Web を利用した連想単語及びモダリティ表現による雑談システム

樋口 真介　ジェプカ ラファウ　荒木 健治
Shinsuke Higuchi　Rafal Rzepka　Kenji Araki
北海道大学大学院情報科学研究科
Graduate School of Information Science and Technology, Hokkaido University

1　はじめに

現在，自然言語処理の分野において，タスク指向の対話システムは数多く存在するが [1][2]，雑談システムのような非タスク指向の対話システムは，あまり盛んに研究されていない．その理由は，システムが特定の知識を用いて，ユーザの発話の内容や話題を予測することが困難であることなど，解決しなければならない難題が数多く存在するからである．しかし，雑談システムのような非タスク指向対話システムは，タスク指向対話システムなどと併用することで，ユーザの満足度を向上させると考えられる．

現存する非タスク指向対話システムとして古典的な対話システム ELIZA[3] や最近では A.L.I.C.E[4] といったものが存在するが，何れも人手によりルールを多数記述する必要がある．ELIZA では，あらゆる入力に対応するが，新たな情報を提供するような応答はなく，情報を要求するのみである．また，A.L.I.C.E. も知識源が限定されており，データベースの作成に多大な労力を要し，AIML といったようなマークアップ言語を習得する必要がある．このように，対話システムを構築するには，人手によるルール記述が最も現実的ではあるが，そのための多大な労力を考慮すると，精度向上には限界があると考えられる．

今現在，Web の発達に伴い，膨大なデータに容易にアクセスすることが可能になっている．Web 上のデータは常に更新されるため，時事的な情報を扱うには適した情報源であると言える．したがって，雑談といった非タスク指向対話のために，ユーザの発話から連想単語を抽出する際などに，Web は適していると考えられる．そこで本稿では，ユーザの発話に対する連想単語を Web から抽出し，その連想単語を用いて命題を生成し，さらにモダリティ表現を付与することで自動的に応答文を生成するという対話システムを提案する．本システムは，人間の発話というのは，命題とモダリティの要素によって構成されるという考えに基づいている．本稿では Web から抽出した連想単語の評価を行い，システムのアルゴリズムとモダリティ表現の有用性についての評価実験を行った結果について述べる．

2　発話に対する連想単語の抽出

本章では、ユーザの発話に対する連想単語を Web を用いて自動抽出することを提案する．本処理では，Google 検索 [5] を利用することで，データベース作成などの事前処理を行わずに，リアルタイムで連想単語を抽出する．

2.1　Web による連想単語の抽出

ユーザの発話から連想単語を抽出するために，まず始めにユーザの発話を MeCab により形態素解析する [6]．形態素解析により得られた名詞，動詞，形容詞，未知語を検索キーワードと定義する．これらを検索キーワードとして選択したのは，発話における自立語の中で，特に重要であると考えられるからである．形態素解析した結果, 名詞が連続している場合は, まとめて 1 つの名詞と定義している．例えば,「自然言語処理」を形態素解析すると,「自然」「言語」「処理」という 3 つの名詞に分解されるが，この場合は「自然言語処理」を 1 つの名詞として扱う．なお，以降の形態素解析方法はすべて同様の手法を用いる．

次に，得られた検索キーワードをクエリとして、Google 検索を行う．得られた検索結果のスニペット中に含まれる名詞を出現頻度順でソートし，単語リストを生成する．これは，共起頻度が高い単語ほど，入力単語との関連性が高いという考えに基づいている．使用するスニペットの個数は 500 とした．本数値は処理時間と出力結果を考慮し，実験的に決定した．得られた単語リストの上位にあるものを連想単語と定義する．表 1 が，発話から自動抽出した連想単語の例である．

表 1: 発話から抽出した連想単語の例

入力：「札幌は寒い。」

出現頻度順位	連想単語	出現頻度 (回)
1	雪	52
2	冬	50
3	気温	16
4	時期	12
5	東京	12
6	天気	11
7	地域	10
8	部屋	10

