情 報 学 基 礎 61－4
自 然 言 語 処 理 142－4
（2001. 3. 5）

発話の働きとキーワードから応答生成を行う事例ベース自由対話系

江部 利明*, 乾 伸雄**, 小谷 善行**
* 東京農工大学大学院工学研究科電子情報工学専攻情報工学講座
** 東京農工大学工学部情報コミュニケーション工学科
E-mail: toshiaki@fairy.ei.tuat.ac.jp

あらまし：
事例ベース自由対話システムは人間と話題を限定しない対話を可能にするシステムである．以前開発したシステムは自由で楽しい対話を実現できたが，キーワードだけを用いて対話事例とのマッチングを行うため文脈的な対話は実現されにくい．そこで，本稿では入力発話の働き (DA) をマッチングの要素とすることにより対話の局所構造に従った自然な応答を実現する．DA を表層的に求めるため，確率的手法を導入する．bi-gram を使った手法により，58.3%の正答率が得られた．事例ベース自由対話システムの使用実験では，システムの応答に対して被験者 6 人中 5 人が以前のシステムより高い評価を与え，応答の向上が認められた．

An Example-Based Natural Dialogue System using Dialogue Acts and Keywords

Toshiaki Ebe*, Nobuo Inui**, Yoshiyuki Kotani**
* Department of Computer Science, Graduate School of Technology,
Tokyo University of Agriculture and Technology
** Department of Information and Communication, Faculty of Technology,
Tokyo University of Agriculture and Technology

Abstract:
An example-based natural dialogue system makes possible to interact a human with a computer unconstrainedly. We previously proposed such a system which provides human to enjyo talking with a computer, though there are no contextual dialogues because of only using keywords. This paper describes a way of generating natural responces using dialogue acts (DA) to extract local structures of dialogues. First, we dertermine DA of a sentence using bi-gram statistic with 58.3% correctness. Second, out dialogue system finds a appropriate example from database using DA, then finall generates a a responce. 5 subjects out of 6 subjects marked higher than our previous system in our experiment. This results showed the effectiveness of using dialogue acts in selecting examples.

1 はじめに

近年，対話システムの実用化や対話型ゲームの登場により，自然言語対話システムは人間に身近な存在になってきている．しかし，計算機と人間が自由な対話を行うことは難しく，現在の対話システムの多くはシステムに「スケジュール調整」「天気情報案内」「列車案内」などの役割を持たせることで扱う知識を限定して人間との対話を行っている．

我々は「人間と自由な対話をロバストに行う」という考えのもと，事例ベース自由対話システムの作成を行った [1]．事例ベース自由対話システムは，ELIZA 型の対話システムで，対話事例とキーワードマッチングを行い応答生成を行う (図 1)．キーワードマッチングとは，文に含まれる動詞・名詞・形容詞の中で最も重要であるものをそれぞれ一つずつ抜きだし，それを使って対話事例とのマッチングを行う方法である．図 1中の () 内がキーワードであり，(名詞, 動詞, 形容詞) の順で並んでいる (×はその品詞のキーワードがないことを示す)．応答生成は最も類似していると判断された対話事例の応答対話事例に同じキーワードが含まれる場合，それを入れ換えるという方法で行っている．図 1では，(ご飯, 食べる, ×) と (うどん, 食べる, ×) がマッチし，その応答事例に「うどん」という語があるので,「ご飯」に入れ換えて応答を生成している．

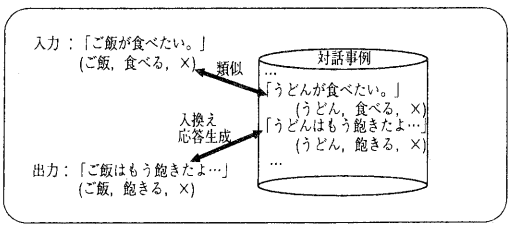

図 1: 事例ベース自由対話システム概要

システムはユーザとの対話を対話事例として追加していくことで新しい応答を覚えるため，パターン化されない応答が可能である．そして，適切な対話事例が選ばれればあたかも人間と対話しているような効果が期待できる．

