安田女子大学紀要　42，223–229　2014.

連想概念辞書とダイクストラ法を用いた単語連想と対話応答システムの構築

坂　口　琢　哉

Constructing a Dialogue System Based on Word Association by Using Associative Concept Dictionary and Dijkstra Algorithm

Takuya SAKAGUCHI

1．は　じ　め　に

近年，少子高齢化や核家族化に伴い，人間同士がコミュニケーションを取る機会が減少しつつあり，これを補う手段として，人間と対話可能なコンピュータの研究がさかんに進められている。一方，PC の小型化やスマートフォンの普及により，従来の GUI に替わるコンパクトな入出力手段として，音声でやりとりできるインターフェイスが注目を集め，「Siri[7]」に代表されるように実用的なソフトも登場しつつある。これらの背景から，コンピュータが人間の発話を正しく理解し，またそれに対して適切な単語や文を応答する対話システムの研究は，今後ますます重要になると考えられる。

実用的な対話システムには，音声認識，言語理解，対話制御，言語生成，音声合成といった様々な技術が用いられる。中でも言語理解から言語生成に至る過程は最も実用化が難しい分野であり，これを実現するためには，コンピュータが人間と同様に様々な知識を保持している必要がある。連想概念辞書[1] は，こうした分野に利用可能な大規模知識データのひとつであり，人間が持つ様々な知識を，連想実験と呼ばれる実験により単語単位で直接収集している点と，各単語間の距離が定量化されている点に特徴がある。連想概念辞書の応用例としては，これまでに比喩理解[4] や文書要約[5]，多義性解消[6] など，比較的高度な自然言語処理を扱ったものが報告されている。

筆者らはこれまでに，連想概念辞書を利用するためのダイナミクスとしてニューラルネットワークに着目し，これを用いて脳の記憶モデルを構築する研究を進めてきた。特に，ニューロン素子のモデルに通常使われる McCulloch-Pitts タイプではなく，時間発展をより忠実に記述できる Integrate-and-Fire タイプを採用することで，実際の脳に近い挙動が可能なモデルを目指した[2]。例えば同モデルでは，人間のプライミング効果を模した形で単語連想を行うことが出来るため，これを利用して人間に近い感覚で比喩を理解できるシステムを提案した[4]。しかし一方で，ダイナミクスが複雑であり，実践的な自然言語処理システムの実現に対して最適なパフォーマンスを維持するためのパラメータ調整が困難である問題があった。

本研究の目的は，実際の脳を意識したモデルではなく，より直線的かつ単純化された手法で連想概念辞書を利用することと，これを用いた簡単な応答システムを構築し，その有効性を検証することである。具体的には，連想概念辞書に含まれる単語とリンクのデータを一つの巨大な有向

グラフとみなし，グラフ理論の分野で一般的に知られるダイクストラ法により，任意の単語間の最短経路を計算する。その結果，従来より簡便な方法であっても，ある程度の連想が可能であることを示す。

2. 提 案 手 法

2.1 システムの概要

提案システムでは，ユーザが入力した自然文に対し，予め登録された単語の中から最も関連があるものを一つ選択し，出力する。例えばユーザがキーボードから「空を自由に飛びたいな」という文を入力すると，システムはまず，この文に含まれる「空」「自由」「飛び（たい）」などの単語を形態素解析により抽出する。尚，この形態素解析には「MeCab[3)]」を用いた。一方，システムには予め応答用の目的語として「タケコプター」「どこでもドア」「タイムマシン」などが登録されており，形態素解析によって得られた単語と，これらの目的語との距離をそれぞれ計算する。その際，各単語間の関係と距離の情報は，グラフ化された連想概念辞書のデータを利用し，またグラフ内における距離の計算にはダイクストラ法を用いる。こうして得られた距離を目的語ごとに合計し，その値が最も小さいものが，ユーザの入力文に対して総合的に最も関連が強い単語であるとみなし，システムによる応答文が出力される。

