5Q-1

マルコフテーブルによる学習を用いた辞書型対話システム

西村 太佑 † 奥村 紀之 †
† 長野工業高等専門学校 電子情報工学科

1 はじめに

本研究では，既存の人工無能の方式の特徴を組み合わた人工無能の提案を行う．既存システムでの会話の中でユーザに唐突さや不自然さを感じさせたり，フラストレーションを与えてしまい，その結果ユーザとの会話が途切れてしまったり，拒絶されてしまう問題を改善することを目指す．

2 会話システム

本システムは，ユーザの入力に対する返信文を自動的に生成するものである．返信文生成にはマルコフ連鎖 [1] を用い，マルコフ連鎖の過程で次の単語を決定するために辞書データベースを利用する．辞書データベースは自動学習により逐次更新される．学習には辞書型と呼ばれるシステムと同じ形式の辞書データを学習データとして使用する．辞書データの例を表 1 に示す．

表 1: 辞書データの例

入力パターン	返信パターン
行ってきます	行ってらっしゃい
ただいま	お帰りなさい

図 1 に本システムの構成図を示す．

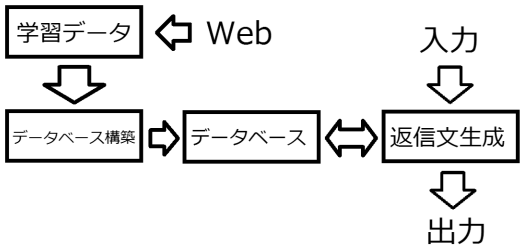

図 1: システム構成

3 自動学習システム

ユーザからの入力に対する返信文を生成するための語彙の学習を行うための辞書データは数多く集める必要があるが，手作業で集めるには非常に手間がかかってしまう．現代では Web を用いて様々な情報を閲覧することができるため，学習データも Web から収集することができないかと考え，以下の二通りの方法によるデータの収集を行った．

3.1 Twitter を利用した学習データの収集

既存の人工無能には IRC ネットワークを用いて学習を行うものもあるが，IRC ではある発言がどの発言に対する返信文であるのかを特定することができない．しかし,「Twitter[2]」ではある発言がどの発言に対する返信であるか特定が可能である．このことから，対となる文章が必要な学習データとして収集するのに優れていると考え，Twitter を用いることとした．

3.2 青空文庫を利用した学習データの収集

一般的に物語中では，発言部分は鍵括弧で括られている．また，連続して会話が続く場合は，改行を挟み連続して鍵括弧で括られた文章が並んでいる．このような特徴に注目し，青空文庫より学習データを収集する．本文中に鍵括弧で括られた文章が 2 行以上連続で登場した場合は，それぞれ入力パターン，返信パターンの対として抽出した．

3.3 データベースの構築

上記の手法を用いて収集した学習データは，そのまま辞書型システムの辞書として利用できるが，入力パターンが長くなってしまうため，検索時のヒット率が低くなってしまうおそれがある．そこで，辞書データを形態素解析し形態素単位の辞書データベースを作成した．辞書データ一行ごとに，入力パターンの全形態素に対し返信パターンの全形態素の組み合わせを登録することで辞書データベースを構築する．構築した辞書データベースの例を表 2 に示す．この際同時にマルコフ連鎖を行うために必要なテーブルの構築も行う．

表 2: データベースの例

入力単語	入力品詞	返信単語	返信品詞	出現回数
今日	名詞	日曜日	名詞	70
今日	名詞	月曜日	名詞	60
今日	名詞	火曜日	名詞	50
今日	名詞	水曜日	名詞	40
今日	名詞	木曜日	名詞	30
今日	名詞	金曜日	名詞	20
今日	名詞	土曜日	名詞	10
今日	名詞	です	助動詞	280

An Artificial Chatting Method based on Automatic Learning System using Web Resources
† Daisuke NISHIMURA
† Noriyuki OKUMURA (noriyuki_okumura@ei.nagano-nct.ac.jp)
Nagano National College of Technology (†)

