2005－CE－82（15）
2005／12／10

反復クラスタリングによる意味ネットワークに基づく作文支援システムの開発

鄭　在玲、三宅真紀、畑中伸幸、赤間啓之
東京工業大学大学院社会理工学研究科

概要：統語論的側面から文法的な誤りを正すためこれまで開発されてきた作文支援システムとは異なり、我々は本研究において、単語の辞書的語義だけでなく単語間の連想−連関についての意味論的情報を提示するという、異なる観点から作文支援を行う新しいシステムを提案、さらにそのシステムを Web アプリケーション化した例を紹介する。我々が開発したシステムは反復クラスター過程によって得られた意味ネットワークを基盤とし、学習者から入力された単語に対する連想単語を提示する形である。

For the Development of Composition Support System based on Semantic Network by Repeated Clustering

Jaeyoung Jung, Maki Miyake, Nobuyuki Hatanaka, Hiroyuki Akama
Dept. of Human System Science, Tokyo Institute of Technology

Abstract Unlike the composition systems to have ever been developed to usually correct the grammatical errors in the syntactical aspect, we propose here the new system to support the composition from the different respect by providing semantic information not on just the meaning of words on the dictionary, but on the associative relations between words. For this new composition system, what we developed first is the system which is as its resource based on the semantic network obtained by the repeated clustering process and which provides as its output learners with association words for the words they input.

1. はじめに

言語学習及び教育においては主に４つの領域、つまり、話し、聞き、読み、書きに分かれて考えられている。コンピュータを通じた言語学習でも、この４つのそれぞれの領域において言語学習を補助するシステムに関する研究や開発があいついで行われている。作文学習を対象とする分野では、多様な観点からそれを補助する方法やシステムが提案されてきた。だが、作文学習支援システムとは、学習者が書いたものに対して誤りの校訂やフィードバックの提供を行うものと通常考えられている。このような観点からの作文支援では、自由に書かれた人間の言語を機械的に解析・処理し、誤りを直すのに実際上は困難な点が伴うのだが、この困難を克服しようと、自然言語処理技術を応用する研究が盛んに行われている(楊, 1999)。

また、杉浦(2002)は、作文支援システムが考慮すべき条件をまとめ、学習対象である言語の語法に関し、学習者が大量の実例に効率的に触れることができるという意味で、コーパスの意義を高く評価している。実際、作

文支援システムにコーパスを用いる研究は近年増えつつある(楊,1999; 戸次,2002)。さらに、作文の過程は言葉を探す連続的行為であるとも言えるという観点から、作文における検索の必要性を重視した研究もある(高林,松本, 2001)。しかし、従来の研究では、ほとんど文法的な面が重視されており、語彙的支援は辞書をつけて単語の基本的な意味を提示するにとどまる場合がほとんどである。我々は作文学習においてまったく異なる視点から学習支援を行う。作文での語彙力の重要性に焦点をあわせ、いままで考案されていない、ある自由な連想に基づく語彙学習を介した作文支援システムを提案する。

2. 意味ネットワークに基づく作文支援システムの構想

言語を学習し、言語スキルを向上させる上で、助けとなる方法やツールはたくさん存在する。とくに、作文を対象とする研究から確かに良い方法論を提案した作文支援システムもある。日本語の作文学習を支援するシステムで既存開発された有用なシステムのひとつは、学習者から入力された文章において,自然言語処理技術を用いて統語論的側面から誤りを検出し、それに対する適切なフィードバックを与えるシステムである。(楊、1999)