以上のようなシステムの概要を，図2.1に示す。

以降の節で，連想概念辞書のグラフ化および，ダイクストラ法による最短経路探索とこれに基づいた連想について解説する。

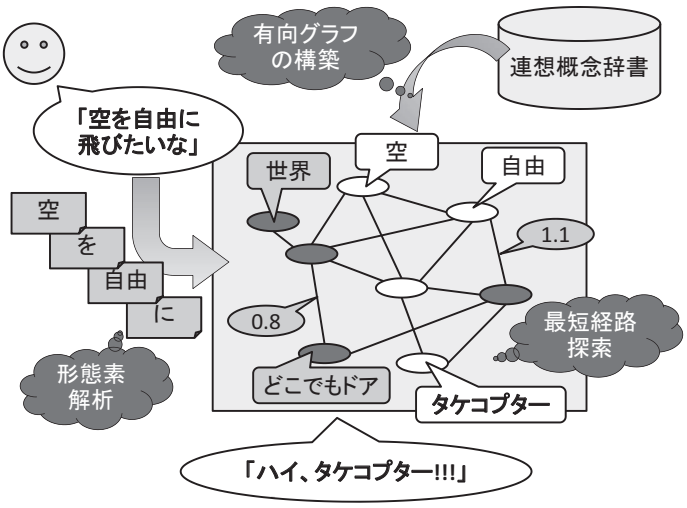

図2.1 システムの概要

2.2 連想概念辞書のグラフ化

連想概念辞書は，コンピュータの様々な自然言語処理に利用可能な大規模知識データの一種であり，その特徴は前述したように，知識の収集手法に連想実験を用いている点と，収集した単語

間の距離を定量化していることである。このうち前者については，被験者に連想の起点となる刺激語を提示し，その上で「上位概念」「下位概念」「動作概念」「環境概念」など計７種類の課題に従って自由に単語の連想を行ってもらう，というものである。一方後者については，刺激語 x から連想語 y への距離 $D(x, y)$ が，式（１）のように定式化されている[6)]。

$$D(x, y) + 0.81\, F(\mathrm{x}, y) + 0.27\, S(x, y) \qquad (1)$$

ただし，$F(x, y)$ は複数の被験者において，刺激語 x に対し連想語 y が回答された割合（連想頻度）であり，一方 $S(x, y)$ は x に対する y の回答順位の平均（連想順位）である。また各項の係数は，線形計画法により最適化された値である。

表2.1に，連想概念辞書に記述されたデータの一部を示す。

表2.1　連想概念辞書のデータ例

刺激語	課題	連想語	回答時間	回答順位	回答率	距離
明かり	環境概念	電気	0.2	1	0.1	***6.49***
明かり	環境概念	暗がり	0.183	1	0.1	***6.49***
明かり	環境概念	家	0.808	4.5	0.2	***4.565***
明かり	環境概念	学校	0.775	5.5	0.2	***4.895***
明かり	環境概念	暗闇	0.275	1.5	0.2	***3.575***
明かり	環境概念	部屋	0.244	1.167	0.6	***1.412***
秋	上位概念	シーズン	0.333	3	0.1	***7.15***
秋	上位概念	期間	1.567	5	0.1	***7.81***
秋	上位概念	時候	0.65	2	0.1	***6.82***
秋	上位概念	四季	0.258	1.5	0.2	***3.575***
秋	上位概念	時間	0.592	2.5	0.2	***3.905***
秋	上位概念	時期	0.594	3	0.3	***3.043***
秋	上位概念	季節	0.228	1.1	1	***0.979***
秋	下位概念	大きい秋	0.3	2	0.1	***6.82***

こうしたデータを展開し，各単語をノード，単語間の関係をノード間のリンクとした有向グラフをコンピュータ上に構築した。その際，同一の単語であれば，刺激語／連想語にかかわらず同一のノードとみなした。一方，リンクについては刺激語から連想語に向けた有向リンクとし，各リンクには単語間の距離の値を重みとして記述した。また，連想実験における課題の種類は実装せず，全ての単語を均等に扱うモデルとした。