このシステムの使用評価を行ったところ，80%以上の被験者がシステムと楽しく対話することができたと回答した一方，90%の被験者がその対話は不自然なものであったと回答した．この原因の一つに，キーワードだけでは「もうご飯食べた?」「どこでご飯を食べたの?」といった同じキーワードを持つ発話の働き違いが判断できないことがあげられる．そのため「どこでご飯を食べたの?」という入力に「はい，食べました.」といった応答をしてしまうので対話が不自然になってしまう．そこで，本研究では事例ベース対話システムに発話の働きの違いを確率的に求め，応答能力の向上を目指す．

2 発話の働き

対話は相手へ働きかけたり，また相手からの働きかけに応答することで進行していく．この働きかけには依頼・真偽を問う質問・未知の情報を問う質問などがあり，また応答には肯定・否定や情報の提供などの働きがある．これらの発話の働きには真偽を問う質問には肯定・否定で，未知の情報を問う質問には情報の提供で応答するといった関係がある．このような発話の働きとその関係は，対話の局所構造の分析や文脈処理 [7]，相手の発話の意図を理解する手がかり [3] として使用されている．このような発話の働きは使用目的などにより異なる分類がされている．本研究では文献 [3, 4, 7, 8] や予備調査の結果を考慮し，表 1のような独自の発話の働きを決定した．

表 1: 本研究で使用する発話の働き

挨拶	別れ	意見
意志	事実説明	理由
wh 疑問	yn 疑問	確認
依頼	提案	肯定
否定	熟考	詫び
驚き	感謝	

以後，この発話の働きを DA(Dialogue Act) と呼ぶことにする．この DA を使用し，対話

事例中の局所構造に従った対話事例を応答に選択することで応答能力の向上が期待できる．

3 文の DA の決定

3.1 DA 付きの単語 n-gram モデル

文の持つ DA を，単語の n-gram モデルを基に決定する．単語列 $w_1 \cdots w_n$ からなる文 s の DA d は，条件付き確率 $P(d'|s)$ を最大にする d であるとする．式 (1) を変形すると，$P(s)$ は d' に依存しないので式 (2) のようになる．

$$d = \operatorname*{argmax}_{d'} P(d'|s) \quad (1)$$

$$= \operatorname*{argmax}_{d'} P(s, d') \quad (2)$$

ある単語 w_n がそれ以前の単語列 $w_1 \cdots w_{n-1}$ に依存すると仮定すると，

$$P(s,d) = P(w_1 \cdots w_n, d) = \prod_{i=1}^{n} P(w_i|w_1 \cdots w_{i-1}, d)$$

となる．ここで，以前の k 個の単語にだけ依存すると仮定すると，

$$P(w_1 \cdots w_n, d) = \prod_{i=1}^{n} P(w_i|w_{i-k+1} \cdots w_{i-1}, d) \quad (3)$$

となる．この確率モデルを使った DA 判定機構を事例ベース対話システムに組み込む．

3.2 DA の判定実験

3.1節の確率モデルを使って判定実験を行う．使用するデータは小説の対話部分を抜きだしたデータとチャットのデータである．会話数・発話数・文の数を表 2に示す．

ここで，文とは句点・疑問符・感嘆符までのものである．発話とは一人の人が一度に話すものであり，複数の文を含むことがある．会話とは発話の並びである．この対話データは形態素解析されており，表 1で示した DA が文ごとに人手により付加されている．使用した形態素解析システムは RWC の品詞体系に基づく品詞分類を行う．このデータの文に付加されている DA を表 3に示す．

表 2: データの対話数・発話数・文の数

	小説	チャット	合計
会話数	49	7	56
発話数	1558	99	1647
文の数	1870	168	2038

表 3: 文に付加されている DA の数

DA	数	DA	数
挨拶	19	依頼	39
別れ	7	提案	29
意見	470	肯定	317
意志	58	否定	51
事実説明	256	熟考	37
理由	117	詫び	10
wh 疑問	163	驚き	26
yn 疑問	248	感謝	9
確認	182		

テストはこの対話データを五つにわけ，四つを訓練データ，一つをテストデータとし，平均的な精度を求める．

3.2.1 uni-gram での判定

式 (3) の $k = 1$ のとき，つまり個々の単語はそれまでの単語に依存しないで決まると考えたときのモデルで判定を行う．このモデルについて，次の二つの単語集合で実験した．

