E-069

語の関連性に基づく内容語選択と文生成パターンの置換を用いた発話文生成

Utterance Generation by Substituting the Operative Words Selected by Word Releation for Sentence Pattern.

佐藤 和* Kazu Sato　福田 雅志* Masashi Fukuda　延澤 志保† Shiho Nobesawa　太原 育夫‡ Ikuo Tahara

1 はじめに

人間と計算機との自由で自然な対話の実現は，自然言語処理研究の当初からの目標の１つである．自由な対話とは，対話の内容を制限せず，対話参加者の自由な発話を許容する対話のことである．本研究ではこのような自由な対話を対象とする．

自由な対話処理において重要なことは対話の流れをいかにして自然に行うかということである．そのための方法として，対話事例を用い，発話とそれに対する応答の対をあらかじめ用意しておく手法がある [1]．しかしこのような対話事例のみを用いる手法では，対話事例に含まれていない文を生成することはできない．

自由な対話という状況では，相手がどのような形式で，どのような内容の発話を行うか予測を立てることが難しく，あらかじめ入力と発話の対応パターンを用意しておくのは困難である．人間同士が雑談を行う際には，話されている内容と関連のある内容の発話をある程度自由な形式で行う．

本稿では，複数人での対話中に自然な発話を行うことを目標として，対話の流れに沿った発話を生成する手法を提案する．すなわち，発話すべき内容を語の集合で表現し，内容を表わす語が適切に選ばれているならば，適切に配列することで発話文が得られると考え，入力された文との関連性にもとづいて発話に用いる語を選択し，選択された語と文生成パターンを用いて発話文を生成・出力する手法を提案する．

2 対話の内容に沿った発話生成

本稿で提案する手法は，対話の流れに沿った語を選択し，選択された語を用いて文生成パターンに含まれる語を置換することで発話文を生成し，出力する．

図１に本手法の処理の流れを示す．

まず，先行する発話文をもとに発話に用いる語（内容語）の候補（内容語候補）をいくつか選択する．内容語候補に対して先行する発話文に含まれる語（手掛かり語）との関係の強さをもとにしたスコアを与え，スコアの上位の語を発話に用いる内容語として出力する．手掛かり語との関係の強さを用いることで，話の流れに沿った語が選択される．

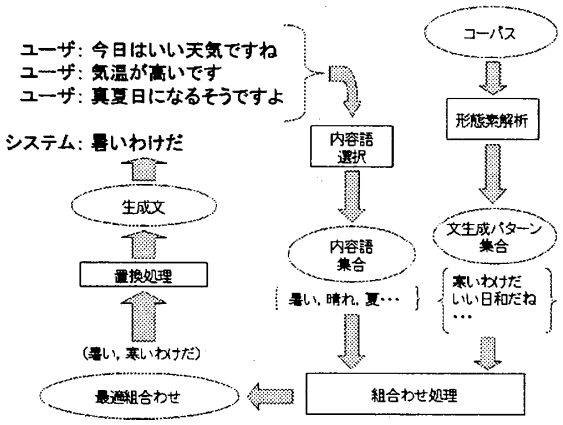

図 1: 対話の流れに着目した発話生成手法

選択された内容語と文生成パターンに含まれている自立語（置換語）を置換し，発話文を生成する．文生成パターンはコーパス中の文を形態素解析したものである．語同士の関連度に基づいて置換語と内容語の最適な組み合わせを決定する．このときの置換語と内容語の関連度をもとに，文生成パターンの適合度を計算する．すべての文生成パターンの適合度を計算し，最も適合度の高い文生成パターンを用いて発話文を生成する．この方法の特徴は，人間の作った文をもとにして関係の強い語で置換することにより，自然な発話文を生成することができる点にある．