しかし、コンピュータを使って作文の誤りを矯正させるのには、技術的に困難な面が伴う。作文とはそもそも文法的知識を利用して正しい文章構造を立てることであると同時に、豊かな語彙で文章構造を満たす作業でもある。作文学習においては確かにどちらも軽視できない。だが、開発面でのバランスを考え、我々は文法中心的学習という観点をひとまず措き、他の比較的緩やかな支援方法を重視するに至った。すなわち「いかに正確に」書くかではなく、いかに「自由に(流れるように)」書くかということを踏まえた新しいシステムを提案してゆく。このシステムは、単語の辞書的な意味ばかりでなく、他の単語との連想情報を提供する点、文法的なエラーチェックのような統語的支援とは一線を画す。単語の自由な連想情報に基づく我々のシステムは、語彙の世界を広げる上で有益な、豊かな言語データを言語学習者にもたらし、思考や意見を様々な言葉で表せるよう促してゆくものである。そのため、我々は、コーパスから意味ネットワークを構築し、後に述べるようにマルコフ・クラスター・アルゴリズムに基づく独自のグラフクラスタリング手法を用いて、セマンティックな側面からの作文支援システムを構築した。さらに単語の直接的な定義ばかりでなく、ゆるやかな概念結合の提示を可能にするため、意味ネットワークに対し、「アリバイ崩しアルゴリズム」と「マルコフ・クラスター最短パス」という独自の手法で、概念間の経路計算を行った。

3. 意味ネットワークの構築

3.1 背景

この章では、本研究における連想作文支援システムを開発する上で我々が考案した、新しいアルゴリズムについて概説する。単語の連想情報をコーパスから取得し、それを様々な形で使用するには、連想情報をひとつのグラフ、あるいはネットワークの形で表現すると便利である。本研究においても、グラフ操作を介して、連想作文支援システムを実現してゆく。しかし、コーパスからのグラフ情報を、人間の思考の自由でゆるやかな流れにマッチさせるためには、解決すべき大きな問題がひとつ存在する。本節ではまず、連想過程におけるこの問題についてとりあげる。

先端ネットワーク科学において、"small-world, scale-free"という特徴が普遍的に捉えられることはよく知られている。それとともに、ノード間の最短パスの問題もまた新たな関心を集めている。たとえば、Steyvers et al. (2003) によれば、単語間の意味ネットワークは、高密度に凝集した近傍と平均して短いパス長の双方によって特徴付けられる「小世界構造」を有している。彼らによれば、Nelson らの連想ネットワークのうち無向なものの平均最短パス長は 3.03、有向なものの平均最短パス長は 4.26、ロジェのシソーラス、WordNet の平均最短パス長はそれぞれ 5.43、10.61 であるという。

そのことは、本研究においてグラフ操作のための語彙連関情報を取得する目的で使用する「石崎概念連想辞書」においても同様である。43 個のランダムに選んだ単語対において、平均最短パス長は 3.442 であった。だがこのような低い値にもかかわらず、単語間の間に挟まった単語のパスをたどる通常の方法では、計算に平均して 1 分以上かかるという問題点がある。さらに、最短パス長の一様に低い値ゆえ、単語間の類似性/距離の指標としては、最短パス長をそのままの形では使用できないということが挙げられる。

3.2 マルコフ・クラスター・アルゴリズム

ところで、上記の最短パス問題を論じるうえで、マルコフ・クラスター・アルゴリズム(MCL)はきわめて重要である。これは、Van Dongen (2000)により提案されたグラフクラスタリングの手法であり、Expansion と Inflation のふたつのステップを、遷移確率行列が収束しグラフ全体が重複のないハード・クラスターに分割されるまで繰り返すものである。

本研究においては、Grid 上の GridMathematica を用い、MCL を前述の「石崎概念連想辞書」に適用する。この辞書は、10 人の被験者の連想に基づき、 33,018 語による 240,093 の単語対から構成されたものである。MCL アルゴリズムを適用する前の処理段階において、我々は希少語を除いた 9,373 語を含む連想対を取り上げ、有意味でバランスよく構成された意味ネットワークを形成させた。これらの重要語による 187,113 個の単語対から 9,373 行 9,373 列の隣接行列を計算し、これに MCL を適用して、16 回の反復計算後に、1,408 個のハード・クラスターからなるほぼ冪等な確率行列に収束させた。これらのクラスターは、それぞれ類似の単語群により維持される「概念」に対応するものである。