2.2　連想単語の評価実験

システムが抽出した単語リストが，連想単語としてふさわしいかどうかを被験者により，評価する．

被験者に自由に発話を入力してもらい，その発話に対してシステムが出力した連想単語１０個を被験者に提示する．各単語が，入力した発話に対する連想単語としてふさわしいかどうかを被験者が３段階で評価をする．被験者は，連想単語として確かにふさわしいと感じた場合３と評価し, 連想単語と扱っても良いと感じた時には２と評価し，連想単語としてふさわしくないと感じた場合に１と評価する．ここでは，２，３と評価された単語が連想単語としてふさわしいとし，有効出力とする．この評価を３人の被験者が１０回ずつ行った．１回の発話につき，１０個の単語を評価するので，計３００個の単語の評価を行った．連想単語を共起頻度でソートし，上位１０単語を対象とした時の結果が表 2 であり，上位５単語を対象とした時の結果が表 3 である．なお，評価の際，連想単語の定義はユーザに委ねている．これは，雑談における連想単語というのは，厳密に定義されているわけではないからである．

表 2: 上位 10 位まで連想単語

	被験者（A，B，C）	全体
評価 3 の数	40，52，57	149
評価 2 の数	37，17，27	81
評価 1 の数	23，31，16	70
有効出力率 (%)	77，69，84	77

表 3: 上位 5 位までの連想単語

	被験者（A，B，C）	全体
評価 3 の数	20，29，36	85
評価 2 の数	17，9，10	36
評価 1 の数	13，12，4	29
有効出力率 (%)	74，76，92	81

単語リストの上位 10 単語を対象とした表 2 の結果をみると，連想単語としてふさわしいと単語は，全体の約 77 %を占める．個別に見ると，評価の平均値はユーザにより若干の個人差はあるが，これは連想単語の定義がユーザごとに微妙に異なるからであると考えられる．また，単語リストの上位 5 単語を対象とした表 3 の結果をみると，連想単語としてふさわしい単語は，全体の約 80 %を占めている．これは，単語リストの上位にあるものほど，連想単語としてふさわしい可能性が高いということを示している．これらの結果から，検索エンジンを用いた連想単語の自動抽出手法は，高い精度を持つといえる．Web から連想単語を自動抽出するメリットは，時事的な単語，固有名詞など専門的な単語に対応できる点である．また，検索エンジンのスニペットを利用しているため，数秒で連想単語を得ることが可能である．本章では，連想単語として名詞のみを扱ったが，名詞よりも抽象的な概念である動詞，形容詞も同様の手法により連想単語として扱うことができると考えられる．

次章では，連想単語の自動抽出を利用した発話に対する応答文の生成方法について述べる．

3　システムの概要

本システムにおける応答文生成過程を，以下に示す．

1. ユーザの発話からキーワード獲得
2. Web から連想単語を獲得
3. 連想単語を用いた命題の生成
4. 命題へのモダリティ付与

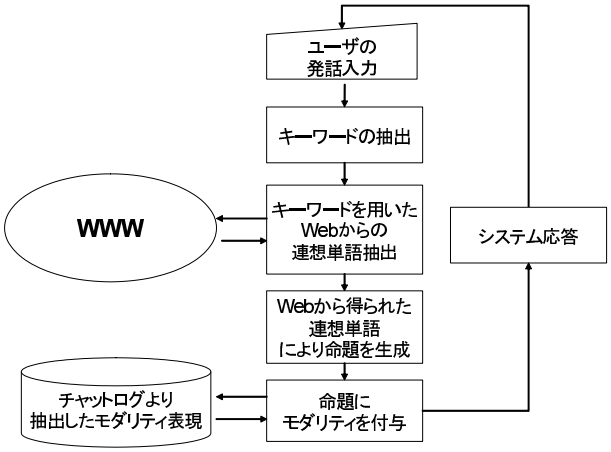

図 1: 応答文生成アルゴリズム

3.1　ユーザの発話からキーワード獲得

2.1 と同様に，最初にユーザの発話を形態素解析し，検索キーワードを抽出する．

3.2　Web による連想単語抽出

次に，得られた検索キーワードをクエリとして、Google による検索を行う．**2.1** と同様に，検索結果のスニペット中に含まれる単語の出現回数でソートする．**2.1** においては，対象単語を名詞に限定していたが，ここでは，名詞だけではなく，動詞，形容詞も扱う．各品詞ごとに，出現頻度でソートし，連想単語リストを生成し，最も出現頻度が高かった名詞，動詞，形容詞を連想単語として用いる．

