グラフ理論を用いた日本語連想作文支援システムの開発と評価

東京工業大学　人間行動システム専攻、鄭在玲、三宅真紀、馬越庸恭、赤間啓之
catherina@dp.hum.titech.ac.jp

１．はじめに

インターネットはマルチメディアを活用し豊富な教材を頒布することで言語学習を興味深く効果的なものにするばかりでなく、計算工学関連の様々な知識と技術によって言語学習の射程そのものを大きく拡張しつつある。本研究では、グラフ理論を語彙関連に適用する試みの結果として、単語の連想情報を検索可能にするウェブベースの言語学習システム ACSS(Associative Composition Support System)を提案する。「連想作文支援システム」ACSS においては、単語の様々な連想情報がユーザにもたらされ、言語表現を想像力豊かなものにするとともに、新しい言語を学習する際にも、ユーザは未知の単語を孤立したものでなく他の単語の星座的配置のうちに位置づけることができる。このような意味論的な支援が、柔軟な単語学習と自由な作文学習に与える効果を調べるため、本研究ではまず、ACSS が出力する単語間の直接関連情報と自由連想情報について主観的な評価実験を行った。

２．「連想作文支援システム」ACSS

２－１．背景

教育工学分野、特に言語学習システム開発において、最近自然言語処理(NLP)技術が盛んに活用されてきている。例えば、楊(1999)［1］の場合、学習者が入力した作文における文法的な誤りを直し、フィードバックをするのに、NLP 技術を用いて最も効果的な結果が得られている。また、高林ら(2001)[2］は検索技術を基にした作文支援システムの開発によって、学習者が簡単な操作で作文における必要な単語の用法や表現などをやさしく検索できるようにした。このように今までの研究では、作文支援の基本的な発想として、Procter(1994)［5］が報告書で言及しているように、単語の用法や文法的な誤りをコンピューターによって正して欲しいという利用者の全般的な希望や要求に合わせるものであった。ところが、本研究では、これとは全く異なる観点から言語学習者の作文を支援する方法を提案する。これは、既存の構文的なサポートとは異なり、意味論的な側面からのサポートである。これは、上手に文章を書くのには文法的な知識ばかりではなく、豊富な語彙力もまた重要な問題ではないかという考え方から来ている。すなわち、学習者が、自分の考えや意見、感情などを多様な単語で表現できるようになれば、作文における流暢性が養われるということである。そして、単語間の連想情報を基にした単語学習により、豊かな lexicon が効果的に構築されれば、表面的な語彙力だけではなく、自然に連想力や想像力もまた向上するのではないかという仮説を踏まえている。

本研究での提案において、キーポイントである単語の意味情報や単語間の連想情報とは、以下のようなものを指す。たとえば、“信頼”の単語についての連想情報は、「信じて頼ること。」(三省堂提供「大辞林 第二版」より）という辞書での定義情報ばかりではなく、「顔見知り, 頼れる,親しい, 女友達, 友達, 友人, 故郷, 男友達, **信頼**, 信頼できる, 知り合い, 知人, クラスメイト, 恋人, フレンド, サークル仲間, 飲み会, 飲み仲間, 遊び友達, 悪友」のような連想関連単語クループをも含む。連想作文支援システム ACSS において、この単語グループは概念グループと呼ばれ、類似しているイメージや概念を表す単語の集合である。また、単語間の連想情報としては、ACSS は二つの単語の間の連想の流れを追う。たとえば, **“コオロギ”**と“**名刺**”のように関連が迂遠に見える二つの単語の連想関係として、「**コオロギ** → 親戚の家 → セミ → 祝い → かさばる → **名刺**」のような一連の単語パスを提示する。先の概念グループの情報を通じ、学習者はそれらが共有しているイメージから未知の単語の意味を推測したり、任意の二つの単語の連想の流れから一連の単語間の関連性を探したりして、単語が持っている表面的な意味ばかりではなく裏のイメージまで透視すること

で、言語に関して想像力や発想力を養う。さらに、提示される連想情報から自然に多くの単語に接することができ、語彙力をさらに広げることができると思われる。このような単語についての連想情報を提供する連想作文支援システム ACSS は、中級程度の日本語学習者を対象として開発され、指示言語は英語で表示されている。