3.3 アリバイ崩しアルゴリズム

このように MCL により分割された概念クラスターは、MCL の最終的な収束段階においては、一つのノードはただひとつのクラスターに属し、その間にはオーバーラップがないので、そのままでは隣接関係をもはや有していない。そこで最短パスを探るには、概念クラスター間の連結を作り出す必要がある。最終クラスターの隣接行列を生成するためには、収束以前のクラスター段階における今や分離してしまった単語ノードの過去の履歴に遡り、概念クラスター間にヴァーチャルな連結を再現・修復せざるをえない。このため我々が提案する遡行的な過程では、まず前提として、各々の概念クラスターそれ自身が、新たな点ノード、あるいはメタ-点ノードとして捉えられ、それが含む単語のうち次数が最大な代表単語ノードにより命名される(同一性を与えられる)ことになる。さらにこの遡行的手続きは、異なる概念クラスターに含まれる各単語ノードが、過去のクラスター段階に

おいてどこかで一緒に帰属していたという「証拠」を集め、今では互いに異質なものと化しているが履歴のどこかで同じ単語ノードを保有していた概念クラスター間で再隣接化を行うということに存している。こうした手続きの側面ゆえ、我々はこのアルゴリズムを、過去の“implication（連累、含み）”の証拠をひとつひとつ取り上げるという意味で、「アリバイ崩しアルゴリズム」と呼ぶことができるだろう。

以下にこのアルゴリズムの各ステップを記すが、これはリカレントタイプのMCLの核心を為すものである。ClusterStageList は MCL のループがまだ回っている段階でのクラスター結果の集合を意味する。ただし、最後の要素である ClusterStage*k* は最終的な収束クラスターを表わすものとする。OverlappingNodes(ClusterStage*i*)という関数は、途中の各 ClusterStage*i* から、 oln(p)と略された多重帰属ノードを見つけるものとする。そして、OverlappingClusters(oln(p))という関数を用い、ClusterStage*i* において oln(p)を含む全ソフトクラスターの合併集合 olc(p) を生成する。各oln(p)に関して、olc(p)の中の過去のすべての共起ノードが列挙され、収束クラスター段階 ClusterStage*k* において conodes(p)を含むクラスターを求めることで、新たに最終クラスター間の隣接関係を設定し直すことになる。

ClusterStagesList=
{ClusterStage$_1$,ClusterStage$_2$,...,ClusterStage$_k$};
OverlappingNodes(ClusterStage$_i$)=
{oln(1),oln(2),...,oln(p),...,oln(m)};
OverlappingClusters(oln(p))=olc(p)=

$$\bigcup_j (\text{ClusterStage}_i(j) \supset \text{oln}(p));$$

For each oln(p){
conodes(p)=olc(p)$\cap\neg$ {oln(p)}
= {con(1),con(2),...,con(q),…,con(n)}};
MakeAdjacency(ClusterStage$_k$(j)$\supset$ conodes(p)); end

3.4 マルコフ･クラスター最短パス

最短パス探査において、幅優先探索（breadth-first）とは、連結グラフから全域木（spanning trees）を構成する形で、出発点のノードより発し、それに隣接する子ノードをすべて走査してゆく方法である。ここで幅優先探索（breadth-first）を採用する理由は、最短パスを使って各単一語の直線的配置ではなく、一連の「同系列要素語群（paradigm）」の配置を代表させようと考えているからである。マルコフ･クラスター最短パス(MCSP)もまた幅優先探索法の一種であるが、他の最短パスと区別される点は、単語ノードのひとつひとつに対してではなく、今度は自分自身が点と捉えられた概念クラスターの隣接行列に対して適用されるということである。