本研究において，連想概念辞書は2003年度時点のものを使用した。更に，システムの応答用として表2.2に示したデータを追加してグラフを拡張し，これらのノードを，後述するダイクストラ法の終点として定義付けた。最終的に構築されたグラフのノード数は23,093個，総リンク数は71,717本であった。

2.3　ダイクストラ法による最短経路探索

ダイクストラ法は，重み付きグラフにおいて任意の２ノード間の最短経路を求めるアルゴリズムである。この方法では，開始地点となるノードを決めた後，そこからリンクされたノードのう

表2.2 システムの応答用データ

刺激語	連想語	距離
夢	もしもボックス	*1.0*
空	タケコプター	*1.0*
場所	どこでもドア	*1.0*
時間	タイムマシン	*1.0*
大きい	スモールライト	*1.0*

ち距離が最短であるものを選択し，そのノードまでの最短経路を確定させる。次に，最短経路が確定したノードに起点を移し，同様の探索を進めていく。こうしたプロセスを繰り返すことで，徐々に各ノードまでの最短経路が確定していき，最終的に目的地点となるノードまでの経路が確定したとき，探索は終了する。

提案手法では，第2.1節で示したような，入力文を形態素解析して得られる全ての単語を開始ノード，一方表2.2において示した各連想語を目的ノードと定め，両者間の距離をダイクストラ法により計算させる。その上で，全開始ノードからの距離の合計を目的ノードごとに集計し，この値が最小であったものを，ユーザの入力文全体に対する応答の最適解として出力させることとした。

尚，ダイクストラ法を適用する際は，グラフを構成する全てのリンクについて重みが 0 以上である事が条件となる。本研究で使用した連想概念辞書の距離は［1, 10］で定義されており[1)]，この条件を満たす。

2.4 動 作 式

式（2）および式（3）に，システムの動作式を示す。

$$c_y = \sum_{x \in I} \sum_{n_k \in L_{xy}} D\left(n_k, n_{k+1}\right) \tag{2}$$

$$o_y = \begin{cases} 1 & c_y = c_{min} \\ 0 & (otherwise) \end{cases} \tag{3}$$

上式において，I は入力文に含まれる全ての単語，すなわち開始ノードの集合を表し，一方 L_{xy} はノード x からノード y への最短経路に含まれる全てのノードの集合を表す。また c_y は，全ての開始ノードから目的ノード y までの距離の合計であり，更に o_y は y に対する最適解のフラグである。距離の合計値 c が最小値 c_{min} と等しいノードにのみ，フラグ o に 1 が代入され，入力文に対する応答として出力される。

3. 実 験 と 考 察

3.1 システムの構築

前章で示した手法やデータに従い，応答システムを構築した。実装は Java で行い，Windows のコマンドプロンプト上で動作するシステムとした。また，開発環境および実験環境としては，

OS が 64bit 版 Windows7，CPU が Core i7 1.90 GHz のノート PC を使用した。

図3.1は，構築したシステムへの入力と応答の様子を捉えたスクリーンショットである。「空を自由に飛びたいな」というユーザの入力に対し，形態素解析によって得られた単語から目的語までの距離を計算して最適解を求め，最終的に「ハイ，タケコプター !!!」と応答していることが分かる。