- すべて形態素レベル
- 選択した語だけ形態素レベルで他は品詞

形態素レベルで扱う語を選択するのは，DA を判定するのに重要ではないを無視するためである．また，テストデータにすべての語が出現するわけではないので，このような一般化を行うことで解析の失敗を防ぐ．疑問代名詞を除いた名詞・「願う」「ください」などを除いた動詞・「ほしい」「すごい」などを除く形

容詞を品詞レベルで扱い，それ以外の品詞は形態素レベルを扱うようにした．

この二つの判定結果の正答率を表4の訓練とテストに示す．

$$\text{正答率} = \frac{\text{正しい } DA \text{ を付加した文の数}}{\text{テストデータの文の数 (2038)}}$$

テストデータでの判定では，学習データに出現しない語があるときには確率値が0になってしまい，判定できない文があった．このような文はすべて形態素解析レベルのときには948文，選択した語だけ形態素レベルのときには285文であった．

3.2.2 品詞階層を使った確率値の補正

3.2.1節で述べたようにテストデータでは学習データに出現しない語がある場合に判定できないという問題がある．そこで，図2のように階層的に分類されている形態素の品詞情報を使って確率値の補正を行う．図2中の'*'は階層を揃えるためにつけた仮想の階層である。

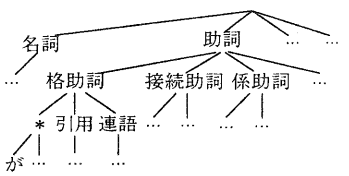

図 2: 形態素の品詞分類

ある形態素 w の一つ上の分類を w' したときの，補正確率値 $\hat{P}(w|d)$ を式 (4) のように定義する．

$$\hat{P}(w|d) = \lambda P(w|d) + (1-\lambda)\hat{P}(w'|d) \quad (4)$$

たとえば，図2の [助詞-格助詞-*-が] の補正確率値は

$$\begin{aligned}
& \hat{P}([\text{助詞} - \text{格助詞} - * - \text{が}]|d) \\
= \; & \lambda P([\text{助詞} - \text{格助詞} - * - \text{が}]|d) \\
& +(1-\lambda)\hat{P}([\text{助詞} - \text{格助詞} - *]|d) \\
= \; & \lambda P([\text{助詞} - \text{格助詞} - * - \text{が}]|d) \\
& +(1-\lambda)\{\lambda P([\text{助詞} - \text{格助詞} - *]|d) \\
& +(1-\lambda)\hat{P}([\text{助詞} - \text{格助詞}]|d)\} \\
= \; & \lambda P([\text{助詞} - \text{格助詞} - * - \text{が}]|d) \\
& +(1-\lambda)\{(\lambda P([\text{助詞} - \text{格助詞} - *]|d) \\
& +(1-\lambda)\{\lambda P([\text{助詞} - \text{格助詞}]|d) \\
& +(1-\lambda)P([\text{助詞}]|d)\}\}
\end{aligned}$$

ここで，$\hat{P}([\text{助詞}]|d) = P([\text{助詞}]|d)$ としたように，いちばん上の階層の補正値は確率値そのものとする．λ を0から1まで0.01刻みにしてテストデータによる判定を行ったところ，$\lambda = 0.99$ のときの正答率がもっとも良い正答率を得た．このときの結果を表4の補正テストに示す．

表 4: uni-gram の正答率

	訓練	テスト	補正テスト
すべて形態素	79.8%	28.3%	46.8%
一部形態素	60.2%	42.2%	46.9%

3.2.3 bi-gram での判定

ここでは式 (3) で，$k = 2$ としたとき，つまり、一つ前の単語に依存する (bi-gram) と考えたときのモデルで判定を行う．

このときの判定結果を表5の訓練とテストに示す．テストデータでは，データが小規模であるためのロバスト性の問題により，すべて形態素レベルで扱ったときは1577文が，形態素レベルで扱う語を選択したときは1188文が判定できなかった．そのためにテストデータでの正答率は uni-gram のときよりも下がった．