4. 意味ネットワークデータの結果と評価

我々はここで、「石崎概念辞書」から、母集団の 1.0e-6 のサイズの標本としてランダムに選ばれた 43 個の単語対に対し、以下に述べる3つのタイプの最短パス計算を行った。a) マルコフ･クラスター最短パス 1(MCSP1)：MCL プロセスから生じた 1,408 個のハード・クラスターのグラフから探査された幅優先探索（breadth-first）結果であり、結果としてクラスターを返すものである。b) マルコフ･クラスター最短パス 2(MCSP2)：マルコフ･クラスター最短パス 1(MCSP1)に基づくが、MCSP1 のクラスター結果をトレースして、その間に介在する単語の詳細なパスを特に返すものである。 c) 通常の幅優先探索パス(SP)であり、クラスター走査を

経ることなく、ローデータグラフから、2つの単語間において隣接関係を持つ単語を出力する。なお、この3つのタイプにおいてコアとなる幅優先探索関数は同一のものである。

結果として認められた傾向は、通常の SP を使うと2つの単語間で相対的に明確な明示的意味的関係をつかむことができるのに対し、MCSP はどちらかというと、単語の拡張的・伴示的な用法による自由連想の結果として連結した大きな意味領域を提示しているということである。量的データはというと、計算に要する平均時間に高度に有意な差が見られた。(Windows XP, 2.01GHz, Mathematica5.0 で、a),b),c)の各々に対し、平均 5.071 秒, 2.342 秒、84.487 秒であり、分散分析の結果は、$F(2,126)=16.066$, $p<.001$ であった). この点、MCSP1とMCSP2は、実用的なシステムに実装したとき、有効であることが判明した。

またパスの平均長は、a) 1.767, b) 17.277 そして c) 3.442 であり、そのことは、以下のようなことを意味している。すなわち、たとえ MCSP2 の結果が、その性質上、近似的かつ冗長的であるのが不可避であっても、これらの性質は、ある状況下では十分な長さの平均長ゆえにポジティブな結果をもたらすということである。それも意味ネットワークの「小世界構造」から帰結される情報不足を補完することによってである。さらに、これらの単語や概念のクラスターが理解する上で自然かどうか、作話する上でインスピレーションを与えるかどうかを見るため、3 人の被験者に 5 段階法でこれら 3 つのタイプの結果をいくつかの観点から評価するように要求したところ、SP の結果は作話インスピレーションより自然的精度の方に有利であり、この傾向は類似度の高いと言語学・言語教育の専門家が判断した単語ペアについてはさらに顕著であった。しかしながら、MCSP 2 の結果にはこのような差異は見出すことができなかった(図1)。

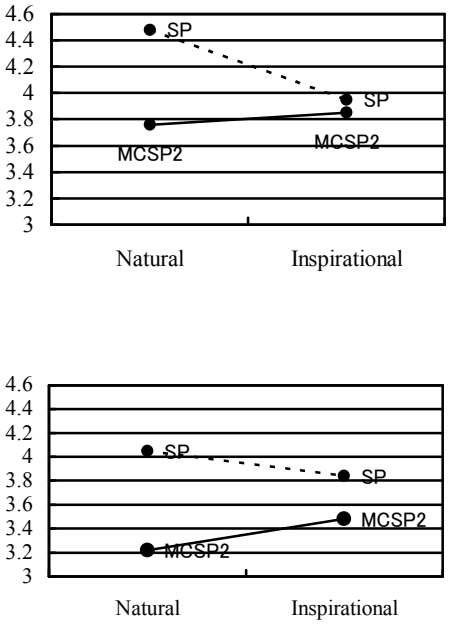

図 1: MCSP2 と SP による単語連関の最短パスの主観評価平均(ここではふたつの観点について5段階評価)。上のグラフは、類似した単語ペアを抽出した際の平均スコアであり、下は単語ペア例すべてに対するものである。