２－２．構造と GUI

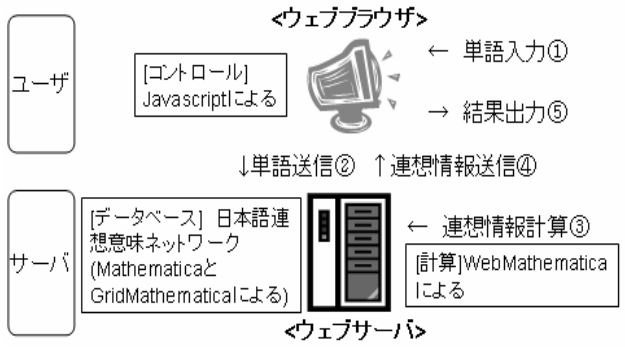

図１．ACSS の構造

ACSS は基本的にウェブベースのアプリケーションである。この構造は図１で示すように、クライアント（学習者）側とサーバ（教師）側に分けられ、ブラウザに学習者が入力した単語が、リモートのサーバ上でそのつど計算される形を取っている。サーバ上でこの連想情報の計算に必要なプログラムは Mathematica で書かれている。ただし参照されるデータベース、特に単語間の隣接情報は大量データなので GridMathematica であらかじめ構築されている。ユーザ側のデータ入出力は、Mathematica カーネルと JSP からなる WebMathematica が処理する。

図２は Web 上で利用できる ACSS の画面の例である。ACSS からは基本的に３種類の連想情報が提供される。1)連想単語グループ　2)直接連想　3)自由連想である。利用方法についてであるが、まず、学習者は連想情報の種類を一つ選び、ひらがなで一つか二つの単語を入力する。ACSS データベースから検索できる単語であるかを調べ、同音異義語の区別をするため、入力した単語のチェックをしてから、情報検索ボタンをクリックする(ACSS では、9,373 個の日本語の単語についての連想情報検索が可能である)。オプションとして、学習者は結果にふりがなを付けることができる。たとえば、連想単語グループを選択し、“**神社**”を入力した場合、「寺，寺院，鐘，**神社**，祈る，教会，偶像，感謝する，参る,参拝する，荘厳，厳か」が検索される。“**歩きやすい**”と“**レジャー**”に関して、直接連想情報としては、「**歩きやすい** → 道 → 旅 → **レジャー**」が、自由連想情報としては、「**歩きやすい** → 大通り → 観光地 → 公共物 → 計画する → 発つ → 険しい → **レジャー**」がACSS によって提示される。二つの単語の連想情報として、直接連想と自由連想のふたつがあるが、それらは、二つの単語を繋げる途中の単語がliteralでdirectな意味関係を基にしているか、metaphoricalでfreeな想像をベースにしているかによって区別されている。ACSSのデータベースが、グラフクラスタリング法を使って構築された連想意味ネットワークを基にしているため、このような連想情報が得られるわけである。

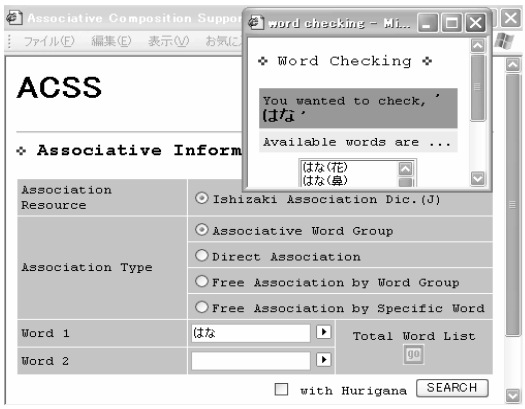

図２．ACSS　GUI

２－３．グラフ理論を用いた連想意味ネットワーク

ACSS の開発における最も核心的な作業として、あらかじめ連想意味ネットワークの構築を行った。連想意味ネットワークとは、単語がそれぞれ連想関係を基にして広げているグラフのことである。本研究では、Jung et al (2006)［3］が提案したグラフクラスタリング方法を利用し、単語レベルばかりではなく、すべての単語

