

言語障害者のための会話支援システム

— Wikipedia からの単語リスト作成 —

山根靖子[†] 服部文夫[‡] 奥健太[‡] 安田清^{‡‡}

[†] 立命館大学大学院 理工学研究科 〒525-8577 滋賀県草津市野路東 1-1-1

[‡] 立命館大学 情報理工学部 〒525-8577 滋賀県草津市野路東 1-1-1

^{‡‡} 千葉労災病院 リハビリテーション科 〒290-0003 千葉県市原市辰巳台東 2-16

E-mail: [†] cc013054@ed.ritsumeai.ac.jp,

[‡] fhattori@is.ritsumeai.ac.jp, oku@fc.ritsumeai.ac.jp, ^{‡‡} fwkk5911@mb.infoweb.ne.jp,

あらまし 昨今の高齢化に伴い認知症や失語症などの言語障害者が日本では急増している。言語障害者との会話の手助けとなるような“楽々自由会話[1]”というコンピュータの画面に単語を表示させるシステムが開発されているが、単語数が少ない事や、登録されている単語に偏りがあるという問題点が挙げられる。本研究では既存のシステムの問題点を改善し、より多くのデータを用いて多様な会話に対応できるように Wikipedia[2]に登録されている単語を抽出し、単語集を作成した。100万を超えるデータを Wikipedia カテゴリを用いて階層化することにより、自然な会話を実現できることを実験により確認した。

キーワード Symbiotic computing, 会話支援, Wikipedia

Conversation Support System for People with Language Disorders

— Making Topic lists from Wikipedia —

Yasuko YAMANE[†] Fumio HATTORI[‡] Kenta OKU[‡] and Kiyoshi YASUDA^{‡‡}

[†] Graduate School of Science and Engineering, Ritsumeikan University

[‡] Information Science & Engineering, Ritsumeikan University

1-1-1 Noji-higashi, Kusatsu-city, Shiga, 525-8577 Japan

^{‡‡} Department of Rehabilitation Medicines, Chiba Rousai Hospital

2-16 Tatsumidai-higashi, Ichihara-city, Thiba, 290-0003 Japan

E-mail: [†] cc013054@ed.ritsumeai.ac.jp,

[‡] fhattori@is.ritsumeai.ac.jp, oku@fc.ritsumeai.ac.jp, ^{‡‡} fwkk5911@mb.infoweb.ne.jp,

Abstract A conversation support system for people with language disorders is proposed. Although the existing conversation support system “Raku-raku Jiyu Kaiwa” (Easy Free Conversation) is effective, it has insufficient topic words and a rigid topic list structure. To solve these problems, this paper proposes a method that makes topic lists from Wikipedia’s millions of topic words. Experiments using the proposed topic list showed that subject utterances increased and the variety of spoken topics was expanded.

Keyword Symbiotic computing, Conversation support, Wikipedia

1. はじめに

現在日本ではほぼ5人に1人が65歳以上の高齢者であり、その中の13人に1人が認知症という割合

である。65歳以上の高齢者がいる世帯では6, 7世帯に1人であり、極めて身近な病気といってよい。また在宅での介護を余儀なくされる方も多く、認知症の度合は様々であるが、介護者である家族の負担も非常に大きい。認知症や失語症などの言語障害者との会話

の手助けとなるような“楽々自由会話[1]”という単語を表示させるシステムが開発されているが、単語数が少ないという問題がある。単語数が少ないと目的の単語に到達できずに会話の途中で話題が途切れてしまう原因となる。また“楽々自由会話”はすべて手作業で単語を入力しているため、入力者によって単語が偏ることや、重複するという問題もある。本研究では既存のシステムの問題点を改善し、より快適に利用していただけるようなシステムを目指して Wikipedia[2]に登録されている単語を抽出して階層構造を用いて単語を表示し検証実験を行った。また、問題点の改善策として抽出したデータベースを整理し、カテゴリ分類を簡素化することにより、多様な会話に適応し、より自然な会話を促すような単語リスト作成を目指した。