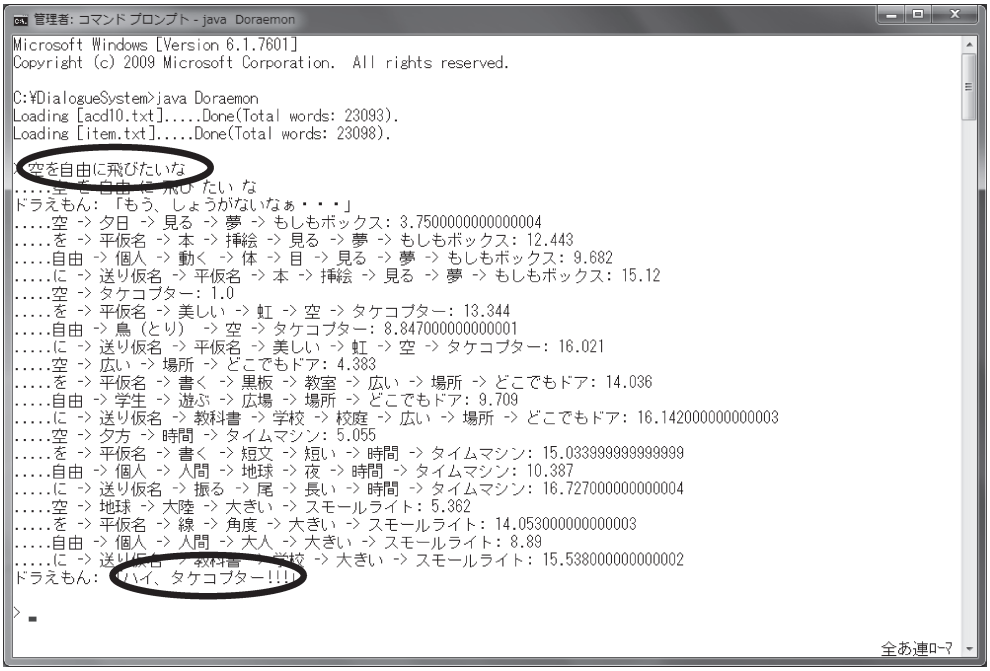

図3.1　システムの入出力

3.2　動 作 実 験

構築したシステムの動作を確認する目的で，簡単な実験を行った。具体的には，幾つかの入力文と期待される目的語のペアを用意し，実際の出力と比較した。表3.1に，これらの実験結果を示す。

多くの入力文において，期待される目的語と同一の出力が得られ，提案手法およびシステムの有効性が示唆された。一方うまく行かなかった例として，表3.1の入力文06を分析した結果，「学校」という単語から「大きい」という単語への距離が短く，このことが期待と異なる出力の原因と考えられた。また，うまくいった例の中でも，例えば入力文08では「ミクロ」という言葉より「世界」という単語が強く影響しており，人間の直観的な判断とは異なるケースが見られた。

第2.2節で述べたように，提案手法では刺激語と連想語の差異や，連想実験の課題による差異を考慮することなく，全ての単語に対して均一的に処理を行う。一方で，実験で使用した入力文にはユーザの希望や欲求を表したものが多く，このような文においては動詞の重要度がより高くなる可能性がある。また，前述した入力文08のように特徴的な単語が存在する場合は，その単語に着目することで精度の向上が期待できる。入力文における各単語の重要度を定式化する手法に

表3.1 システムの応答用データ

ID	入 力 文	期待される目的語	システムの応答
01	「空を自由に飛びたいな」	タケコプター	「ハイ，タケコプター !!!」
02	「鳥のように素早く移動したい」	タケコプター	「ハイ，タケコプター !!!」
03	「昔の自分に会いたいな」	タイムマシン	「ハイ，タイムマシン !!!」
04	「100年後の未来はどうなってるかな？」	タイムマシン	「ハイ，タイムマシン !!!」
05	「色々な場所を旅してみたいなあ」	どこでもドア	「ハイ，どこでもドア !!!」
06	「学校に忘れ物を取りに戻らなきゃ」	どこでもドア	「ハイ，スモールライト !!!」
07	「この荷物，重すぎて運ぶのが大変！」	スモールライト	「ハイ，スモールライト !!!」
08	「ミクロの世界を覗いてみよう」	スモールライト	「ハイ，スモールライト !!!」
09	「おやつを好きなだけ食べてみたいぞ！」	もしもボックス	「ハイ，もしもボックス !!!」
10	「どうかこの恋が実りますように…」	もしもボックス	「ハイ，もしもボックス !!!」