5. 連想作文支援システム ACSS

以上で構築われた意味ネットワークをもとに、我々は Mathematica を利用し、まず石崎連想辞書を実装して、連想作文支援システム ACSS(仮称、Associative Composition Support System の略)の開発に着手した。ACSS は、意味ネットワークに基づく作文支援システムのひとつとして、まだベータ版の段階ではあるが、単語の連想・共起情報から単語間および概念間の最短パスを提示し、学習者の作文を支援するシステムとなっている(図2)。現在開発されているシステムでは学習者が Web 上からアクセスし、2 つの単語を自由に入力すると、3つの連想情報タイプに基づいてそれらの周囲の単語が出力される。すなわち、各単語と一定範

囲でひとつの概念を形成する類似語、2 つの単語の間にダイレクトな最短パスを引く中間介在語、さらに 2 つの単語の間でより自由な連想を作り出す中間介在語である。たとえば、「論文」という言葉と「システム手帳」という言葉を入力した場合には、ダイレクトな最短パスは、論文，→図→メモ→システム手帳と出力されるが、より自由な概念連想を選択した場合には、中間に介在する単語数が増え、論文→{書，本，アイデア，あとがき，古本屋，ダイアリー，読み物，書籍，電話ボックス，理解する，捲る，手帳，専門書，写真集，レポート，身分証明書，カード，マガジン，記述する，記録する，確認する，ハードカバー，書斎，出版社，印刷物，パラパラ，単行本，参考書，週刊誌，接待，シール，背,表紙，書店，栞，スケジュール，スケジュール帳，ページ，資料，書き込む，本文，掌，文庫本，文献，メモする，合成革}→システム手帳が結果として返ってくる。

ここで、本システムの特徴を簡単に述べておく。Web アプリケーションなのでオンライン操作が可能であり、Internet Explorer など、ブラウザの文字コード処理能力を最大限に利用できることが挙げられる。すなわち、ACSS は、Web サイトにインタラクティブな計算機能を登載可能な WebMathematica を用いて開発を行った。WebMathematica は、Mathematica のカーネルと、Java Servlet 技術に基づいて開発されたツールであり、我々はすでに、これを用いた単語の共起情報取得システム、Tele-COEX などの開発実績がある(三宅,2004)。