3.2.4 uni-gram を用いた確率値の補正

bi-gram の確率を次のように uni-gram で補正する．この確率補正値 $\hat{P}(w_{i-1}|w_i, d)$ は式 (5) のように定義する．

$$\begin{aligned}
& \hat{P}(w_{i-1}|w_i, d) \\
= \; & \gamma P(w_{i-1}|w_i, d) + (1-\gamma)\hat{P}(w_i|d) \quad (5)
\end{aligned}$$

$\hat{P}(w_i|d)$ は3.2.2節で説明した方法による補正確率である．このときの λ は3.2.2節で最も正答率が高かった 0.99 を使用した．

uni-gram のときと同様，γ を 0 から 1 まで 0.01 刻みにして open データによる判定を行ったところ，$\gamma = 0.99$ のときが最も正答率が高かった．そのときの結果を表 5の補正テストに示す．

表 5: bi-gram の正答率

	訓練	テスト	補正テスト
すべて形態素	98.1%	20.1%	56.3%
一部形態素	91.6%	31.0%	58.3%

選択した語だけ形態素レベルで他は品詞としたときに uni-gram を使って補正をしたときの bi-gram が最も正答率が高かったので，この手法を対話システムに組み込むことにした．このときの DA ごとの再現率と適合率を表 6 に示す．

表 6: 採用した条件での DA 判定精度

DA	再現率 (%)	適合率 (%)
挨拶	57.9	91.7
別れ	0.0	0.0
意見	83.8	42.8
意志	13.8	61.5
事実説明	36.3	44.5
理由	15.4	58.1
wh 疑問	63.8	87.4
yn 疑問	66.1	58.2
確認	38.5	76.1
依頼	41.0	88.9
提案	13.8	66.7
肯定	81.4	92.1
否定	37.3	90.5
熟考	37.8	100.0
詫び	0.0	100.0
驚き	50.0	100.0
感謝	44.4	100.0
正答率 (%)	1190 / 2038 : 58.3	

4 対話の局所構造を考慮したマッチング

マッチングは，まず DA を使った局所構造マッチングを行い，局所構造が類似している対話事例を集める．そして集めた対話事例に対してキーワードマッチングを行い，最も類似した対話事例を決定する．マッチングに使う DA とキーワードは，入力発話だけでなく，その一つ前の発話も使用する．

4.1 局所構造マッチング

局所構造マッチングでは，発話を文ごとに付加された DA 列として考える．ここで，DA 列の中で重要な DA を考える．DA 列の最初の DA は相手の発話に影響され，また DA 列の最後の DA は相手の発話に影響を与えることが多い．そこで，DA 列の最初と最後の DA は重要であるとして，局所構造マッチングでは次のような優先順位でマッチする対話事例を集める．

I 一つ前の発話の DA 列と入力発話の DA 列が一致する発話

II 一つ前の発話の DA 列の最初と最後の DA が同じで，入力発話の DA 列が一致する発話

III 一つ前の発話の DA 列の最後の DA が同じで，入力発話の DA 列が一致する発話

IV 一つ前の発話の DA 列と，入力発話の DA 列の最初と最後の DA が一致する発話

V 一つ前の発話の DA 列の最初と最後の DA と，入力発話の DA 列の最初と最後の DA が一致する発話

VI 一つ前の発話の DA 列の最後の DA と入力発話の最初と最後の DA が一致する発話

VII 入力発話の DA 列が一致する発話

VIII 入力発話の最初と最後の DA が一致する発話

IX 入力発話の最初の DA が一致する発話

4.2 キーワードマッチング

キーワードマッチングは，文ごとに選択した発話のキーワード列を使ったマッチングである．このキーワード列のマッチングにはDPマッチングを用いる．

動詞キーワード X，品詞情報 y をもつ名詞キーワード Y，形容詞キーワード Z の組を $(X, Y(y), Z)$ と表現するとき，二つのキーワードのコストを式 (6) のように計算する．

$$\begin{aligned} &\text{コスト}\left((X, Y(y), Z), (X', Y'(y'), Z')\right) \\ = &\begin{cases} \delta_{XX'} + \delta_{ZZ'} + 1 & Y = Y' \\ \delta_{XX'} + \delta_{ZZ'} + 0.5\delta_{yy'} & \text{otherwise} \end{cases} \end{aligned}$$

