Japio

自然言語処理と常識の使用

人間の連想に基づく常識の抽出

慶應義塾大学環境情報学部教授
石崎　俊

PROFILE
1970 年東京大学工学部計数工学科卒、同助手を経て 1972 年通産省工業技術院電子技術総合研究所（現、独立行政法人産業技術総合研究所）勤務、1985 年推論システム研究室室長、自然言語研究室長を経て 1992 年から慶應義塾大学環境情報学部教授、1994 年から政策メディア研究科教授兼任、現在に至る。言語処理学会会長、日本認知科学会会長を歴任。

ishizaki@sfc.keio.ac.jp TEL

慶應義塾大学 SFC 研究所上席研究所員
岡本　潤

PROFILE
1997 年慶應義塾大学環境情報学部卒。1999 年同大学院政策メディア研究科修士課程修了。同年同 SFC 研究所上席研究所員（訪問）、現在に至る。

juno@sfc.keio.ac.jp TEL

慶應義塾大学 SFC 研究所上席研究員
寺岡　丈博

PROFILE
2005 年慶應義塾大学環境情報学部環境情報学科卒業。2007 年同大学院政策・メディア研究科修士課程修了。同年同 SFC 研究所上席研究員（訪問）。2008 年より同政策・メディア研究科後期博士課程に在籍、現在に至る。

teraoka@sfc.keio.ac.jp TEL

1 はじめに

21 世紀に入って Web 情報の普及や世界のグローバル化の進展に伴って、自然言語処理技術は多くの場面で使用されるようになり、それに伴って自然言語処理関連技術が著しく発展してきている。しかし、人間の言語能力と比べると、現在の自然言語処理のレベルはまだかなりのギャップがあるのが実情である。機械翻訳システムを試しに使ったことがある人は、人間はしないような、いわば初歩的な誤りを見ることがあると思われる。

Web における言語情報の使用や言語資源の拡充によって、大量の言語データを自然言語処理研究に使えるようになっている。それらによって統計的な手法を多用して自然言語処理システムの性能を向上させる研究が行われてきている。私自身も以前に音声情報処理の研究をしたことがあり、統計的な手法を駆使してシステムを構築したことがある。従来の研究開発で人間の手作業で書いていたルールを自動化し、パラメータの最適化に大変有効である。このようなやり方は自然言語処理システムの工学的研究開発の立場からはオーソドックスなアプローチということができる。しかし、そのような性能の向上は、人間の言語処理における仕組みや機能にはほとんど直接的な関係がない。人間に近いある一定レベル以上の性能を目標にする場合は、人間の大脳における言語処理の複雑な過程を参考にする必要がある。

その一つのアプローチとして、まず、言葉の背景にある膨大な構造化された知識を利用することがある。それは、文章として並べられたものではなく、概念体系といわれるもので、近年ではオントロジーと呼ばれている。このような概念体系の例としては、米国では WordNet が有名であり、ヨーロッパ諸国でもいくつかの言語ごとに構築されている。我が国でも EDR による概念体系がある。

しかし、これらのオントロジーでは階層関係における上位と下位という体系だけの情報しかないため実際の自然言語処理応用システムには使いにくい状況である。もし、二つの概念間の近さを計量しようとすると、上位ま

たは下位へ辿った枝の数を利用することが考えられる。ところが、枝の数は概念体系における局所的な粒度（概念の細かさ·粗さ）によって決まるため、客観的メジャーとして使用できないのが理由である。

人間の場合は大量の概念が複雑に相互結合して意味ネットワークの構造をしており、幼少のころから学習によって紡ぎあげられた概念体系になっている。それをモデル化したものとして連想概念辞書 [1] がある。これは人間の直感に基づいた大規模な連想実験のデータに基づいて、刺激語とその背景にある多様な概念を体系化したものである。刺激語と連想語の間の連想距離を定量化することによって、自然言語処理システムへの応用をやりやすく工夫している。

本稿では、そのような連想概念辞書の構築法を解説し、それを用いるアーキテクチャとして神経回路網を用いて人間の脳記憶モデルを作る。それを用いるメタファー理解システム、文書の自動要約システム、曖昧な語の意味解析システムなどを紹介し、最後に、最近進めている動詞の連想概念辞書のその応用について触れる。

2 連想概念辞書

連想実験システムは、慶応義塾大学湘南藤沢キャンパス (SFC) のコンピュータネットワーク上で多数の被験者が同時に実行可能なシステムで、大量の連想データを収集し、それを用いて連想概念辞書を構築するものである。連想のための刺激語は小学校の教科書に出現する基本名詞を中心にしている。刺激語に対して上位概念、下位概念、部分／材料概念、属性概念、類義概念、動作