ついては，今後の検討課題である。

4. おわりに

4.1 まとめ

本研究では，連想概念辞書を利用するためのダイナミクスについて，従来用いていたニューラルネットワークよりも簡潔な手法としてダイクストラ法に着目した。すなわち，連想概念辞書のデータを展開して，各単語をノードとした有向グラフを構築し，任意のノード間の最短経路をダイクストラ法により効率的に求める。また提案手法を用いて，ユーザの入力文から適切な目的語を選択して応答する対話システムを構築した。システムの動作実験において幾つかの自然文を入力した結果，良好な出力が得られ，提案手法の有効性が示唆された。

4.2 今後の展望

システムに対する今後の検討課題としては，以下の要素が挙げられる。まず，第3.2節でも言及したように，入力文における単語の重要度を考慮することで，システムによる応答の精度向上が期待できる。また，現環境においてシステムの平均応答時間は約4秒程度であるが，より自然な対話の実現に向けてアルゴリズムを改善し，応答時間を短縮させることも検討したい。本研究で用いたダイクストラ法は，リンク数 m，ノード数 n のグラフに対し一般的に $O(n^2)$ の計算量がかかるが，優先度付き待ち行列を用いて最短経路の確定を効率的に行えば，計算量を $O(m+n\log n)$ 程度に抑えられる事が知られている。これらの仕組みをシステムに実装することで，よりスムーズな対話が可能になると考えられる。最後に，目的語を拡充し，また応答文のパターンも複数用意することで，より複雑な対話が可能なシステムへと昇華させたい。

参 考 文 献

1) 岡本 潤，石崎 俊，“概念間距離の定式化と電子化辞書との比較”，自然言語処理，Vol. 8，No. 4，pp. 37–54，2001.

2） T. Sakaguchi and S. Ishizaki, “A Japanese Semantic Network built on a Pulsed Neural Network with encoding Associative Concept Dictionaries”, The 19th International Conference on Computational Linguistics, Proceedings of the Workshop on SEMANET, Building and Using Semantic Networks, pp. 23–29, 2002.
3） 工藤　拓，山本　薫，松本裕治，“Conditional Random Fields を用いた日本語形態素解析”，情報処理学会研究報告「自然言語処理」，Vol. 2004，No. 47，pp. 89–96，2004.
4） 坂口琢哉，“連想概念辞書のニューラルネットワークへの符号化と比喩理解システムへの応用”，安田女子大学紀要，No. 38，pp. 169–179，2010.
5） J. Okamto and S. Ishizaki, “An Associative Concept Dictionary for Natural Language Processing: Text Summarization and Word Sense Disambiguation”, Journal of Cognitive Science, Vol. 12, pp. 259–276, 2011.
6） 岡本　潤，石崎　俊，“文脈ネットワークを用いた語の多義性解消”，Keio SFC journal，Vol. 12，No. 1，pp. 97–111，2012.
7） Siri, http://www.apple.com/jp/ios/siri/, 2013.

Summary

In this study, we have mentioned to dijkstra algorithm and suggested a new method to deal with data of associative concept dictionaries, which was more simple and reasonable way than neural network architecture adopted in some previous studies.　We have constructed a computational dialogue system based on a suggested method, containing a digraph structure of an associative concept dictionary with 23,093 nodes for words and 71,717 links for relations of words. When a user inputs a sentence to the system, it gets some start nodes with morphological analysis and starts searching in the digraph structure with dijkstra algorithm to find the shortest path to some goal nodes.　The system calculates the total distance from all start nodes for each goal node, considering one with the minimum distance as the most appropriate answer to output.　We evaluated the system with some example sentences to input, getting good results in most cases to imply the availability of our suggested method.

〔2013．9．26　受理〕