ただし，

$$\delta_{XX'} = \begin{cases} 0 & X = X' \\ 1 & X \neq X' \end{cases}$$

このDPマッチングを一つ前の発話と入力発話に対して行い，合計コストがいちばん低い対話事例を類似対話事例とする．

5 対話システム使用実験

使用評価実験では，マッチングにDAを利用することで以前の対話システムより応答能力が向上したかを調査した．

5.1 使用実験方法

使用実験は被験者6人に対して，次のような手順で行った．

(1) DAを利用していないシステムAとの対話

(2) DAを利用したシステムBとの対話

(3) 発話の自然さの採点とアンケート記入

システムとの対話の際，被験者にはDAを利用したシステムか利用していないシステムなのかは知らせていない．また，対話時間や発話数の設定は行わず，被験者は好きなだけ対話でき，また好きなときに対話を終了させることができるようにした．なお，二つのシステムには双方とも同じ対話事例を持たせた．

発話の自然さの採点は，ユーザとシステムの対話から図3のように＜システム＞＜ユーザ＞＜システム＞の順で三つの発話の組をとりだし，その個々の組の対話の自然さに対して0～10点で採点してもらった．

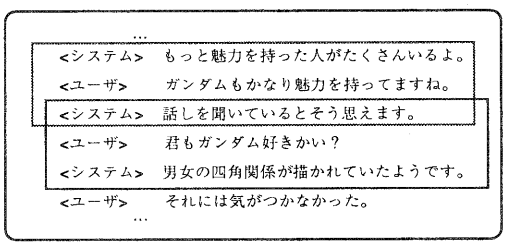

図 3: 三つの発話の組

このような採点は被験者ごとに異なった基準があるので，その基準を測るため，本実験では使用した対話事例やその他の対話コーパスから発話の三つ組を30組無作為に選択し，それに対しても採点してもらった．

アンケート調査では，AとBのそれぞれのシステムの応答に対して，全然ダメ・ダメ・まあまあ・良い・大変良いの五段階の評価と，それぞれのシステムがどのくらいの割合で自然な応答を行ったのかについて評価を行った．

5.2 実験結果

使用評価実験での使用時間・発話の自然さの採点・アンケートの結果を示す．

5.2.1 使用時間

被験者 (1)～(6) がシステムを使用した時間と入力した発話数を表7に示す．システムBの方が後で使用したにも関わらず，使用時間は平均で75秒長くなった．また，発話数は平均で25増えた．

表 7: 被験者のシステム使用時間

被験者	使用時間 (秒)		入力発話数	
	A	B	A	B
(1)	1040	1339	37	38
(2)	1134	601	59	49
(3)	741	2284	34	92
(4)	954	681	34	42
(5)	1635	1644	24	70
(6)	396	1800	28	75
平均	983.3	1391.5	36.0	61.0

5.2.2 発話の自然さの採点結果

被験者が採点した三つの発話の組の数と平均点を表 8に示す．被験者 6 人中 5 人がシステム B の方に高い得点をつけたが，対話コーパスから無作為に選択した発話の三つの組の得点には及ばなかった．

表 8: 対話ログの採点結果

被験者	A	B	無作為選択
(1)	4.97	6.64	9.27
(2)	6.53	5.89	6.90
(3)	1.93	3.50	3.00
(4)	4.60	5.50	8.93
(5)	2.52	3.71	5.93
(6)	4.78	6.03	8.87

採点された中で高得点のものを図 4に示す．

システム　はじめまして．
ユーザ　こんにちは．
システム　今日はいい天気ですね！

システム　うん．僕だってそうだったさ．
ユーザ　ご飯だったのですか?
システム　ご飯？

システム　そうですね．
ユーザ　さっそくお願いします．
システム　はい．こちらこそ．

図 4: 高得点の発話の三つ組

5.2.3 アンケート調査結果

被験者 (1)～(6) の，システム A と B の応答に対する評価を表 9に示す．割合とは，たとえば 1/20 のときは 20 応答に一つの割合で正しい応答であると感じたことを示している．