3.3　連想単語を用いた命題の生成

得られた連想単語を利用して，命題を生成する．ここで，命題とは客観的な事柄を表した表現のことであり，[名詞 は 形容詞] のような命題テンプレートに単語を当てはめることにより生成する．ここでは，人手により８種類の命題テンプレートを用意した (表 4)．この命題テンプレートは，名詞，動詞，形容詞による基本的な命題が生成されるように，第一著者の主観により作成した．命題テンプレートは，

上から順番に適用する．生成される命題は，必ずしも自然な形になるとは限らない．そこで，Google のフレーズ検索を用いて，出現頻度が低く不自然と考えられる命題を淘汰する．これは，Web 上に数多く存在している命題は信頼できる，という考えに基づくものである．命題が不自然と判断された場合は，命題テンプレートを変えて，再び命題を生成する．今回は，実験的にフレーズ検索のヒット数が 1,000 件を超えた命題を採用した．

表 4: 命題テンプレート

名詞 は 形容詞
名詞 が 形容詞
名詞 が 動詞
名詞 は 動詞
それ は 形容詞
名詞
形容詞
動詞

3.4 命題へのモダリティ付与

最後に，このようにして得られた命題に対して，モダリティ表現を付与する．モダリティとは表現者の主観的な判断・態度を表す表現であり，副詞や文末に強く表れやすい [8]．本システムでは，文頭副詞と文末表現のペアをモダリティとして定義した．

3.4.1 モダリティ表現の抽出

モダリティの分類方法は，様々であるが，ここでは，雑談におけるモダリティとして疑問表現と伝達表現の 2 つに分類した．疑問表現は相手に情報伝達を求める表現であり，伝達表現とは相手に情報を伝達するための表現である．これらのモダリティを，予め IRC(Internet Relay Chat)[7] のチャットログ（10 万発話）から自動的に抽出した．そのモダリティ抽出方法は，以下の通りである．

- 文末に現れる助詞と助動詞の組み合わせを文末表現とする．
- 文末に『?』が付与された文末表現を疑問表現とする．
- 文頭に現れる副詞，感動詞，接続語と文末表現の組み合わせを伝達表現とする．
- 得られた候補を出現頻度でソートする．

得られた伝達表現は，686 パターンであり，これらのモダリティを第一著者が評価を行った．ふさわしい表現と評価されたのは，550 パターンであり，約 80%が適切と考えられる出力であった．一方，疑問表現は 396 パターンが得られた．これらのうち 292 パターンがふさわしい表現と評価され，約 73%が適切な出力であった．得られた候補を出現頻度でソートしてみたところ，上位の表現は，適切であることが多かったが，1 度しか出現しなかった表現も適切な表現と判定されることが多かった．例えば，疑問表現「〜じゃなかったでしたっけ？」は，正しい表現と考えられるが，10 万発話の対話ログの中には 1 度しか出現していない．このことからも，チャットログには多種多様のモダリティ表現が存在するが，出現頻度の少ないモダリティ表現が，不適切なモダリティ表現であるとは限らないと考えられる．モダリティ表現の例を表 5，表 6 に示す．

表 5: 得られた伝達表現の例

伝達表現	出現頻度
まぁ〜けど	21
まぁ〜だな	16
まぁ〜ですが	16
そこで〜ですよ	15
まぁ〜だが	14
まぁ〜ですよ	12

表 6: 得られた疑問表現の例

疑問表現	出現頻度
〜ですか？	232
〜かな？	90
〜だっけ？	87
〜ますか？	69
〜なの？	68
〜とか？	55

3.4.2 モダリティ付与

3.4.1 で得られたモダリティ表現を 3.3 で生成された命題に付与することで，システム応答文を生成する．これは人間の発話は，命題とモダリティ表現の要素から成立しているという考えに基づいている．[8] モダリティ表現は，表現の中からランダムに選択する．例えば，命題 [冬は寒い] が生成され，モダリティ [いや〜：ですよ] が選択された場合，「いや〜，冬は寒いですよ.」という発話が生成され，これをシステムの応答文とする．ただし，命題とモダリティ表現の組み合わせによっては，「冬は寒いだよねぇ。」のような不自然な文末表現になる場合がある．それを防ぐために，Google 検索を利用して，自然な文末表現が生成されるようにする．具体的には，文末表現をフレーズ検索し，閾値を上回った場合の発話を応答文として採用する．閾値を下回った場合の発話は，再びモダリティ表現を選択し直し，応答文を生成する．今回は，実験的に閾値を 100 件と定めた．