2. 楽々自由会話

2.1 システムの概要

楽々自由会話とは、失語症などの言葉の障害が原因で、会話が滑らかにできない人のためにデータ化された単語集をHTML形式で一覧表示したシステムである(図1)。単語集の中から意中の単語を探し、その単語に指を差すなどして意思を伝えるという方法で会話を進める。また、楽々自由会話に収録されている単語は、日常的な会話をするために必要な単語だけではなく、思った言葉が口に出せない言語障害を持つ患者が、何を伝えようとしているかを調査し、言語聴覚士の監修のもと、効率的に言葉を引き出すような単語を厳選している。

失語症などの言葉の障害をお持ちの方のための単語集
「楽々自由会話」

1. エンターテイメント	2. スポーツ
3. 趣味と旅行	4. 家庭と生活
5. 社会とニュース	6. 人文科学
7. 自然科学	8. 健康と医学
9. 千葉県・日本	10. コンピュータと資料

分類

図1 楽々自由会話 単語表示画面

2.2 楽々自由会話の課題

楽々自由会話の課題として、単語数の限界があげられる。会話の途中で話題に行き詰ってしまうことや、目的の単語が列挙されていない場合もある。また、カテゴリ分類が曖昧なため、話題の目的の単語がスムーズに探し出せないことや、そこにたどり着くまでの時

間がかかることも問題である。楽々自由会話はボランティアなどで単語の登録を行っているので、単語数やカテゴリにとっても偏りがある。これらの課題を克服し、患者さんと介護者の方とのコミュニケーションが、より円滑に行われるような手助けとして、Wikipediaベースの単語リストを提案する。

3. Wikipediaからの単語リスト作成

成

3.1 Wikipedia カテゴリによる階層化単語リストの生成

Wikipedia から単語を抽出するにあたり、階層構造を実現する Wikipedia カテゴリに着目した。Wikipedia に登録されている単語はカテゴリと記事に分類できる。カテゴリ自体は記事をもたず、そこに属する単語とサブカテゴリをもつ(図2)。そのカテゴリを下っていくことで、階層構造を実現している。

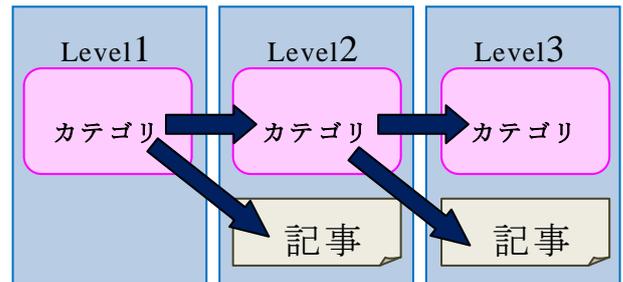


図2 Wikipediaの階層構造

本システムでは Wikipedia から抽出した単語とカテゴリと双方の相互関係を利用して、階層構造化された単語リストを作成する。データの抽出には Wikipedia データ解析ツール Wik-IE[3]を使用した。Wikipedia 上で一般に公開されているダンプデータを取得し、データベースに格納した(図3)。抽出したデータは Wikipedia に登録されているカテゴリを含むすべての単語約103万語(2009年6月1日現在)と、それらの単語全てにおける相互関係(IDを用いた親子関係データベース)の二つである。それぞれのテーブルを“Name”と“Relation”とする。この2つのテーブルを用いて各階層や単語間のリンクを使用して、階層構造を再現する。