表 9: アンケート調査結果

被験者	A		B	
	自然さ	割合	自然さ	割合
(1)	まあまあ	1/3	良い	1/2
(2)	まあまあ	1/5	ダメ	1/10
(3)	まあまあ	1/15	良い	1/7
(4)	ダメ	1/6	まあまあ	1/1.5
(5)	ダメ	1/8	まあまあ	1/6
(6)	ダメ	1/8	まあまあ	1/4

6 使用実験の考察

使用評価実験では，被験者に好きなだけシステムを使用してもらったが，このような場合には被験者が実験に飽きてしまい，後に行うシステムの方が使用時間が短くなると考えられる．しかし，表 7の使用時間と発話数から，被験者は後に使用した DA を利用したシステムの方を長く使用し，また多くの対話を行ったことがわかる．この二つの結果から，DA を利用したシステムでは，以前のシステムよりもユーザの興味を引く対話が可能になったと考えられる．

発話の自然さの採点では，表 8を見てもわかるとおり，DA を利用しないシステムよりも利用したシステムの方が高い得点を得た．図 4を見てもわかるとおり，高得点のものは対話の局所的な構造に従ったものである．

アンケート調査の結果では，システム A とシステム B の比較において，被験者 6 人中 5 人が B の方を一段階高く評価した．以上の結果から，DA を利用することにより対話システムの応答能力の改善に成功したことがわかる．

7 今後の課題

今後の課題として，話題管理機構の組み込みがあげられる．今回の手法では，発話の働きにより局所的な構造に従った応答を行うことができたが対話の大局的な流れが掴めずに全体として不自然な対話となってしまう．図5では，システムはユーザの**yn疑問**に対して**肯定**や**確認**で応答しているが不自然な対話となっている．

```
...
ユーザ　　ご飯にしますか?
システム　うん．僕だってそうだったさ．
ユーザ　　ご飯だったのですか?
システム　ご飯?
...
```

図 5: 不自然な対話の流れ

8 おわりに

本研究ではDAを利用することで事例ベース自由対話システムの応答能力の向上を目指した．発話の働きは確率モデルを使って判定した．この機構をシステムに組み込み，発話の働きを用いて入力発話と対話事例とのマッチングを行った．使用評価実験では被験者6人に対して，使用時間・発話の自然さの採点・アンケートを行った．使用実験の結果，使用時間は平均75秒延びた．発話の自然さの採点では対話コーパス中の対話には及ばなかったものの，5人が以前のシステムより高得点をつけた．また，アンケート調査でも5人が以前のシステムよりも高い評価を行った．これにより，発話の働きを利用することが事例ベース自由対話システムの応答能力の向上に有効であったことがわかった．

参考文献

[1] 江部, 小島, 乾, 小谷 : 会話データとのキーワードマッチングを行い、応答文を決定する対話システム, 情報処理学会第58回全国大会講演論文集 (2), pp.281-282, 1999.

[2] J.Weizenbaum : ELIZA – A computer program for the study of natural language communications between men and machines, Communications of the Association for Computing Machinery Vol 9, pp.3–45, 1966.

[3] James F. Allen et al. : The TRAINS Project: A case study in building a conversational plannning agent, TRAINS Technical Note 94-3, 1994.

[4] Norbert Reithinger, Martin Klesen : Dialogue Act Classification Using Language Models, Proceedings of EuroSpeech-97, pp.2235-2238, 1997.

[5] Norbert Reithinger et.al. : Predicting dialogue acts for a speech-to-speech translation System, ICSLP-96, pp.654–657, 1996.

[6] Andreas Stolcke, Elizabeth Shriberg et al. : Dialogue Act Modeling for Conversational Speech,1998 AAAI Spring Symposium, AAAI Press, pp.98–105, 1998

[7] 高野敦子他 : 対話における文脈の定式化と文脈処理の枠組み, 情報処理学会論文誌 Vol.34 No.1, pp.88–97, 1993.

[8] 荒木 雅弘, 伊藤 敏彦, 熊谷 智子, 石崎 雅人 : 発話単位タグ標準化案の作成, 人工知能学会誌 Vol.14, No.2, pp.251–260, 1999

[9] 泉子・K・メイナード : 会話分析, くろしお出版, 1993.