3 連想辞書

語の選択および文生成パターンに含まれる語の置換は，語同士の関連性に基づいて行われる．語同士の関連性は連想辞書で定義される．

連想辞書は語と語の連想関係を定義した概念ベースの一種である．概念ベースを用いることで，注目する特徴を概念同士がどれほど共有しているかを比較できる．

連想辞書は見出し語と連想語，そして見出し語と連想語の対に対して定義される連想度によって構成されている．連想辞書はコーパスから自動的に作成され，コーパスに含まれる所定の品詞の語を見出し語とし，見出し語と同じ文で共起する語をその見出し語の連想語とする．

図２に連想辞書の構成を示す．図の上部の楕円“犬”，“猫”がそれぞれ見出し語であり，それらとリンクで結

*東京理科大学大学院, Graduate School of Science and Technology, Tokyo University of Science
†武蔵工業大学, Musashi Institute of Technology
‡東京理科大学, Tokyo University of Science

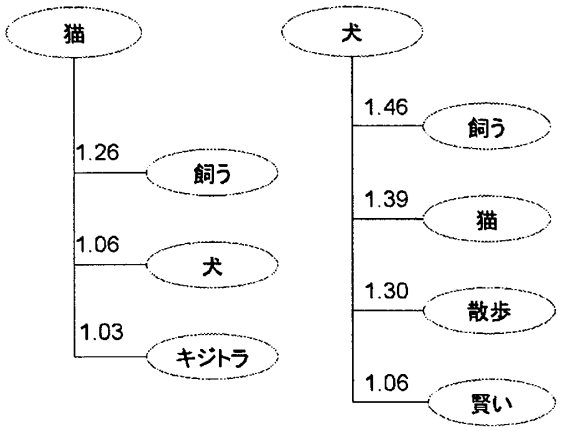

図 2: 連想辞書の構成

ばれている楕円が連想語であり，リンクに付けられた数字が連想度を表している．連想度は，見出し語との共起頻度と共起の偏りをもとに計算される．見出し語 w と連想語 w_i との連想度 $W(w_i, w)$ を式 (1) によって定める．$nsco(w_i, w)$ は w_i と w が共に出現する文の数，$nso(w)$ は w が出現する文の数，nw は連想辞書に含まれる見出し語の総数，$nwh(w_i)$ は w_i を連想語に持つ見出し語の数である．

$$W(w_i, w) = \frac{nsco(w_i, w)}{nso(w)} \cdot \log \frac{nw}{nwh(w_i)} + 1 \quad (1)$$

$\frac{nsco(w_i,w)}{nso(w)}$ は文に w が含まれていた場合に w_i が含まれているときの条件付確率であり，この値が高いほど w が含まれている文に w_i が含まれていることが期待できる．

$\frac{nw}{nwh(w_i)}$ は辞書に含まれる見出し語のうち，w_i を連想語として持っている見出し語の割合の逆数であり，この値が高いほどに w_i は見出し語 w とだけ共起する傾向が強いと言える．

連想度は，w から見て共起頻度が高く，かつ w とだけ共起する w_i ほど大きい値となる．なお，連想度はそれぞれの見出し語の持つ連想度の合計が 1.0 になるように正規化する．

4 語同士の関連度

内容語選択に用いられる内容語候補のスコアや文生成パターンの置換に用いられる適合度などを連想辞書から得られる語同士の関連度に基づき計算する．語同士の関連度としては渡部らの提案した $MatchW$ を用い，式 (2) で定義する [2].

$$MatchW(w^A, w^B) = \frac{\frac{\sum_{w_i^A = w_j^B} W_i^A}{\sum_{l=1}^{L} W_l^A} + \frac{\sum_{w_j^B = w_i^A} W_j^B}{\sum_{k=1}^{K} W_k^B}}{2} \quad (2)$$

ここで，w_i^A, w_j^B はそれぞれ w^A, w^B の連想語であり，W_i^A, W_j^B はそれぞれ w_i^A, w_j^B の連想度で，L, K はそれぞれ w^A, w^B の連想語の数である．