4　返信文生成の手法

返信文の生成は，ユーザにより入力された文に含まれる単語全てに対し，辞書データベースから検索を行ない，返信文の候補となる単語を決定する．得られた候補語は二次のマルコフ連鎖により連結され，文章として出力される．

5　評価実験

5.1　実験データ

入力データとして，挨拶文，会話文、単語などを 36 通り用意した．

5.2　評価方法

作成したシステムに実験データを入力し，生成されたそれぞれの返信文に対し人手により評価を行う．評価の観点は以下の通りである．

- 返信文としてふさわしい単語が含まれているか
- 文法が正しいか

返信文としてふさわしいか否の判定は，返信文が入力文に対し関連を持つかを評価基準としている．

5.3　結果

作成したシステムに実験データを入力し得られた返信文に対し評価を行った．入力した文章と，生成された返信文の例を表 3 に示す．評価の結果を表 4 に示す．

表 3: 生成された返信文の例

入力した文章	生成された返信文
おはよう	おはようニャン
カレーを食べます	それ最近お気に入りか？なう．
こちょこちょ	きゃっ///
学習	技術ほんといみわからん

表 4: 出力結果の評価

評価の観点	良いと評価された割合
出力単語のふさわしさ	25/36
文法の正しさ	14/36

入力として,「おはよう」「おやすみ」などの簡単な挨拶を入力とした場合は，高確率で適切な挨拶の単語が含まれる返信文が得られた，10 文字程度の日常会話文を入力すると，入力文に関連があり，返信として悪くない単語は含まれるものの，余計な語が接続されていたり，文法が乱雑な文章が生成され，会話を行うに及ばない出力となった．

6　考察

評価実験より，入力された文章に対し，関連のある語が生成された返信文に含まれる場合の存在が確認できた．既存のマルコフ型人工無能では，文章の生成を乱数に頼って行うことが多く，入力文に関連する返信を的確に出力することが難しかった．本システムでは，返信文生成における単語の決定に学習データの出現頻度情報を利用しているため，学習量を増やすほど，入力文に関連の強い単語が返信文として出力される頻度が高くなると考えられる．今回の実験で人手により良いと判断された結果の多くは短い文章，あるいは 1 単語の出力であったが，本システムの手法は，返信文に含むべき単語の候補を選び出す際に有用であると推察される．文章の先頭に記号が含まれてしまうなど，マルコフテーブルや辞書データベース内のノイズの影響が結果に大きく現れた．この問題について，Twitter から収集した学習データには記号や顔文字などが多く含まれていることが原因と考えられる．また青空文庫から収集した学習データには古い日本語の表記やかな表記を多く含む作品からの抽出データが多く含まれていたため，形態素解析での誤りや，文章を生成する過程での誤った接続が行われてしまったと考えられる．

7　まとめ

本稿では，人間が入力した文章に対し，返信文を生成するための新しい手法として形態素単位での学習を行う手法を提案した．web 上でのユーザー同士の交流の履歴を用いて学習させ，統計的手法で返信文を作成することで，人間と計算機の間の交流の幅を広げる可能性を探りながら評価・実験を行った．今後は的外れな発言の繰り返しによるユーザのモチベーション低下を緩和することを目標とし，システム改善を行うと共に，自然な会話のシミュレートの実現を目指す．

謝 辞

本研究の一部は科研費（23720222）の助成を受けたものである．

参考文献

[1] 森部 敦, 毛利 公美, 森井 昌克：自動会話システム (人工無能) の開発とその応用 Web テキストからの会話文生成と会話形成に関する研究 (マルチメディアシステムの品質, 一般), 電子情報通信学会技術研究報告. OIS, オフィスインフォメーションシステム 105(283), 11-16, 2005-09-08

[2] Twitter ：http://twitter.com