Relation			Name		
children	parent	relation	NO	NAME	Relation
1076039	617718	target	10	言語	leaf
1498358	907547	target	100	インドネシア	leaf
1466960	346822	target	10000	1726年	leaf
1585759	346822	target	1000000	アムステルダム防衛線	leaf
927756	346822	target	1000001	キイウ	redirect
702415	NULL	target	1000003	ベームスター千托地	leaf
1507340	1141435	target	1000004	専ブラ	leaf
1142268	1141435	target	1000005	ヒグホオジロ	leaf
1167270	1330313	target	1000006	スワートルスクワート	leaf
907954	585193	target	1000007	ベストグループ	leaf
698717	585193	target	1000009	ジェーン	leaf
585194	585193	target	100001	東京都道427号	redirect
1455669	1455665	target	1000010	フエリチカ岩塩坑	leaf
603064	394161	target	1000014	パンダルガン国立公園	leaf
603062	394161	target	1000016	category:東京慈恵会医科大学	node
573362	394161	target	1000018	イモオジロ	leaf
1094494	394161	target	1000019	ほんわかテレビ	redirect
661522	394161	target	1000022	東京都道405号	redirect
434857	242283	target	1000020	六分橋立水産高等学校	redirect
			1000021	ラブレラの岩屋教会群	leaf

図 3 Wikipedia データベース

ゴリとして検索するか、もしくは単語として検索するかを選択すると、入力した文字を含むカテゴリや単語が表示される。検索結果画面を図 5 に示す。正規表現を用いる事も可能で前後一致や部分一致などを指定することができる。



図 5 検索結果画面

3.2 システムの実装

本システムでは単語リスト表示、画像表示、検索機能の3つの機能の実現を目指す。システム作成には、Adobe が開発している Flex builder3[4]を使用した。

① 単語リスト表示

単語をカテゴリと記事に分類し、各階層ごとにカテゴリと単語を分けて表示させる。階層構造の最上層として、最初に表示させる TOP ページのカテゴリは Wikipedia の主要カテゴリ 9 つ（総記、学問、技術、自然、社会、地理、人減、文化、歴史）を使用する。各カテゴリをクリックすると次の階層のカテゴリと単語が表示される。

② 画像表示

各単語をクリックすると Wikipedia に登録されているその単語（記事）の画像が表示される。その記事が画像を持たない場合は”No Image”と表示する。単語と画像が表示される例を図 4 に示す。また選択した単語の実際の Wikipedia の WEB ページも表示することができる。単語の意味が分からない時や、より詳しい情報を調べたい時に表示させる。



図 4 表示画面

③ 検索機能

目的のカテゴリや単語を直接表示させるために検索機能を作成した。検索したい単語を入力して、カテ

4. 検証実験 I

作成したシステムの有効性を確認するため、既存の単語リスト“楽々自由会話”との比較実験を行った。被験者は千葉労災病院に通院する認知症患者 3 人である。

4.1 実験方法

事前に各被験者と会話をする話題をヒアリングの上、3つ用意しておく。それぞれの話題について、最初に楽々自由会話で会話を行い、次に今回作成した Wikipedia からの単語リストを用いて会話をした。各話題の会話時間は3分間で終了とし、各話題において被験者がモニターを見て発言した単語数を計測した。

4.2 実験結果

実験結果を表 1 に示す。

まず、平均値から見て、いずれの被験者も Wikipedia からの単語リストを用いた時のほうがひとつの話題に対する発言単語数が多くなっている。またそれぞれの被験者がこの実験中に一番多く発言した話題も全て Wikipedia からの単語リストを用いた時であった(表中黄色網掛け)。

表 1 実験結果

被験者	話題	楽々自由会話		Wikipedia	
		回数	Avg.	回数	Avg.
S	A	11	Avg. = 7.0	15	Avg. = 13.3
	B	7		18	
	C	3		7	
F	D	2	Avg. = 5.7	4	Avg. = 10.3
	E	5		17	
	F	10		10	
I	G	15	Avg. = 12.0	22	Avg. = 18.3
	H	14		19	
	I	7		14	

4.3 考察と問題点

上記の結果より、被験者全員に対して、Wikipedia 単語リストを用いた会話における発言単語数が増加している事がわかる。いずれの被験者も、実験全体の9分×2の会話の中で、一番多くの単語を発言したのはWikipedia 単語リストである（表中の黄色網掛け）。