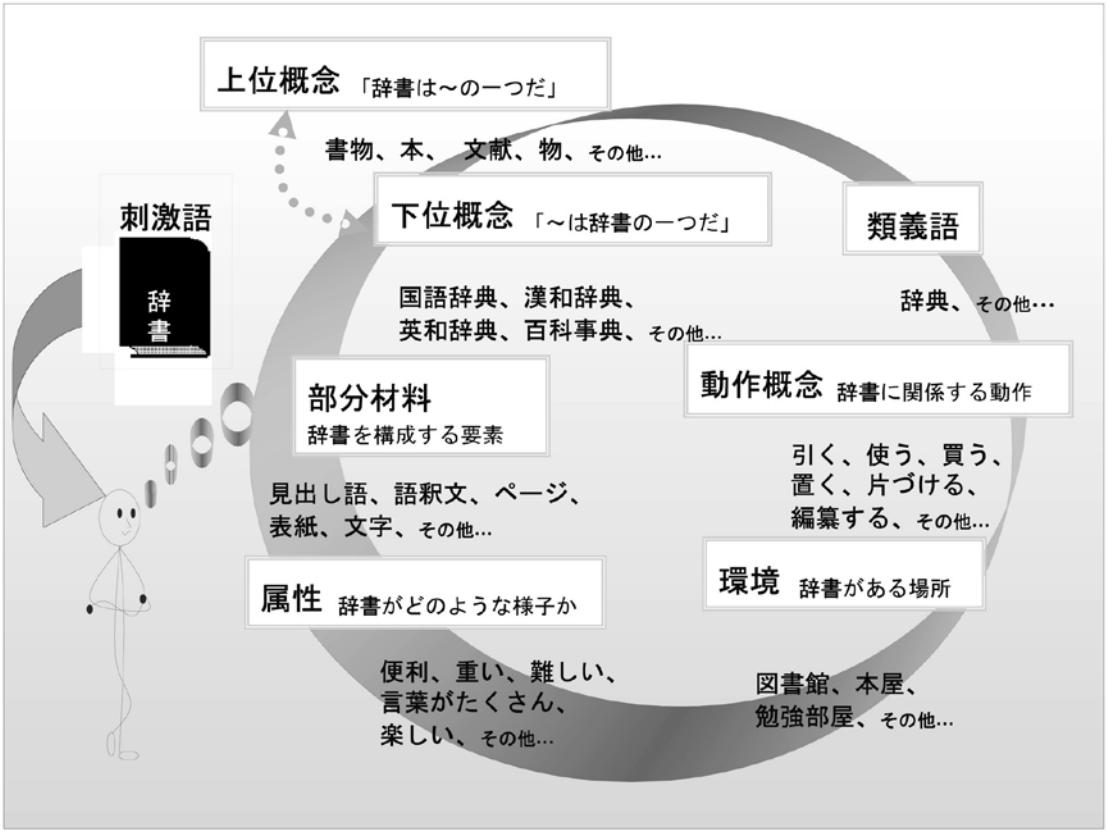

図 1　連想実験システムの概要

概念、環境概念という７つの課題を与えて連想させている。被験者は同キャンパスの学部生と大学院生で、毎年多数の協力を得て連想データを蓄積し、整備している。　この連想概念辞書は無料で一般公開しており、要望のあった大学の研究室や企業の研究所に配布しているので、興味のある方は石崎 (ishizaki @sfc. keio.ac.jp) までご連絡いただければ案内を送ることができる。現在公開中の連想概念辞書の規模は、刺激語数は約 1650 語余り、連想語は延べ約 13 万語、異なり語で約３万３千語となっており、被験者数は１刺激語あたり 10 名である。

なお、この公開版とは別に、１語当たり 50 人の被験者の連想概念辞書を構築中であり、刺激語数は 1055 語、延べ連想語数は約 25 万語、異なり語数は約５万６千語で、それを用いた応用システムを試作している。

刺激語と連想語の距離が計算できれば、応用範囲が極めて広いことが分かっているので、線形計画法という手法を用いて果物や乗り物などの身近な概念について概念間の距離を計算する方式を決め、連想概念辞書全体に適用してその有効性を示してきている [1]。図２はブドウを中心にした上位概念と下位概念の距離を見やすく２次元表示したものである。

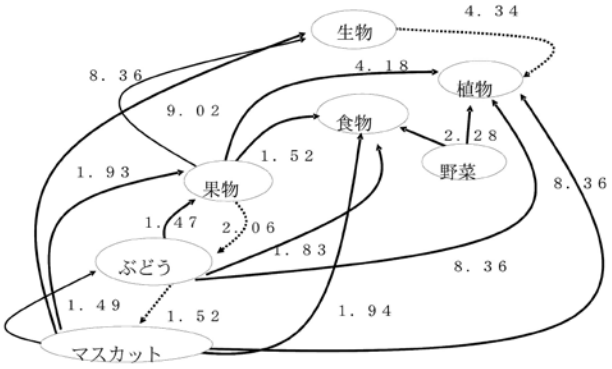

図２　ブドウを中心とした概念間の距離

3　連想概念辞書の応用

3.1　比喩の解析　[2]

