博士学位論文審査要旨

2008年1月15日

論文題目： 大規模概念ベースの自動構築と概念連想方式の研究

学位申請者： 奥村 紀之

審査委員：

主査：	工学研究科	教授	河岡 司
副査：	工学研究科	教授	三木 光範
副査：	工学研究科	教授	渡部 広一

要 旨：

コンピュータに人間のような連想能力を持たせることができれば、わずかな入力からでも柔軟に連想し、人間と機械の意思疎通を支援することが可能となる。本研究の目的は、連想メカニズムの実現であり、その基盤となる概念ベース構築法と概念連想方式を確立することである。

本研究では、電子化新聞等を用いて約 12 万語規模の大規模概念ベースを構築する手法を確立した。また、概念ベースに登録されている任意の 2 つの概念間の関連の強さを定量化する方式として、より人間の感覚に近い結果の得られる動的関連度計算方式を開発した。これらの研究成果により、日常会話で使われる殆どの語に関する概念連想機能を実現することができた。

1 章の序論では、研究の背景と研究の目的を明らかにしている。2 章では、概念ベースと関連度計算プログラムからなる連想メカニズムの全体構成と研究の基本となる概念の表現法および概念と概念の関係を定量的な関連度により表現する方法について述べている。3 章では、日常会話に必要な大規模概念ベースの構築方式について論じている。電子化国語辞書の語彙説明文から作成出来る基本概念ベース（約 4 万語）は日常的な会話に対応するには語彙が不足しているという問題点があった。残り 8 万語に対して、語の属性を電子化新聞記事より取得する手法を開発した。この方式では、国語辞書の語彙説明文から抽出できる属性に比べて質の悪い属性が取得されるため、精度の高い精錬技術が必要となる。属性信頼度と概念価値の考えに基づく新しい重み付与方式を考案しこの問題を解決している。新方式で構築した大規模概念ベースは殆どの日常会話語彙に対応できるばかりでなく評価実験により精度についても約 20%の向上が見られることを確認している。4 章では、大規模概念ベースをシソーラスにより、さらに改善する方法を論じている。シソーラスは新聞記事などの文書情報と比べ、親子関係など特定の関係について信頼性の高い属性を取得可能である反面、冗長性が高くなり絞り込みが必要となる。実験により、属性として、親ノードは全て、子ノードは 10 語、仲間語を 15 語追加することによってほぼ最適な概念ベース（73.3%の評価値）を構築できることを明らかにしている。5 章では、大規模概念ベースを用いて、語と語の関連の強さを、より人間の感覚に合うように定量化する動的関連度計算方式について論じている。従来方式では、個々の概念の固定的な属性で関連の強さを評価するのに対し、提案方式では、評価対象概念が決まった時点で、2 つの概念に共通する属性を重視して評価する人間的な特性を考慮した動的関連度計算方式を考案し実験によりその有効性を示している。6 章では、構築した大規模概念ベースと関連度計算を用いて、提示された語から関連の強い複数の語を想起させる語連想システムを構築しその有効性を実験により確認している。

7章では、連想メカニズムの適用例として時間判断メカニズムについて述べている。時間判断メカニズムは、「秋刀魚の季節→秋」、「日が沈む→夕日」、など会話の中で現れる時間に関する文や語から常識的な季節や時間を連想するメカニズムで精度は約80%である。時間判断の他、場所判断、感覚判断、感情判断などのメカニズムにも適用され、コンピュータの常識的な判断に基づく会話処理の実現に大きな役割を果たせることを確認している。

本研究で開発した大規模概念ベースと関連度計算プログラムで構成される連想メカニズムは、常識的な判断能力をもち人間と自然な会話のできるコンピュータの実現に重要な役割を果たすものであり、人間と機械とのインタフェイスの改善に大きく寄与することが期待できる。

よって、本論文は、博士（工学）（同志社大学）の学位論文として十分な価値を有するものと認められる。

総合試験結果の要旨

2008年1月15日

論文題目： 大規模概念ベースの自動構築と概念連想方式の研究

学位申請者： 奥村 紀之

審査委員：

主査：	工学研究科	教授	河岡 司
副査：	工学研究科	教授	三木 光範
副査：	工学研究科	教授	渡部 広一

要 旨：

本論文提出者は、本学大学院工学研究科知識工学専攻博士課程前期課程を2005年に修了後、2005年4月より本学大学院工学研究科知識工学専攻博士課程後期課程に在学している。この間、各年度において優れた研究成果をあげ、英語の語学試験に合格し、ドイツ語についても十分な能力を有すると認定されている。また、本論文の主たる内容は、自然言語処理学会論文誌（3編）、同志社大学理工学研究報告（2編）、Proc. of KES（4編）、Proc. of AIA（1編）等に掲載されており十分な評価を得ている。