を概念別にグループ化し、概念レベルでのネットワークを作成した。連想意味ネットワークのリソースとしては、刺激語から自由連想反応語を集めて作られた石崎概念連想辞書(日本語)(Okamoto, 2001)［4］を使用した(総単語数 33,018 語、単語ペア 240,093 個で構成されている)。しかし、実際の計算は、意味ネットワークの密度や計算条件を考慮し、頻度数が低い単語を外して、9,373 語だけを対象に行った(それによって単語ペア数は 187,113 に減った)。ここで選択された単語の隣接行列(9,373*9,373)をまず Markov Cluster Algorithm(MCL)にかけた。MCL とは Van Dongen(2000)［7］によって提案され注目を集めているグラフクラスタリング方法のことであり、random walk による Expansion と Gamma operator による Inflation という二つのステップを繰り返すことで、結果的に単語を重複なく(ハードクラスタリング法と呼ばれる)、概念別にグループ化させる特徴を持つ。この MCL によって、およそ 16 番目のクラスタリング段階で 9,373 語は 1,408 個のグループに分かれ、すなわち 1,408 個の概念クラスタが得られた。MCL による概念別グループ化の結果は、ACSS の‘連想単語グループ’という連想情報タイプのもとになる。次に、MCL 概念クラスタが、そのアルゴリズムのハードクラスタリングの特性のためクラスタ間に連結関係を失ってしまうことから、Recurrent Markov Cluster Algorithm(RMCL)(Jung et al 2006)［3］を考案し、単語間でなく単語グループ（概念）間の連想意味ネットワークを生成させた。RMCL は、MCL におけるクラスタリング途中の段階に戻し、最終クラスタ間の重複履歴を見つけることで､MCL 概念クラスタを再連結させるアルゴリズムである｡RMCL には原理を異にする二つのタイプのアルゴリズムがあり、本研究では MCL によるクラスタリング全過程から最終クラスタ間の過去の重複履歴を明らかにする Alibi-breaking タイプを使った。図３は MCL と RMCL による、概念別グループ化とそれの連結回復の結果を簡単に示している。

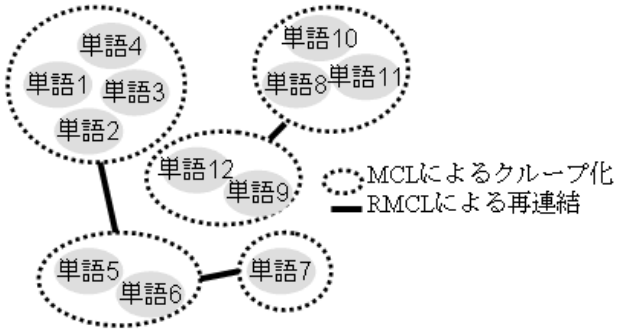

図３．MCL と RMCL モデル

３．システムによる連想情報の評価

３－１．Shortest Path

ACSS においては、二つの単語間の連想情報は、連想意味ネットワークからそれらの単語を最短でつなげるパスを探し、そのパス上の単語を示すことで求められる。最短パス問題は、もともとネットワーク科学の分野で主に扱われていることである。これを単語の意味ネットワークに適用すると、言語世界は small-world, scale-free で構成されていることが知られている(Steyvers et al 2003)［6］。本研究では、さらに、最短パスを計算する新しい方法を使用することで、独特の興味深い単語関係が得られた。単語 A と B の最短パスは、単語をノードとしている意味ネットワークから単語レベルで連結パスを探すのが一般的な方法である。ところが、新しい検索法では、MCL と RMCL によって構築された単語グループ(概念クラスタ)をひとつのノードとして考え、その概念ノード間の意味ネットワークから単語 A の単語グループと単語 B の単語グループを最短でつなげるパスを検索する。本論文では、最短パス計算方法について、前者を ordinary breadth-first shortest path(SP)で、後者を Markov Cluster shortest path(MCSP)と区別して呼ぶ。単語グループベースの新しい MCSP の方法は、単語レベルベースの SP とは異なり、計算時間を短縮させるだけではなく、豊かな自由連想関係を求めることを可能にする。たとえば、“**夏**”と“**許す**”の最短距離において、通常の SP は 9,373 単語のノードグラフから、「**夏** → 祭り → 神 → **許す**」が、MCSP は 1,408 概念グループをノードとしているグラフを基にして、「**夏** → 汗をかく → けなす → パンチパーマ → 仏 → **許す**」が ACSS によって求められる。すなわち、SP による結果は直接関連情報を、MCSP によるものは、自由連想情報を意味することになる。