日常生活でたいへん多く用いられている比喩表現は、たとえる語 (Source 概念) と、たとえられる語 (Target 概念) があって、それぞれの名詞に共通する属性概念を発見して人間は比喩理解する。

（１）比喩理解のアルゴリズム

連想概念辞書では、名詞を刺激語にしたときに、その名詞の属性概念として連想した概念が多量に蓄積されている。名詞から属性概念への距離も定量的に使用できる。そこで、ニューラルネットワーク (神経回路網) を設定して、ニューロンに語を割り当てて相互結合し、ニューロン間の長さを、連想で得られた距離に対応させて情報伝搬の仕組みを実現する。

「まるで S のような T だ」という形式の比喩文を対象とし、まずこの S（Source）と T（Target）から連想された属性における共通概念の検索を行う。たとえば「まるで雪のような肌だ」という比喩文がシステムに入力されたとする。「雪」と「肌」のニューロンをそれぞれ発火させ、発火の活性伝播による連鎖ダイナミクスを発動させる。比喩の場合は形容詞のような属性を表す表現が中心なので、図３の右下に示すように刺激語の名詞層からつながった属性層のみを活性化させる。

システム内では「雪」と結合した属性層のニューロンとして「白い」、「冷たい」、「サラサラ」等が、一方「肌」と結合したものとしては「白い」、「きれい」、「スベスベ」等が活性化する。特に「白い」は「雪」、「肌」の共通概念であり、両者からの刺激を同時に受けるので、他と比較してニューロンの活性値が大幅に上昇する。その結果、システムではこれを最適な比喩理解候補と考え、最終的に「その肌はとても白い」という理解

結果に至る。

この様に、比喩理解候補はシステム内で高い活性値を獲得したニューロンに相当する。出力された比喩理解候補それぞれに対し、表１のように確信度というパラメータを、比喩理解候補に対するシステム自身の信頼性を表わすものとして与えることができる。

例文	比喩理解候補	確信度（%）
まるで雪のような肌だ	白い	95
	冷たい	41
	きれい	11
まるで雪のような心だ	冷たい	95
	白い	39
	きれい	11
まるで鬼のような人間だ	怖い	95
	強い	88
	大きい	78

表１　比喩理解システムの出力結果

一方、Source 概念と Target 概念との間には影響力の違いがある。Source 概念の影響力をより強調したいときは、システムへのそれぞれの概念の入力タイミングに時間差を設ける事によって実現することができる。たとえる語と、たとえられる語を入れ替えた場合の例として、

「まるで人の様なロボットだ」 → 「賢い」

「まるでロボットの様な人だ」 → 「冷たい」

がある。このような比喩文では、たとえられる語よりも、たとえる語の影響の方が強いために比喩理解結果にも効果が現われ、直感的な理解に近い結果になる。

3.2　多義語の曖昧性解消 [3]