2008年1月12日午後1時から約1時間半にわたって提出論文に関する学術講演会（博士論文公聴会）が開かれ、種々の質疑討論が行われたが論文提出者の説明により十分な理解が得られた。

さらに、講演会終了後、審査委員により論文に関連した諸問題につき口頭試問を実施した結果、十分な学力を有することが確認された。

よって、総合試験の結果は合格であると認める。

博士學位論文要旨

論文題目： 大規模概念ベースの自動構築と概念連想方式の研究
氏名： 奥村 紀之

要 旨：

コンピュータはネットワークと結合し、日常生活のあらゆる分野で使用される重要な社会インフラの一つとなっている。インターネット検索や文書処理に便利なパソコンの機能も高度化・多様化する反面、複雑で使い難いものとなってきている。また、自動車、航空機、携帯電話、カメラ、エアコンなど多くの製品にもマイクロコンピュータが組み込まれ、多種多様な機能が提供されているが、機能の高度化・多様化・複雑化に伴い、人間が正確に指示を与えることが困難になってきている。これらの問題を解決するには、コンピュータにも自然会話能力や常識的な判断能力を持たせることが必要となる。本論文は、これらの機能の中核となる概念連想メカニズムの実現方式に関するものである。

人間は、長年の経験によって培った「語」に関する膨大な情報を記憶しており、その情報の中から、状況に応じて必要な語を連想することによって、コミュニケーションを成立させている。コンピュータに人間のような連想能力を持たせることができれば、わずかな入力からでも柔軟に連想し、人間と機械の意思疎通を支援することが可能となる。本研究の目的は、連想メカニズムの実現であり、その基盤となる概念ベース構築法と概念連想方式を確立することである。連想メカニズムには、ある語からその語と密接な関係のある語を連想する機能と、その語に代替できる同義語あるいは類義語を連想する機能の2種が必要となる。前者は、会話等において使用される語の幅広い意味解釈に、また、後者は使用された語に関する知識がない場合にも知識の定義されている語に代替することにより会話の意味を適宜に理解するために必要となる。

本研究では、すでに開発済みの電子化国語辞書から作成した小規模概念ベース(約4万語)を基本に、電子化新聞を用いて約12万語規模の大規模概念ベースを構築する手法を確立した。また、概念ベースに登録されている任意の2つの概念間の関連の強さを定量化する方式として、より人間の感覚に近い結果の得られる動的関連度計算方式を開発した。これらの研究成果により、日常会話で使われる殆どの語に関する概念連想機能を実現することができた。

1章の序論では、概念連想に着目し研究を行うに至った背景と共に研究の目的を述べ、目的達成に必要な技術課題を中心に論文の各章構成について述べている。

2章では、会話処理や常識的判断を行う上で重要となる連想メカニズムについて述べる。連想メカニズムは大きく概念ベースと関連度計算方式によって構成されている。概念ベースは、語の意味(概念)を、その語の特徴(属性)と特徴の重要度(0~1)で表現した離散ファジィ集合として定義した知識ベースである。また、関連度計算プログラムは、語と語の同義・類義関係のみならずあらゆる関係を対象に、語と語の意味的関連の強さを定量的に評価する手法である。概念連想メカニズムは、これら二つの要素を用いることにより、人間の感覚に似た常識的で柔軟な連想を行う仕組みである。

3章では、特殊な固有名詞や専門用語を除く一般的な日常会話において出現するほぼ全ての国語辞書に掲載されている語彙(約12万語)に対応できる大規模概念ベースの構築方式について論じる。これまでに開発されていた基本概念ベースには、およそ4万語の概念が定義されているが日常的な会話に対応するには語彙が不足しているという問題点があった。この基本概念ベースは、電子化国語辞書の語彙説明文から作成されおり、辞書に記載されている約12万語の内の残り8万語については殆ど属性を抽出することができない。これらの8万語については、基本概念ベ

ス構築方式では対処できず新たな方式の開発が必要となる。そのため、基本概念ベースを基に残り8万語の概念に対する属性を電子化新聞記事より取得する大規模概念ベース構築手法を開発した。この方式では、国語辞書の見出し語に対する語彙説明文から抽出できる属性に比べて質の悪い属性が取得されるため、精度の高い精練技術が必要となる。このため、取得した属性に対し、属性信頼度と概念価値の考えを導入し新しい重み付与方式を考案した。基本概念ベースと新たに構築した大規模概念ベースの連想精度に関する比較評価実験では、両概念ベースに共に定義されている概念に関しては、基本概念ベースの評価値54.8%、大規模概念ベースの評価値73.2%を得ることができた。また、拡大した概念を含む全体的評価においては、基本概念ベースの評価値7.4%、大規模概念ベースの評価値71.2%を確認できた。開発した新方式により基本概念ベースの性能を大幅に向上させると共に殆どの日常会話の語彙に対応できる約12万語の大規模概念ベースの構築に成功した。

