは指定困難である。そこで、概念の判定を文字情報の一致から、「それらしい」文字列の判定へと拡張する。これを概念マッピングと呼ぶ。

これを実現するためには、ユーザの入力した文字列に対して、複数の記号情報から総合的に概念ベース中の概念を指定する方式を検討する必要がある。これについては、現在検討中であるが、漢字変換 FEP や OCR の研究における知見が利用できると考えている。

3.2 概念合成［既知語の概念の組み合わせ］

人間は、複合語の意味を知らなくても、その構成語の意味を知っている場合、複合語を近似的に理解できる。そこで、概念ベースに含まれない語を含まれている語に分解し、分解した語の概念を組み合わせて、概念を合成する方法論を採用する [3]。合成の規則としては、構成語間の構文関係等を利用する方法があるが、我々は、構成語の表す概念のベクトルを単純和し、観点に応じた複合語の概念ベクトルを変調して表現する。

3.3 拡張の複合化［合成とマッピング］

上記のマッピングと概念合成を組み合わせて、より多くの単語の概念を表現する。ユーザからの入力「運どう会」に対して、対応する概念ベース中の概念を指定する部分について、図 3 にその方式の一例を示す。

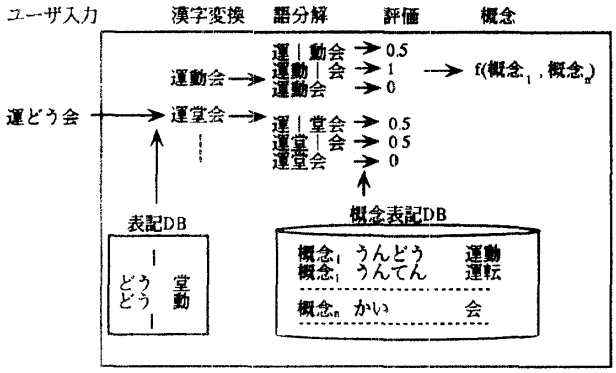

図 3: 複合語の概念マッピング

3.4 概念知識の複合化［概念知識の拡張］

上記の拡張によって、任意の単語の概念を表現することが可能であるが、入力文字列全てを概念ベース中の日常語の組合せで表現できない場合には、合成された概念の精度が低くなる。そこで乱暴ではあるが、漢和辞典を用いて、個々の漢字の概念を日常語の概念と同じ形式で表現・獲得し、漢字概念ベースと組み合わせてユーザの入力文字列に応じた概念を表現する方針を取り、現在、漢字概念ベースを構築中である。

また、専門語や固有名詞などの場合には、概念の表記とそれが指す実体・現象が大きく異なっている場合が多い。そこで、このような特殊性が高い単語については、あらかじめ、日常語との対応表をあらかじめ作成しておき、それを用いて概念を表現する。

さらに、新聞記事中の単語同士の共起関係に基づいて、概念を表現し、連想検索に用いる研究も同時に進められている [4]。これにより、流行語や時事語の概念の表現が可能となり、ユニバーサル概念ベースで用いる予定である。

3.5 獲得された知識の精密化［概念の精錬］

このように多数の概念を表現し、利用することができると、その利用結果に基づいて、基本となる日常語の概念や、その他表記情報に関する問題点が明らかになる。この結果を負の教師データとして、概念の質を向上することが可能となる。また、現在、概念ベースを用いたネットワーク型言葉遊びを開発し、それを用いた知識獲得を提案している [5]。これをユニバーサル概念ベースに適用することも可能と考えられる。

4 おわりに

本稿では、ユーザからのどのような文字列の入力であっても、日常語の概念ベースの拡張によって概念を表現するための、ユニバーサル概念ベースの構想について提案を行なった。その方法論として、入力文字列の概念マッピング、複合語の概念の合成方式、日常語以外の概念を獲得する方法の検討を段階的に進める構想である。今後は、実際に構成部分のそれぞれについて方式検討を行なう予定である。

References

[1] 松澤, 石川, 湯川, 河岡: アバウト推論 －「常識的な推論」を目指して－, AI 学会人工知能基礎論研究会, Vol. SIG-FAI-9401-1, pp. 1–8 (1994).

[2] 笠原, 松澤, 石川: 国語辞書を利用した日常語の類似性判別, 情処論文誌, Vol. 38, No. 7 (1997).

[3] 永森, 金杉, 笠原, 松澤: 概念ベースを用いた複合語概念の合成, 第 55 回情全大, Vol. 2R-04 (1997).

[4] 松澤, 飯田, 松田, 今井: 想起型情報検索システムの基本構想, 第 53 回情全大, Vol. 1T-4 (1996).

[5] 金杉, 笠原, 阿部, 松澤: 「ネットワーク型言葉遊び」の知識獲得への応用, 第 55 回情全大, Vol. 2R-05 (1997).