今回の検証実験で、発言単語数が全員増加したことについて、これは一度に表示される単語数が多いこと、またあらゆる分野を網羅する単語が表示されたことや、画像などを用いて会話の促進に役立てたことが大きな要因であると考え（表 2）。画像を表示することは、新たな話題へと会話が広がり、単語の意味を理解するという目的だけでなく、会話自体を盛り上げる役割をされると考えられる。そのほかにも、表 2 に示したように一画面に表示する単語数が平均して約 2 倍になったことで、画面を移動することなく、同じ画面で話題を保てることが要因であると考え

表 2 楽々自由会話との比較

	楽々自由会話	Wikipedia 単語
総合単語数	1 万語	1 0 2 万語
一画面の平均単語数	1 0	2 0
階層構造の最大数	4	9

しかし、今回の実験を通して、本システムを実際に使用する際の問題点として、階層構造の見直しが必要である事が明らかになった。データベースの構築にあたって、Wikipedia の項目となる単語間のマッチングを用いて親子関係のデータベースを作成し、これを用いて動的に単語リストを生成する手法を用いたため、次のカテゴリにいくまでにとっても時間がかかってしまった。そのため、本来被験者に見ていただきたい画面が現れる前に勝手に会話が進んでしまう事や、待っている間に被験者の集中力が切れてしまい、違う話題にそれてしまう事もあった。また、単純に字面だけのマッチングを行ったため、同義語など不必要な単語まで表示され、必要以上に 1 画面に表示される単語数の増加を招いてしまっていた。

5. カテゴリ構造の見直し

上記の問題を解決するために、カテゴリ構造の見直しを提案する。Wikipedia 上のすべてのデータを網羅して単語数増加を実現することは変わらず前提に置き、それを踏まえて、TOP である主要カテゴリ以降のカテゴリの連結を改良した。

これまでは id だけで表記した親子関係のデータベ

ースを用いて、単語のテーブルと強制的に連結させていたため（図 6）、SQL 文を実行する際にデータベースに存在するすべての単語を検索していた。これは全体的な処理速度の遅延につながっている。他にも、”Homepage”と”ホームページ”などの同一の言葉であっても Wikipedia 上では別々の id を持つため異なる単語としてみなされる。この場合、同じ意味の単語であるから別々に表示する必要はない。

そこで、Wikipedia 上の単語間の相互のカテゴリリンクを利用することによって階層構造を構築したデータベースを用いた（図 7）。このデータベースを使用することにより、従来よりも約 30 万語の不要な単語を削除することができた（表 3）。なお、使用するデータベースは東京大学の中山助教に頂いた。

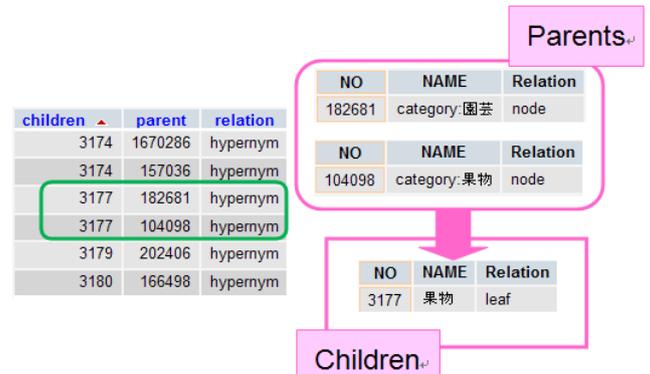


図 6 従来の階層構造の再現

cl_from	cl_to	cl_to_id
67	自然言語処理	201526
68	言語の分類	420989
68	言語学	103679
69	プログラミング言語	104919
69	コンピュータ言語	186553
69	プログラミング	142738
70	人工知能	185049
70	心の哲学	513792
70	コンピュータのユーザインタフェース	972111
70	SF	138222
74	オーストリア	139067
74	内陸国	424105

