E-067

自由対話のための語の関連性に基づく応答文生成

Sentence Generation Based on Word Association for an Automatic Conversation System

佐藤 和* 福田 雅志* 佐野 智久† 延澤 志保‡
Kazu Sato Masashi Fukuda Tomohisa Sano Shiho Nobesawa
太原 育夫§
Ikuo Tahara

1 はじめに

人間と計算機が自由な対話を行うことを目標とした研究は，計算機で言語を扱いはじめた初期から現在に至るまで行われている [6]．本稿では 2 名間の雑談を対象とし，先行する発話を基に適切な応答文を生成するため，与えられた応答候補語を利用して応答文を自動生成する手法を提案する．

近年の研究では，小磯らはマッチング用の文と応答用の文からなる応答対を用い，ユーザの入力したテキストと応答対のマッチング用の文を比較し，最も適切なマッチング用の文を持つ応答対の応答用の文を出力するシステムを開発した [4]．また，黒田らは対話事例から応答文生成ルールと表層文生成ルールを自動的に学習するシステムを開発した [3]．このシステムは入力に対して応答文生成ルールを適用することで応答文に用いる語を選択し，応答文に用いる語に対して表層文生成ルールを適用することで応答文を得ている．また，ルールを遺伝的アルゴリズムと帰納学習によって汎用的なルールを生成している．

実際の人間の対話である対話事例を用いて応答を行うことは，自然な応答を行うために有効と考えられるが，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．また自由な対話では複雑で多様な文が出現し，汎化によって有効なルールを得るためには多くの対話事例が必要であると考えられる．機械翻訳における翻訳文生成の手法として，複数の依存構造を組み合わせて 1 つの依存構造を構成し文を生成する手法がある [1]．しかし自由対話では入力文に対して出力文の依存構造が定まらないため用いることは難しい．

本稿で対象とする雑談は明示的な話題を持たず，対話相手の発言やそれまでの対話の内容と関連性を持つ内容を自由に発言する点に特徴がある．自由な対話といっても，発言と発言の間には何らかの関連性が保たれていなければならない．先行する対話の内容と関連性のない発言を行えば，対話に混乱をきたすからである．そのため，応答を行う際には，先行の発話との関連性を保つことが重要である．

*東京理科大学大学院,Graduate School of Tokyo University of Science
†慶應義塾大学大学院, Keio University Graduate School
‡武蔵工業大学, Musashi Institute of Technology
§東京理科大学, Tokyo University of Science

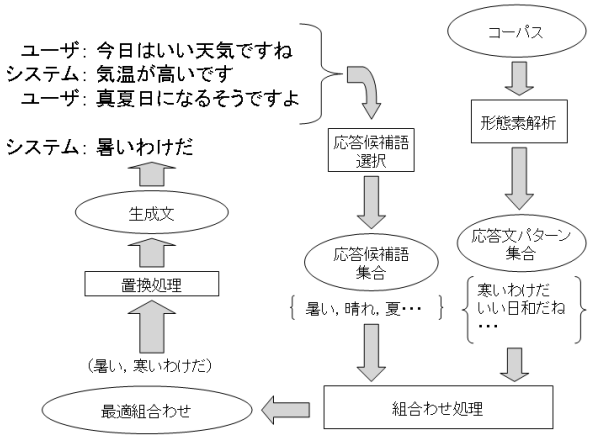

図 1: 発話の流れに着目した応答文生成手法

そこで，対話の内容に関連のある語を用いて文を生成することで，対話の内容に合った文を生成することができると考えた．

2 発話の流れに着目した対話システム

2.1 提案手法の概要

ここで提案するシステムにおいて想定する対話システムは，先行する対話文の列を入力とし，対話の内容や流れを反映して応答文に用いる語の候補を選択し，候補語を用いて生成可能な文の中で最も優れた文を応答として出力するものである．図 1 に本手法全体の処理の流れを示す．本稿では対話の流れに沿う語を用いて応答文を生成する手法を提案する．従って，本手法の入力は対話文の列ではなく，先行対話文の列をもとに得た応答候補語集合とする．

