1S-8

会話パターンの機械学習と知識ベースを用いた会話システムの研究

菊地啓夫† 田胡和哉‡

†東京工科大学コンピュータサイエンス専攻

1 はじめに

人間とコンピュータ間のチャット形式の会話システムは,雑談相手や情報を得るための簡易操作インターフェイスなど,多くの目的で利用されて来た.なかでも,言語の意味に立ち入ることなく,表面的に会話を成り立たせることを目的とした人工無能が広く利用されている.従来の人工無能の返答アルゴリズムを大別すると「マルコフ文生成型」,「ログ型」,「辞書型」が挙げられる[1,2].特に「ログ型」は過去ログを利用し,入力された発言と類似した過去の発言を,品詞など特定のキーに注目して探索し,これを元に返答文の生成を行う経験的な手法である.

一方,近年のコンピュータリソースの増加に伴い,大量のデータをコンピュータ上で扱うことが容易となった.その様ななかで,大量のサンプルデータに対しデータ解析を行い,新しい知識や規則を獲得する,機械学習などの統計的な手法に注目が集まっている.

本研究では,過去ログに対し,逐次的な学習を行い,新しい知識や規則を発見,利用することで,返答文を生成する手法を提案する.また,実際にそれを用いたシステムを実装し,学習と知識の評価を行ったので報告する.

2 システムの概要

本研究で提案するシステムは,ユーザが入力した会話文を,学習を行いながら逐次的に知識ベースへ記憶し,知識の獲得とこれを元にした返答文を生成を行うものである.学習により逐次的に得られる知識は,知識ベース内に記憶された文の中で,複数回出現した同じ並びの文字とその出現回数,文脈における前後関係である.

本稿では,入力文の記憶と学習を同時に行う知識ベースの構築,返答文生成の際に用いる文脈情報の定義,これらを用いた実際の簡単な返答文生成の手順について述べる.

2.1 文字列包含関係階層構造データベース

本システムにおいて,入力された文を逐次的に処理し,学習による情報の取得と整理を同時に行うものを,文字列包含関係階層構造データベース(以下”知識ベース”)と呼ぶ. 情報の整理は文字の並びに着目して行い,包含関係による階層構造で表現する.学習により逐次的に得る知識は,知識ベース内に記憶した文の中で,複数回出現した同じ文字列のうち,それを包含する文字列の最下位の階層に位置する文字列と出現回数が異なる文字列(以下″言い回し”)とその出現回数である.従って,知識ベース内では,入力された文と言い回しが,それぞれユニークな文字列要素(以下″知識要素”)として存在する.また,出現回数は上位に位置する知識要素と下位に位置する知識要素で異なった値をとることになる.

これを実現するアルゴリズムを以下に示す.

1. 文を入力として受け付ける
2. 入力文が知識要素として存在する場合
   [出現回数の更新処理]へ
3. ＜事前処理＞
   入力文から文字数を1ずつ減らす形で分解を行い,文字数nの時,n(n-1)/2個の文字列要素を格納した文字列要素配列Nを定義
   知識要素を格納する空の配列Mを定義
4. Nの中で最も文字数の大きい要素を要素Xとして定義
   Nに要素が存在しない場合
   [出現回数の更新処理]へ
5. Xが知識要素として存在する場合
   その知識要素X'を取得し,これをMの要素に対し包含関係に従う形で整理
   X'をMに格納,NからXを削除し,[手順3]へ
6. Mに要素が存在しない,若しくはXを包含する知識要素がMの要素以外に存在せず,且つXが入力文内に出現する回数が,Xを包含するMの要素が入力文内に出現する回数の,いずれよりも大きかった場合
   Xを知識要素X'として定義し,これを包含関係に従う形で整理
   また,上位階層に位置する知識要素の出現回数をX'の出現回数の初期値として与える
   X'をMに格納,NからXを削除し,[手順3]へ
7. Xを包含する知識要素がMの要素以外に存在し,且つ,これがMの要素を包含しない,またはXが入力文内に出現する回数が,Xを包含するMの要素が入力文内に出現する回数の,いずれよりも大きい,またはXを包含する知識要素がお互いを包含しない状態で,Mに二個以上存在する場合
   Xを知識要素X'として定義し,包含関係に従う形で整理
   また,上位階層に位置する知識要素の出現回数をX'の出現回数の初期値として与える
   X'をMに格納,NからXを削除し,[手順3]へ

[出現回数の更新処理]

1. ＜事前処理＞
   入力文から文字数を1ずつ減らす形で分解を行い,文字数nの時,n(n-1)/2個の文字列要素を格納した文字列要素群Nを定義
2. Nの中で最も文字数の大きい要素を要素Xとして定義
   Nに要素が存在しない場合,処理を終了
3. Xが知識要素として存在する場合

Machine learning of conversation pettern and development of conversation system by Knowledge-Base

† Yoshio KIKUCHI(yoshio.kikuchi.513@gmail.com)

† Kazuya TAGO(ktago@cs.teu.ac.jp)

Department of Computer Science,Tokyo University of Technorogy

1404-1 Katakura,Hachioji, Tokyo 192-0982, Japan‡

その知識要素の出現回数を1増加
NからXを削除し,[手順2]へ

2.2 文脈情報の定義

返答文を生成する際の情報の一つとして,過去に行われた会話の文脈情報を保持する.本研究では,発言Aに対し,別のユーザの発言Bが行われた場合,知識要素Aに対し,知識要素Bへ向けての重み付きリンクを与える形で定義を行う.この時の重みは,初期値を1とし,同じ応答パターンが繰り返された場合,重みを1ずつ増加させる.