３－２．評価

ACSS によって提供される SP と MCSP の連想情報における特性を明確にするため、客観的な側面と主観的な側面からの評価を行った。この評価には、構成可能な全単語ペアを母集団として、その約 1.0e-6 のサイズということで 9,373 単語からランダムで単語ペア 100 個を取り、サンプルセットを作った。次に 3 人の日本語母国語話者に、単語ペア間の主観的な距離をまず評価させ(1=一番近いから 5=一番遠いまで５尺度による)、単語ペアの距離が一番近い単語 12 ペア(2.333 以下のレート)と一番遠い単語 13 ペア(4.333 以上のレート)を求めておいた。次に、サンプルセットからそれぞれ SP と MCSP の計算を行った結果(100 個の単語ペアのうち、主観的な距離がいちじるしく遠い二つの単語のパスがメモリ不足のため計算できなかった)、平均パス長は、SP では 3.28、MCSP では 13.41 であった。だが、最も大きな差はその計算時間の差であり、平均すると SP には 84.43sec、MCSP には 15.55sec かかった(Windows XP, 1.67GHz で Mathematica5.1 による; $t(96)=5.97$, $p<.001$)。最後に、SP と MCSP の形で求められた連想情報について、20 名の日本人大学生が主観的に評価を行った。対象はサンプルセットからランダムに選ばれた 25 個の単語ペアである(一番主観距離の近い 12 個の単語ペアと一番遠い 13 個の単語ペアからなる)。具体的な評価は、“自然であるか”と“インスピレーションを与えるか”という２つの側面において、１=そう思わない ２=あまりそう思わない 3=どちらとも言えない 4=すこしそう思う 5=そう思うとする５段階からひとつ選択するという形を取った。２(MCSP vs.SP) x ２(主観距離近い vs.遠い)の分散分析を行った結果、“自然さ”の側面では、最短パス計算方法と単語間の主観的距離の双方の主効果が有意であり、交互作用は見られなかった($F_1(1,46)=47.03$, $p<.001$; $F_2(1,46)=20.32$, $p<.001$)。一方“インスピレーションを与える”という点では、どちらにも統計的に有意な差は見られなかった($F_1(1,46)=0.83$, $p<.1$; $F_2(1,46)=1.91$, $p<.1$)が、MCSP がわずかに SP の評価を上回る健闘を見せている。

４．まとめと今後の課題

本研究では、日本語連想辞書にグラフ理論を適用することで構築された連想意味ネットワークを基に、ACSS というウェブベースの作文支援システムを提案した。単語の連想情報による語彙学習は、語彙力のみならず連想力や想像力も向上させ、より流暢な作文を可能にすると想定される。今後の課題としては、英語の連想意味ネットワークの作成、より使いやすいシステムの実装、作文練習支援機能の追加、システムを通じた学習の認知的・教育的面からの評価などが考えられる。

５．参考文献

[1] 楊接期，赤堀侃司， 文章の結束関係を用いた科学技術日本語テキストの作成支援システム，第１５回日本教育工学会大会講演論文集，1999，pp. 323-324

[2] 高林 哲，松本 裕治， 検索技術を用いた作文支援，言語処理学会 第 7 回 年次大会発表論文集，2001，pp. 127-130

[3] Jung, J., Miyake, M. and Akama, H. (2006). Markov Cluster Shortest Path Founded upon the Alibi-breaking Algorithm, CICLing-2006, LNCS 3878, Springer Verlag Berlin Heidelberg, pp.55-58

[4] Okamoto, J., & Ishizaki, S.: Associative Concept Dictionary and its Comparison Electronic Concept Dictionaries 2001 http://afnlp.org/pacling2001/pdf/okamoto.pdf

[5] Procter, M. (1994). A Report on Software to Support Writing Instruction, University of Toronto

[6] Steyvers, M., Tenenbaum, J.: The Large Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth, Cognitive Science, 29(1) 2005, pp.41-78

[7] Van Dongen, S.: Graph Clustering by Flow Simulation. PhD thesis, University of Utrecht 2000,http://www.library.uu.nl/digiarchief/dip/diss/1895620/inhoud.htm