図 7 カテゴリネットワークデータベース

表 3 階層構造の変化

	既存の DB	改良した DB
総リンク数	2212544	1924820
総単語数	約 102 万語	約 102 万語
一画面の平均単語数	2 0	1 7
階層構造の最大数	9	7

6. Wikipedia カテゴリネットワーク

を用いた単語リスト作成

6.1 カテゴリネットワークの利用

このデータベースを用いて、単語リストを作成する。今回は画像や検索機能を排除した単語リストを作成した。PHPを用いて、階層構造のTOPをWikipediaの主要カテゴリの9個とし、それ以降の階層において、カテゴリネットワークを用いてWikipediaカテゴリを再現する。

このカテゴリネットワークは、ボトムアップ形式のリンクの連結であるため、一つのテーブルでは単語まで一緒に抽出できない。今回作成した階層構造（トップダウン形式）に対応するためには、Wikipediaのページ情報だけを抜き出したPagesテーブル（図8）とも連結される必要があった。必要なidを抽出したのちに、テーブルPagesとリンクさせることで、単語表示を実現した。

page_id	page_namespace	page_title	page_restrictions
64	0	形式言語	[BLOB-0バイト]
65	0	文脈自由言語	[BLOB-0バイト]
66	0	正規言語	[BLOB-0バイト]
67	0	自然言語処理	[BLOB-0バイト]
68	0	自然言語	[BLOB-0バイト]
69	0	プログラミング言語	[BLOB-0バイト]
70	0	人工知能	[BLOB-0バイト]
74	0	オーストリア	[BLOB-0バイト]

図8 pages テーブル

6.2 システム概要

作成したシステムの単語表示画面を以下に示す（図9）。主要カテゴリをTOPとし、Wikipediaカテゴリネットワークに沿って階層を下っていく。最終階層にたどり着いた場合は、直接Wikipediaにリンクを貼っている。

このように、ひとつの画面に単語のみを表示させ、平均してやく20個の単語が一度に表示されるように調整し、できる限り大きな文字を使用することで、言語障害者の方の視覚的困惑を防ぐ。



図9 カテゴリネットワークを用いた単語リスト

7. 検証実験II

改良を施したカテゴリネットワーク単語リストを用いて千葉労災病院に通院する軽度から中度の認知症患者5名と検証実験を行う。

7.1 実験方法

4章の検証実験Iと同様に、事前に各被験者と会話をする話題をヒアリングの上、3つ用意しておく。それぞれの話題について、Wikipedia単語リストとカテゴリネットワーク単語リストを順不同に用いて会話をした。各話題の会話時間は3分間で終了とし、各話題において被験者がモニターを見て発言した平均単語数と平均発言時間、ならびにシステム待ち時間（画面を移動する際の会話が停止する時間）を計測した。

7.2 実験結果

実験結果を表4に示す。

まず、平均発言数、平均発言時間ともに、カテゴリネットワークを用いた会話実験の方が、より多く被験者が発言している。発言単語数は2倍以上に増加し、発言時間に関しては全体として約1.5倍の長さに伸びている。

また、システム待ち時間に関しては、カテゴリネットワークを用いた会話実験においては約2秒と、日常生活におけるインターネット使用時の待ち時間と変わらないまでに短縮することができた。逆に、Wikipedia単語リストにおいては平均して20秒のシステム待ち時間があるので、被験者の集中力がきれてしまうことや、話題がそれてしまう原因となることが改めて確認できた。

単語リストの方が、言語障害者との会話に適している。

本システムでは、単語数の充実、カテゴリ分類の一般化、会話に適切な単語表示に関して最適なシステムであると考えられる。

今後の展望としては、最下位層の単語の扱いについてより工夫が必要であると考えられる。今回は Wikipedia へのリンクを貼り付け、必要な場合に Wikipedia のページを用いて会話を行ったが、これは介護者側に対するアプローチであり、実際に言語障害者が見る画面としては適切ではない。Wikipedia のページから、その単語の概要と画像、また関連する単語のみを抽出するなど、必要最低限の情報を抽出し、表示させることで、介護者と同じ画面、情報を共有できる。これは会話のさらなる発展要素になりうる、と考える。