4章では、大規模概念ベースをシソーラス(日本語語彙体系)により、さらに改善する方法を論じている。国語辞書の語義説明文と新聞記事から構築した大規模概念ベースの平均属性数は約30個であるが、概念の中には数個の属性しか付与されていないものや、400を超える属性が付与されているものがあり、その分布は疎らである。そのため、より連想に適した概念ベースに改善する手法として、シソーラスからの属性追加方法を考案した。シソーラスは、人手で作成された語彙に関する大規模な木構造の知識ベースであり、新聞記事などの文書情報と比べ、親子関係など特定の関係については信頼性の高い属性を取得することが可能である。しかしながら、一方では、同じノードに所属する語をすべて属性として付与する場合、例えば、「ほ乳類」のノードに所属する「犬、猫、猿、ライオン、虎、…」などの語は、その性質が多少の変化を持つものの、概念の属性としては、同種の属性が大量に付与され概念を効率よく特徴づけることができない。そのため、属性として、概念の親ノード、子ノード、兄弟語(共通する親ノードを持つノードまたはリーフ)をどの程度の割合で追加するのが適切かを実験により明らかにした。その結果、3階層上位までの親ノードは全て、子ノードは10語、兄弟語を15語追加することによってよりほぼ最適な概念ベース(73.3%の評価値)を構築できることを明らかにした。

5章では、大規模概念ベースを用いて、語と語の関連の強さを、より人間の感覚に合うように定量化する動的関連度計算方式について論じている。従来の方式では、対象となる2語それぞれの重み上位30属性における一致度を基に関連度を算出している。さらに、特定の条件下において正しく関連度を算出できないという問題点があった。これに対し、提案方式では個々の概念の固定的な属性で関連の強さを評価するのではなく、評価対象概念が決まった時点で、2つの概念に共通する属性を重視する考えに基づき、重み下位の属性であっても、共通属性を優先的に抽出し、残りを重み順で対応づけることにより、人間の評価により適合する関連度が得られることを実験により示した。動的関連度計算方式の開発により大規模概念ベースの評価値を75.3%まで向上させることに成功した。

6章では、構築した大規模概念ベースを用いて、1つの単語から複数の語を想起させる1語連想、および、2語の入力から関係の深い複数の語を連想させる2語連想メカニズムについて論じている。1語連想においては、対象となる概念の重みの大きい1次属性と1次概念(対象とする概念を属性として持つ概念)を候補語として入力2語と候補語の平均関連度を評価する方式により、精度84.2%、再現率70.0%を得た。また、2語連想においては、対象とする2語に共通する2次属性、並びに2次概念を候補語とし、精度65%、再現率45%を得ることができた。

7章では、連想メカニズムの応用について論じている。代表的な応用例は時間判断システムの構築である。時間判断システムは、「秋刀魚の季節→秋」、「日が沈む→夕日」、「海で泳ぐ→夏」など会話の中で現れる時間に関する文や語から常識的な季節や時間を連想するメカニズムで、その正答率は69.4%、精度は約81.6%である。時間判断の他、場所判断、感覚判断、感情判断などのメカニズムにも適用され、コンピュータの常識的な判断に基づく会話処理の実現に大きな役割

を果たせることを確認している。

他の重要な応用例は概念ベースにもシソーラスにも定義されていない特殊な固有名詞や新語(未定義語)をシソーラスのノードに対応付け意味を同定するメカニズムの構築である。本研究では、WEB から未定義語に関連する語を収集し、それらに連想メカニズムを適用することで未定義語の意味を同定するメカニズムを開発した。大規模概念ベースに定義されていない 200 語をテストデータとして実験した結果、66%の割合で未定義語を所属すべき親ノードに割り付けることができた。この未定義語の意味同定メカニズムは概念ベースやシソーラスの自動拡張に重要な役割を果たす。

8章では、論文全体を総括し、今後の展望について述べている。

本研究で開発した大規模概念ベースと関連度計算プログラムで構成される連想メカニズムは、常識的な判断能力をもち人間と自然な会話のできるコンピュータの実現に極めて重要な役割を果たすものであり、今後の情報社会における人間と機械とのインタフェイスの飛躍的な改善に大きく貢献することが期待できる。