2.3 返答文生成の流れ

会話において文が入力されると,知識ベースへ記憶される.この際,アルゴリズムに従い階層付けが同時に行われる.返答文の生成はこの階層情報に従って行われる.返答文生成の流れの例を図1に示す.

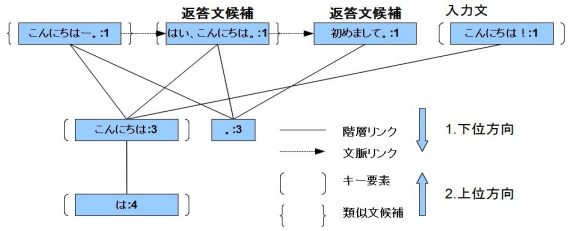

図 1:返答文生成の流れ例

返答文生成の流れを以下に示す.

1. 入力文を示す知識要素に対し,下位方向への探索を行い,得られた知識要素を類似キー要素として取得
2. 類似キー要素から,上位方向への探索を行い,この中で文脈情報を保持する知識要素を類似文候補として取得
3. 類似文候補の保持する文脈情報から,示される知識要素を返答文候補として取得
4. 返答文候補の中から選択を行い,返答文として出力

3 システムの実装

システムの構築はJavaにて行い,Javaの実行環境が導入済み環境で動作する.システムは入力文を受け取った際,知識ベースへの記憶を行い,類似キー要素,類似文候補,返答文候補の3項目を重みと共に出力する.

4 評価

インタビュー形式による日本語会話データベース[3]のコーパス・サンプルにあるテキスト100人分を学習用データとして用いた.この際,「。」と「?」と改行を1つの文の終端とし,文の合間にある相槌の表現は無視した.

29183個の文を入力した結果,164472個の知識要素と16157個の文脈情報を保持する知識ベースが構築できた.

ここでは,「こんにちは。」と「今日は何をしていましたか?」の二つの入力文を例として用い,得られる知識要素と類似文候補からシステムの評価と考察を行う.

取得した類似キー要素の階層毎の一覧を表1に,各階層の類似キー要素から得られる類似文候補の一部を表2,3に示す.

表 1:類似キー候補の一覧

	候補キー要素	回数
入力文	こんにちは。	100
下位1階層	こんにちは	137
	ちは。	102
下位2階層	は。	749
	ちは	188
	こんにち	140
	にちは	139

	候補キー要素	回数
入力文	今日は何をしていましたか?	1
下位1階層	いましたか?	9
	をしていました	5
	していましたか	3
	今日は何を	3
	は何をしていま	2
下位2階層	今日は	178
	いましたか	59
	ましたか?	44
	をしていま	34
	は何を	23
	していました	9
	ていましたか	6
	何をしていま	6
	は何をしてい	3

表 2:「こんにちは。」の類似文候補

こんにちは。					
入力文		下位1階層		下位2階層	
文	回数	文	回数	文	回数
こんにちは。	100	こんにちはー。	30	は。	748
あ、こんにちは。	12	はい、こんにちはー。	8	はは。	280
はい、こんにちは。	5	あ、こんにちはー。	4	ははは。	135
⋮		⋮		⋮	
合計10		合計14		合計363	

表 3:「今日は何をしていましたか?」の類似文候補

今日は何をしていましたか?			
下位1階層		下位2階層	
文	回数	文	回数
今日は何をつくりますか。	1	今日は	178
お兄さんは何をしていますか。	1	今日はどうもありがとうございました。	31
あの時まだソウルにいましたか?	1	今日はありがとうございました。	11
⋮		⋮	
合計11		合計209	

表1より,「いましたか?」や「今日は何を」といった,品詞などの特定のキーワードに依存しない要素を類似キー要素として,自動的に獲得できている事が分かる.また,これらの要素は下位にあたる要素と出現回数の比較を行うことで,どの程度の割合で要素が変化するのか,といった規則を統計的に判断することができる.例えば,「こんにちは」は知識要素として137回出現し,下位にあたる「こんにち」や「にちは」はこれに非常に近い割合である.従って,これら下位の要素は「こんにちは」に変化する可能性が高いことを意味し,文字の推定に利用できる可能性がある.

表2,表3より,階層構造を利用した全文に対する類似文探索が行えていることが分かる.また,「こんにちは。」に対し,形的にも意味的にも類似性の高い文が得られている.しかし,「今日は何をしていましたか?」に対しては類似性の高い候補を得ることができず,類似文候補を決める手法の改善や統計的な情報の解釈,学習量の増加を図って行く必要があると考えられる.

6 おわりに

提案する知識ベースに会話文を記憶させていくことで,特定のキーワードに依存しない類似文の全文探索や,統計的な情報により推定を行う事のできる会話システムについて報告した.今後は課題である類似文候補の獲得手法改善や学習量の増加,意味的な要素を付加することでの会話システムとしての完成度向上を行っていく.

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-:IEICE Technical Report CQ2005-36,OIS2005-21,IE2005-29(2005-09)
[2]人工無能は考える
http://www.ycf.nanet.co.jp/~skato/muno/index.shtml
[3]インタビュー形式による日本語会話データベース
http://www.env.kitakyu-u.ac.jp/corpus/