表 7: 実験結果

	システム α（命題）						システム β（命題＋モダリティ）					
評価項目	A	B	C	D	E	F	A	B	C	D	E	F
被験者 a	1	3	2	2	4	2	4	4	3	4	3	5
被験者 b	1	3	1	2	1	1	4	4	4	5	4	3
被験者 c	1	2	1	2	1	1	1	2	1	2	1	1
被験者 d	1	3	1	3	1	2	4	3	1	3	3	4
被験者 e	1	4	1	1	2	1	3	2	2	4	5	4
平均	1.0	3.0	1.2	2.0	1.8	1.4	3.2	3.0	2.2	3.6	3.2	3.4

4 実験結果及び考察

ユーザの入力に対して，自動生成した命題のみを出力するシステム α と，命題にモダリティを付与して出力するシステム β を用意した．5 人の被験者が各対話システムと 10 ターン以上の対話を行った後，アンケートによる 5 段階評価を行った．アンケートの評価項目は，対話を続けたいかどうか (A)，対話が文法的に自然であるか (B)，対話が意味的に自然であるか (C)，システムの語彙が豊富かどうか (D)，システムが知識を持っているように感じるかどうか (E)，システムが人間らしいか (F) である．各スコアの平均をとったのが表 7 である．モダリティ表現を用いたシステムは，モダリティを用いないシステムよりも高い評価を得られた．実際の対話例が表 8 である．実際に対話を行った被験者からは，「α α は β β。」のような表現は機械的であるが，「まぁ」，「もう」などの表現が加わることで機械的な印象が和らいでいるように感じた，といった意見が得られた．このように，システムの応答に日常的な会話表現であるモダリティを加えることで，ユーザの印象は向上すると考えられる．

表 8: システムの対話例

ユーザ:	今日は少し肌寒いよね。夏なのに。
システム:	朝は肌寒いとか?
ユーザ:	朝から肌寒かったね。
システム:	ああ、今日は早いなあ。

5 まとめ

本稿では，雑談のような非タスク指向対話を実現するために，Web から連想単語を自動抽出し，実際に連想単語としてふさわしいかどうかを被験者が実験により評価を行った．さらに，得られた連想単語を利用することで，命題を生成し，予め対話ログから自動的に抽出したモダリティ表現を組み合わせることで，応答文を自動生成するという対話システムを提案し，モダリティ表現の有効性についての評価を行った．人間の対話から自動的に抽出したモダリティ表現が，ユーザ，システム間の対話において有効であったことから，自動的にルールを生成するような対話システムを構築することが十分に可能であると考えられる．ただ，雑談のような対話を行うためには，連想単語だけではなく，他にも様々な知識が求められる．今回は，連想単語を命題テンプレートに当てはめ，モダリティを付け加えることで発話生成を行ったが，より高度な対話を行うためには，意味や文脈を考慮する必要がある．今後は，ユーザの発話に含まれる単語に反応するだけでなく，挨拶に対して挨拶で応答する，疑問に対して答える，相手の発話に対して肯定する，などのような相手の発話のモダリティを認識し，それに適した応答が可能となるような対話システムを目標としたい．

参考文献

[1] 駒谷和範，上野晋一，河原達也，奥乃博，”音声対話システムにおける適応的な応答生成を行うためのユーザモデル，”電子情報通信学会論文誌，Vol. J87-D-. No.10 pp.1921-1928,2004

[2] 川島宏文，土屋誠司，黒岩眞吾，任福継，”実用会話システムにおける対話型案内コンテンツの構築，”情報処理学会研究報告. 自然言語処理研究会報告 Vol.2007, No.76(20070724) pp. 1-5,2007

[3] J.Weizenbaum,”ELIZA—A computer program for the study of natural language communication between man and machine,” Commun. ACM, vol.9, no.1,pp.36—45,1966.

[4] Wallace，R. The Anatomy of A. L. I. C. E. http://www.alicebot.org/anatomy.html

[5] Google, http://www.google.co.jp/

[6] MeCab: Yet Another Part-of-Speech and Morphological Analyzer, http://mecab.sourceforge.jp/

[7] Internet Relay Chat Protocol , http://www.irchelp.org/irchelp/rfc/rfc.html

[8] 仁田義雄・益岡隆志，日本語のモダリティ，くろしお出版，1989.