本手法での $MatchW$ は，関連度を測ろうとする 2 語が互いに同じ語と共起する傾向が強いほど大きな値となる．

5 内容語選択のためのスコアリング

内容語は先行する発話に含まれる語との関連度に基づいて選択されるが，先行する発話に含まれる語すべてが内容語の選択に等しく影響するわけではない．新しい発話に含まれる語ほどその影響は大きいと考えられる．そこで，スコアリングに用いる手掛かり語には出現位置に基づく重みが与えられる．手掛かり語 w^{key} の重み $Weight_H(w^{key})$ を式 (3) で定義する．

$$Weight_H(w^{key}) = \sum_{t=0}^{n} exist_t(w^{key}) \cdot f(t) \quad (3)$$

$exist_t$ は t 発話前の発話に語が含まれていれば 1 を，さもなくば 0 を返す関数であり，$f(t)$ は $t = 0$ の時 1 である単調減少関数である．$t = 0$ は最新の発話を指す．

内容語候補と手掛かり語との関連度を計算し，さらに手掛かり語重みをかけた値が，1 語の手掛かり語から内容語候補が受け取るスコアである．すべての手掛かり語から受け取るスコアの総和がその内容語候補のスコアとなる．内容語候補 w^{pic} のスコア $score(w^{pic})$ を式 (4) で定義する．

$$score(w^{pic}) = \sum_{i=1}^{n} MatchW(w^{pic}, w_i^{key}) \cdot Weight_H(w_i^{key}) \quad (4)$$

すべての内容語候補のスコアを計算し，スコアの高い語を内容語として後段の文生成に用いる．

スコアの高い内容候補語はそれまでの発話文に含まれている語と似た話題の際に出現する語であると言える．

6 文生成パターンの置換のための適合度計算

内容語の集合 S と文生成パターン R の適合度 $fit(S, R)$ を式 (5) で定義する．

$$fit(S, R) = \max_{\varphi \in \Phi} \prod_{j=1}^{N} rel(\varphi(w_j^R), w_j^R) \left(\prod_{k=1}^{N} aff(\varphi(w_j^R), \varphi(w_k^R)) \right) \quad (5)$$

ここで，w_j^R は R に含まれる置換対象語，N はその総数，φ は $\exists w_i^S \in S$ に対して $w_i^S = \varphi(w_j^R)$ なる一対一写像，Φ はそのような写像の集合である．内容語の数より置換対象語の数が多い場合，発話文生成は行わないものとし，φ による対応付けに余った $w_k^S \in S$ は無視する．また，$rel(w_i^S, w_j^R)$ は式 (6) で与えられる．

$$rel(w_i^S, w_j^R) = sim(w_i^S, w_j^R) \cdot MatchW(w_i^S, w_j^R) \quad (6)$$

ここで，$sim(w_i^S, w_j^R)$ は w_i^S と w_j^R の品詞類似度である．品詞類似度は 2 語の品詞階層が完全に一致すれば 1 を，一致しなければ 0 をとる．品詞階層とは茶筌 [3]

の出力する“名詞-一般”のように，ハイフンで区切られた品詞の各階層である．

また，$aff(\varphi(w_j^R), \varphi(w_k^R))$ は共起することが不自然な語を用いることを避けるために用いる語同士の親和度である．文に含まれている語から見た $\varphi(w_j^R)$ の連想度を用いて式（7）で定義する．

$$aff(\varphi(w_j^R), \varphi(w_k^R))) = W(\varphi(w_j^R), \varphi(w_k^R)) + 1 \quad (7)$$

見出し語から見た連想度を用いることで，同じ文中に出現することが不自然な語の組み合わせを排除することができる．

内容語の集合と文生成パターンの適合度は，以下の操作によって決定する．また，この操作により個々の文生成パターンの置換対象語と置換対象語を置換する内容語（置換語）の組み合わせを決定する．本手法で置換対象語となるのは，品詞が名詞，形容詞の語である．動詞を除外するのは，動詞を置換することで文の内容が文生成パターンの元々の内容から大きく変化してしまい，非文を生じやすくなることを避けるためである．