9. おわりに

本論文では言語障害者の方との会話の手助けとなるようなシステムとして Wikipedia ベースの単語リストを提案した。画像表示や検索機能を付加することにより、会話の幅が広がり発言する数が増加することが実験より検証された。同時に階層構造の構築手法によって動作に必要以上に時間を要すること、カテゴリ分類が複雑であること、表示される単語数が多すぎる事を改善すべく、不要な単語を除き、単語同士の連結方法を変えたことで、階層構造の改善として Wikipedia のカテゴリネットワークを使用してカテゴリ構造のみなしを行い、実際に作成した単語リストを持って被験者実験を行い、よりスムーズな会話を実現できるようにシステムを改良した。

謝辞

処理速度の改善のために、東京大学知の構造化センターの中山助教より Wikipedia データベースの提供を受けた。また、実験に協力していただいた千葉労災病院リハビリテーション科の被験者の方々に感謝する。

参考文献

- [1] 言語障害者のための単語集 「楽々自由会話」, <http://rousai.kuronowish.com/>
- [2] フリー百科事典 「Wikipedia」 <http://ja.wikipedia.org/>
- [3] Wikipedia データ解析ツール Wik-IE <http://wik-ie.sourceforge.jp/>
- [4] Flex builder 配布ページ <http://www.adobe.com/jp/products/flex/>
- [5] 中山浩太郎, 伊藤雅弘, Erdmann Maïke, 白川真澄, 道下智之, 原隆浩, 西尾章治郎: Wikipedia マイニング: Wikipedia 研究のサーベイ, 情報処理学会論文誌 Vol.2 No.4 49-60(Dec.2009)

表 4 実験結果

被験者	評価項目	Wikipedia	カテゴリネットワーク
K	平均発言数	2 個	4 個
	平均発言時間	60.3 秒	101.3 秒
	システム待ち時間	22.6 秒	1.3 秒
U	平均発言数	1.6 個	4.8 個
	平均発言時間	53.3 秒	111 秒
	システム待ち時間	22.3 秒	2.3 秒
Y	平均発言数		3.3 個
	平均発言時間	※測定不可能	108.3 秒
	システム待ち時間		2.6 秒
H	平均発言数	2 個	5 個
	平均発言時間	62 秒	109.3 秒
	システム待ち時間	36 秒	2 秒
H	平均発言数	3.3 個	4.6 個
	平均発言時間	63.6 秒	83.6 秒
	システム待ち時間	24.3 秒	4.3 秒

(※被験者の事情により途中で中止した)

7.3 考察

今回の検証実験で、前回のシステムから被験者の発言率が大きく伸びた要因として、表示画面の簡素化、システム待ち時間の短縮、カテゴリの連結方法を改善したことにより、選択した単語の下位層が以前よりもシンプルな表示結果であることが挙げられる。実際に会話をしていて、システム待ち時間が長いと会話のテンポが停滞するのみならず、今何について会話していたのか分からなくなってしまうケースが多くみられた。被験者が話すタイミングに合わせて単語を表示することで、会話の流れを止めることなく、本人が心地よく話続けられる要素を提供できたと考えられる。

8. 考察と展望

本研究における二つの検証実験を経て、カテゴリネットワークを用いた単語リストが被験者との会話において、より発言率を高める効果があると検証された。

Wikipedia 単語リストで実現した単語数の充実ならびに画像表示や検索機能などを付加した、情報量の多い単語リストを用いた場合と比べて、カテゴリネットワークを使用した単語リストにおける、単語数の充実を実現しつつ、よりシンプルな画面表示、ならびに処理速度を改善しシステム待ち時間を大幅に減少させた