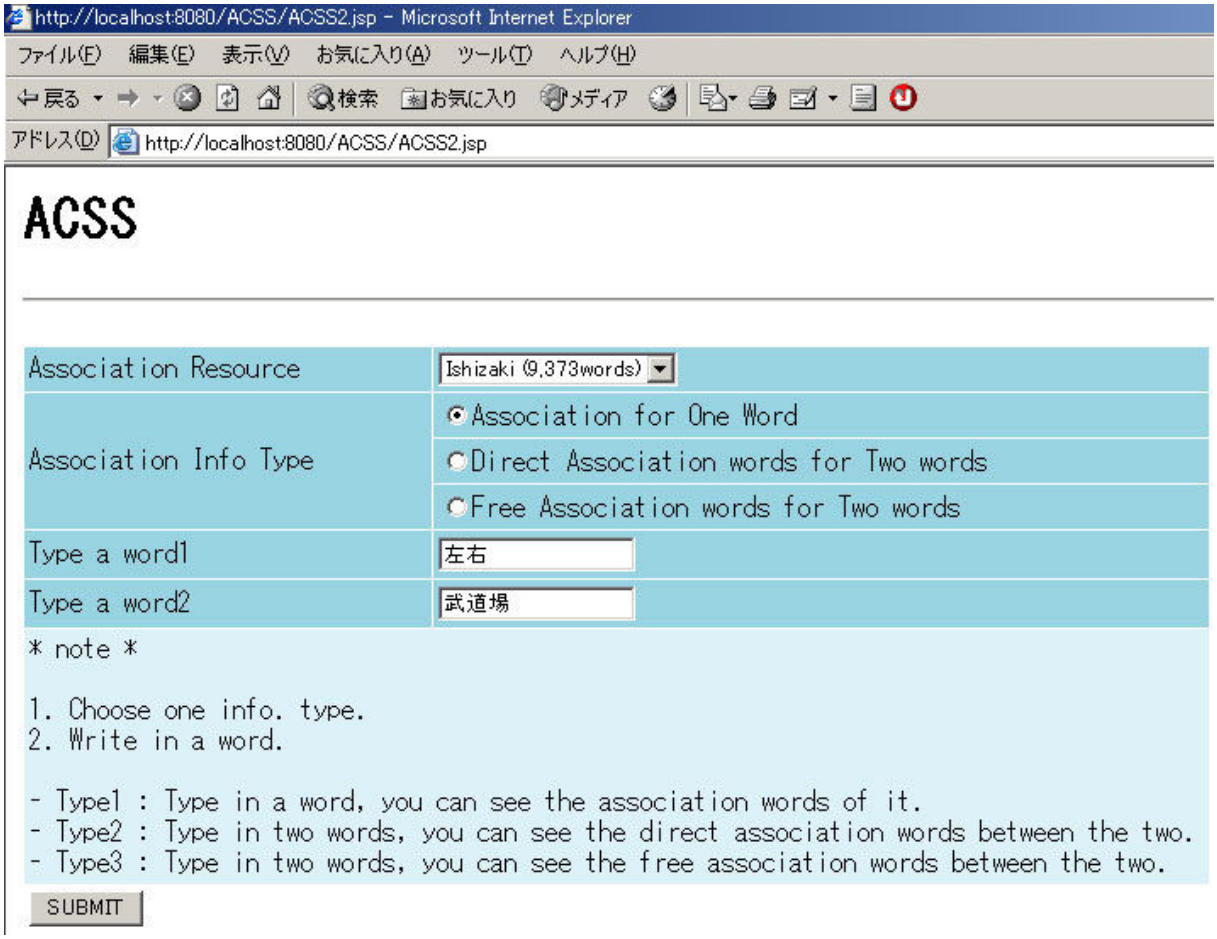

図2：ACSS（Associative Composition Support System）の GUI

6. まとめと今後の課題

我々は MCL を用いてその過程により生成する概念クラスターから、アリバイ崩しアルゴリズムによるマルコフ･クラスター最短パスをもとめ、言語全体の意味ネットワークやその詳細な小世界構造を作った。そして、この意味ネットワークに基づく作文支援システムを提案、そのベータ版である ACSS という連想作文支援システムをまず開発した。このシステムで得られる単語の情報は、単語の辞書的な表層の意味ばかりでなく、拡張的、連想的かつ伴示的な意味であり、また学習者に提示することで有益であると考えられる。

我々は今後、日本語ばかりではなく、英語などの多様なデータを使って、我々が開発したアルゴリズムを適用し、より豊かな連想リソースを構築したいと考えている。さらに学習者が直感的に理解できるよう、出力結果をグラフの形でビジュアル表示することも今後の課題である。このようにして、ACSS の作文支援システムを完成し、教育工学や認知科学の観点からそれを評価する予定である。

7. 謝辞

本研究は、21 世紀 COE プログラム（研究拠点形成補助金）「大規模知識資源の体系化と活用基盤構築」の言語・文献知識資源分野に関する研究の一環として行われたものです。また、データとして連想概念辞書の使用を許可してくださった石崎俊先生（慶応大学環境情報学部教授）に深く感謝致します。

【参考文献】

[1] 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

[2]杉浦 正利, コーパスに基づいた外国語作文支援システム, 上田博人編『日本語学と言語教育』東京大学出版会, 2002, pp.149-172

[3]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

[4]Norihisa Totsugi, Kikuko Nishina. Development of a System for Composition in Japanese by Utilizing the Dependency Structure Analyser--Focusing on Adjectives. 3rd International Conference on Computer Assisted Systems for Teaching & Learning/Japanese, 2002, pp.67-70.

[5]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

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

[8] 三宅真紀, 赤間啓之, 中川正宣, 馬越庸恭 単語の共起データに基づく共観福音書の特有性の分析, 情報処理学会, 人文科学とコンピュータ研究会, vol.78, 2004 pp.23-30