1. 適合度を 0 とする．
2. 置換対象語と内容語，置換語のリストをそれぞれ作る．
3. 最大部分適合度を 0 とする．
4. リストに含まれている置換対象語と内容語を組み合わせる．
5. 4 で組み合わせた語同士の部分適合度を計算する．置換語リストに含まれる語と 4 で組み合わされた内容語との親和度を計算し，部分適合度に乗算する．部分適合度が現在の最大部分適合度よりも大きいならば最大部分適合度を更新する．
6. すべての組み合わせの計算が終了していないならば 4 へ戻る．

   そうでなければ，このときの最大部分適合度の値を適合度に乗算する．また，このとき組み合わされている語の置換対象語を置換対象語リストから削除し，内容語を置換語リストに加え，内容語リストから削除する．
7. 置換対象語のリストが空になっていれば操作を終了する．このときの適合度が内容語の集合と文生成パターンの適合度である．

   そうでなければ 3 へ戻る．

この操作をすべての文生成パターンについて行い，もっとも適合度の高い文生成パターンの置換対象語を操作によって決定された置換語によって置換することで，発話文を生成する．

7 発話文生成実験

7.1 実験方法

本稿で提案した手法の有効性を検証するため，語の関連性に基づく内容語選択と文生成パターンの置換を用いた発話文生成の実験を行った．すなわち，先行する発話文を入力とし，入力をもとに生成した発話文を出力とした．入力の発話文はインターネット上で行われたチャット対話文を用いた．1 つの発話文を生成するために 5 発話の入力を用いた．55 セットの入力から 55 文の発話文を生成した．

文生成は入力 1 セットごとに行う．1 セットの入力を茶筅 [3] によって形態素解析し，含まれる名詞，形容詞，動詞を手掛かり語とする．手掛かり語の重みに用いる関数（式（3）の $f(t)$）は 0.7^t（t は語の出現位置）とした．連想辞書から手掛かり語の連想語を抽出し，内容語候補とする．内容語として出力する語はスコアの上位 20 語とした．得られた内容語と，すべての文生成パターンにより文生成を行い，最も 適合度·内容語利用率 が高い文を出力する．内容語利用率とは，与えられた内容語のうち，実際に置換語として用いた語の割合であり，適合度だけで判定した場合，短い文ほど出力されやすくなる傾向があるのでこれを抑制するために用いた．

連想辞書と文生成パターンは，入力とは別のコーパスから作成した．連想辞書・文生成パターンの作成に用いたコーパスはインターネットの電子掲示板から収集した．コーパスに収録された文は 32318 文である．連想辞書を作成した後，各見出し語に対して連想度の上位 100 位未満の連想語を削除する操作を施した．連想辞書に収録された見出し語は 7571 語である．文生成パターンはコーパスに含まれる文のうち，未知語を含む，アルファベットを含む，連想辞書に収録されていない自立語を含む，少なくとも 2 語の置換対象語を含まないの 1 つ以上に該当する文を除いた文 12182 文から作成した．

本実験では自然な文を生成するため，1 度内容語に選択された語のスコアは，恒久的に 1/2 を乗じている．こうすることで，かつての話題として出現した語で，再び出現した語（すなわち大域的な話題に関わる語）を文の生成に用いず，現在の新しい話題に沿った語（局所的な話題に関わる語）によって文を生成することが可能になる．しかし，こうした処理を行っても同じ文が何度も生成されることがある．同じ文を何度も出力することは不自然であるため，出力文は未出の文のうちで最も 適合度·内容語利用率 が高い文を出力することにした．

7.2 実験の評価方法

実験結果の評価は，先行する対話文と生成された発話文を被験者に示し，

1. 先行する対話文と関連性のある語が使われていること
2. 発話文として容認できないほど文法的に誤っている，または意味的にずれた文でないこと

表 1: 実験結果

評価	文数（文）
3 つの条件を満たす文	6
条件 1 と条件 2 のみを満たす文	12
条件 2 のみを満たす文	18
条件 1 のみを満たす文	11
すべての条件を満たさない文	8
合計	55