日本語でも基本的な語ほど、複数の意味をもっていて文中では曖昧な場合が多く、人間は文脈によって直観的に意味がわかるが、コンピュータにとっては大きな問題

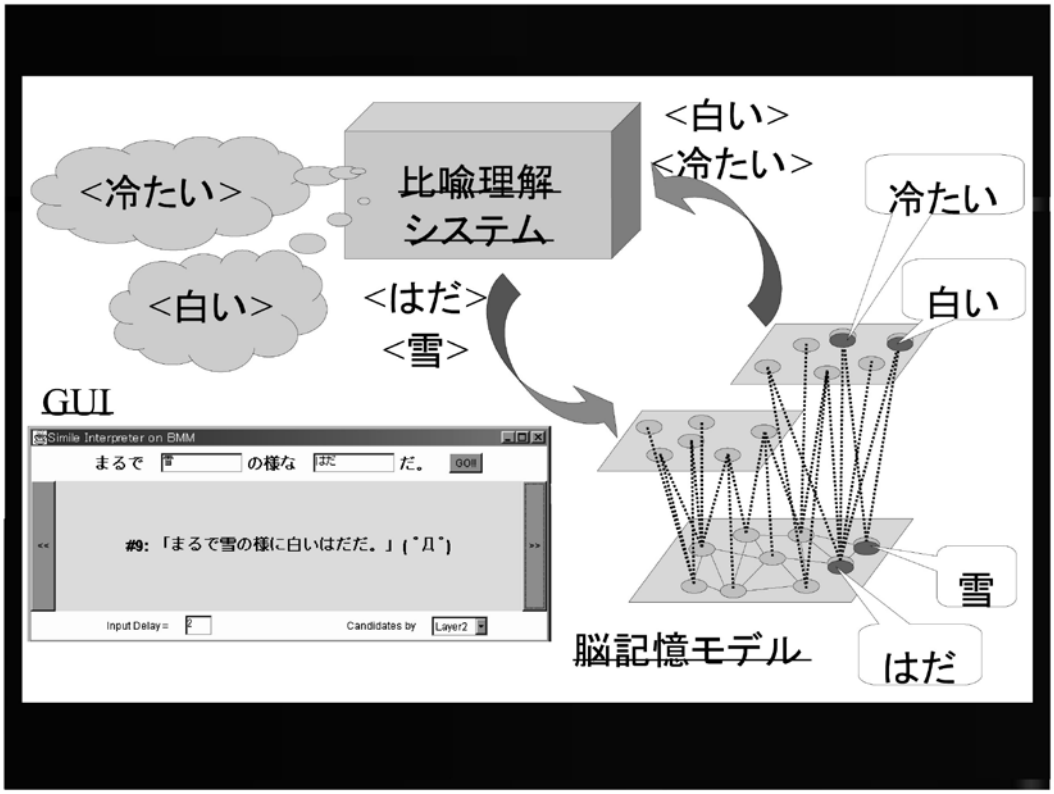

図３　比喩理解システム

である。

文中の語の意味的な関係の構造は、脳の神経細胞の結合のモデルであるニューラルネットワークを用いて自然に表現できるので、3.1 節と同様に、文中の語をニューロンとして相互結合してニューラルネットワークを構成する。各ニューロンは活性値と呼ぶ数値を持つことにして、ニューロン i の活性値を Vi とし、ニューロン i とニューロン j の間の相互結合には重み Wij を係数として与える。この重みは、連想距離に対応する。単位時間ごとに活性値を隣り合うニューロン間でやりとりをすると、次式のようにネットワーク上を活性値が伝わって収束していく。ここで、 n は 1 から増えていく自然数で単位時間の倍数である。

$$V_i(n+1) = V_i(0) + \sum_j WijVj(n)$$

多義語の意味をそれぞれ別の節点として、関連する語をそれらの周りに結合して、ニューラルネットワークを構成する。上記の活性伝搬が収束したときに、活性値の大きい方を選択すれば、多義語の意味の中から適切な意味を選択することができる。

3.3　連想距離を用いる重要文の抽出 [4]

テキスト中の刺激語と連想語の距離と、語の頻度を総合することによって、注目する語の重要度を計算し、重要文の抽出を行なうことができる。

まず、テキストに含まれるすべての語の重要度を大きさの順に選択することによって、重要な語を順番に取り出すことができる。つぎに、このように計算した語の重要度を文ごとに合計することによって、今度は文の重要度を計算することができる。重要度を用いて並べれば、テキストの中から重要な文を選んで要約することも可能になり、従来の表層的な情報を用いるやり方よりも、人間の要約結果に近いことが示されている。

4　おわりに

最近の研究では、動詞を刺激語にした連想実験を大量に実施して、動詞の持つ深層格にしたがった連想語を構造化して、動詞連想概念辞書を構築している。ある動作の動作主、対象、場所、時間などから連想語を得て、刺激語と連想語の距離を定量化する。第 2 章で述べた名詞の連想概念辞書と一緒に用いると、たとえば、「家に財布を忘れたので、友達から借りました」という文で、「借りた物」を推論して、「かね」を導くことができる [5]。

このように本稿では、従来の工学的なアプローチの研究だけでは実現できないが、人間が直観的に理解するような仕組みの実現を目指す自然言語処理システムの研究をご紹介した。

参考文献

[1] 岡本潤 , 石崎俊 , 概念間距離の定式化と既存電子化辞書との比較 , 自然言語処理 , 第 8 巻 ,4 号 ,pp.37 - 54,2001

[2] 坂口琢哉 , 石崎俊 , 連想概念辞書の実装による意味ネットワークと比喩理解システムへの応用 , 情報処理学会研究報告 2003-MPS-46, pp.21-24,2003

[3]Jun Okamoto, Kiyoko Uchiyama, and Shun Ishizaki, A Contextual Dynamic Network Model for WSD Using Associative Concept Dictionary, LREC, Morroco,2008/5.

[4] 岡本潤 , 石崎俊 , 連想概念辞書の距離情報を用いた重要文の抽出 , 自然言語処理 , 第 10 巻 ,5

号 ,pp.139 - 151,2003

[5] 寺岡丈博、岡本潤、石崎俊、“動詞連想情報を用いる省略語の推定法と評価—動詞連想概念辞書の構築と応用—、” 言語処理学会第 15 回年次大会 ,2009 年 3 月、鳥取

[6] 天野真家、石崎俊、宇津呂武仁、成田真澄、福本淳一、自然言語処理、(株) オーム社、2007