表 2: 3 点の条件を満たす文の例

F: 寝まぁす！　(*　　)　またね♪ F さんが退室しました。 C: またね～ B: 獣医さんが言うには、 米は、すぐにエネルギーに代わるから、 体力付くんだよ～って言ってたらしい。。。 D: そうですね。 でも、タオルとかも食べちゃうので、 そっちのほうが体に悪そうですが…
生成文: 獣医さんも人によって言うことが

3. 先行する対話文の内容に沿った内容であること

以上の 3 点について個別に評価してもらった．各発話文についてこれら条件を満たしているかどうかは，被験者の半数以上の判断により決定した．

7.3 実験結果

実験結果の評価は 8 名の被験者が行った．結果を表 1 に示す．表 2 に 3 点の条件を満たす生成文の例，表 3 にいずれの条件も満たさない文の例を示す．

対話の内容に関連のある語を用い，発話文として容認できる文は全体の 32.7%であり，そのような文の 33.3%が，対話の内容に沿った内容であると判定された．

この結果から，本手法によって対話の内容と関連のある語を選択し，選択された語を用いて自然な発話文を生成することが可能であることを示している．

7.4 考察

対話の内容に関連のある語を用い，発話文として容認できる発話文のうち，対話の流れに沿った内容の発話文は 33.3%であった．このことは，対話の内容に関連のある語を自然に配列することで，対話の流れに沿った自然な発話生成が可能であることを示している．

一方，条件 1 を満たさない文，すなわち先行する対話の内容に関係のある語を含まないと判定された文の合計は全体の 47.3%に上る．これは，出現頻度の高い一般的な語を用いたほうが文生成パターンとの適合度が高くなりがちな傾向があるために，そのとき話されている内容よりも一般的な内容と関係のある語が文生成に用いられることが影響していると考えられる．

表 3: いずれの条件も満たさない文の例

C: ｗｗｗｗｗ D: ちゃんと許可とってひざの上に乗るの？えらいね～ B: なので、膝を叩いて、おいで～って言ったら、 乗ってきてちゅ～してきたので寝るのを許可した www C: うちのは上半身だけ私の片膝に乗ってくる A: うちのは勝手に乗るなあ
生成文: 人の猫やたらと最近

条件 2 を満たさない文，すなわち発話文として容認できないほどの文法的誤りまたは意味のずれを含むと判定された文の合計は全体の 34.5%である．このような文には“猫もキジトラです”のような宣言的な文，2 つの名詞のみからなる文，スラングを含む文，途中で終わっている文などがある．宣言的な文に関しては 2 つの語がどのような関係であるかを記述した知識を持たないこと，そのほかの文に関しては自動作成した文生成パターンの中に不適切なパターンが含まれていることが原因であると考えられる．

8 おわりに

対話の内容に関連のある発話が自然な発話であるという考えに基づき，対話の内容に関連のある語を選択し，選ばれた語を適切に配列することで発話文を生成する手法を提案した．先行する対話文に含まれる語を手掛かりとしたスコアリングによる内容語選択と，それにより選択された語と文生成パターンに含まれる語を置換する文生成手法によって，対話の内容に沿った自然な発話文を生成できることを示した．このことから，対話の内容に関係のある語を適切に配列することで得られた文は，対話の内容に沿った自然な発話文となり得ると言える．

しかし，実験結果は本手法だけでは不十分であることも示している．自由な対話であっても，対話参加者の発話には意図があり，対話参加者が互いに意図を理解することで，自然な対話が行われると考えられる．そのため，自然な発話生成の実現のために，相手の意図を理解し，発話生成に反映させることが必要である．

参考文献

[1] 小磯拓也，乾伸雄，小谷善行，“相互情報量を用いた話題語集合による対話の応答選択，”情報処理学会研究報告，2004-NL-160，pp.101–108，2004.

[2] 渡部広一，河岡司，“常識的判断のための概念間の関連度評価モデル，”自然言語処理，Vol.8，No.2，pp.39–54，2001.

[3] 松本裕治，北内啓，山下達雄，平野善隆，松田寛，高岡一馬，浅原正幸，“形態素解析システム『茶筌』version 2.3.3 使用説明書，”2003.