2.2 先行発話の解析

佐藤らは自然な対話における発話と発話の間には関連性が維持されていると考え，人間の発言に対する応答文に用いるための適切な語を選択する手法を提案した [5]．この手法は対話の流れを考慮して先行する発言に含まれる語との関連度を基に応答に用いる応答候補語を選択し出力する．すなわち, この手法では，対話の流れを過去の対話文に含まれていた語の集合として表現している．これらの語には重みがつけられており，新しい語ほど大きな重みが与えられ，過去に遡るほど重みは小さくなる．重みの大きい語ほど話の流れ

の中で重要な語であるとし，出力語の選択に大きな影響を持たせている．この重み付きの語集合を手がかりとし，出力候補の語に対して相互の関連度に基づいたスコアリングを行い，スコアの高い語を応答候補語として出力している．この手法によって出力された語のうち，39.5%の語が応答文に用いる語として適切であるという結果が出ている．

2.3 応答文の生成

本稿では，対話の内容を反映した語が応答文に用いる語の候補として与えられているという前提で，コーパスから作成された応答文パターンに含まれる自立語を，与えられた語で置換することで新たな文を生成する手法を提案する．置換の際には応答文パターンの置換対象の語と与えられた語との間の適合度を計り，最も適合度が高い語によって置換を行う．適合度は後述する連想辞書を元にして計算される値である．既知の文に含まれる語を置換することによって文生成を行うことで非文が生じにくく，また置換の際に適合度を用いて置換する語の判定を行うことで内容にまとまりがある文を得られるという利点が考えられる．

応答文パターンはコーパスから自動的に作成した文のテンプレートである．本システムは応答文パターンに対して，適合度に基づいて入力語の中から応答文パターンに含まれるそれぞれの語に対して最適な語を選択し，選択された入力語で応答文パターンに含まれる語を置換し，文生成を行う．この処理を全ての応答文パターンに対して行い，最も適合度の高い文を出力する．

3 関連度に基づいた応答文パターンの置換による応答文生成

3.1 語の関連度

語の関連度とは注目する特徴に関して語がどれほど似ているかを示す値である．本稿では文に含まれている語を関連度の高い語で置換しても文が成り立つ必要がある．

ある文に含まれる語を，似た文に出現する別の語で置換した場合，2 語は似た文で出現する傾向が強いことから，文が自然な形になることが期待できる．また，名詞を動詞で置換するなど異なった品詞で置換することは非文を生じさせる原因となる．

そのため文の内容が近いという点から共起語の類似度，また用いられ方という点から品詞の類似度という 2 つの特徴を用いて関連度を定義する．ある 2 語の共起語が似ているということは，その 2 語は似た語と共に用いられるということであり，その 2 語が出現する文同士は似た語で構成されていると言える．似た語によって構成されている文は，内容も似ていると考えられるため，共起語の類似性の高い 2 語は，似た文に出現する可能性が高いと考えられる．

3.2 連想辞書

連想辞書と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．概念ベースは概念を多数収録したデータベースである．概念の表現としては，概念の持つ属性と，その属性の概念との関連の強さを表す重みの対の集合で表す手法が用いられている [8, 2]．概念ベースを用いることで，概念同士がどれほど注目する特徴を共有しているかを比較できる．

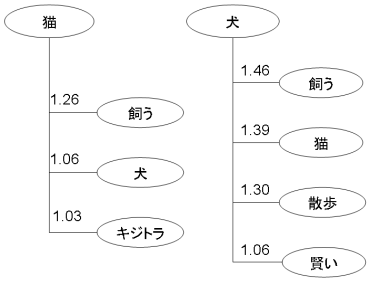

図 2: 連想辞書の構成

連想辞書はコーパスから自動的に作成することが可能である．コーパス中の同一の文中で共起する語同士を互いの連想語と定義する．適合度は同一の連想語を多く持つほど高くなる値として定義している．また，それぞれの連想語には重みが定められている．重みは見出し語に対してどれほど重要な連想語であるかを示す値である．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語重みを表している．

3.3 連想語の重み

連想語の重みは見出し語ごとに定め，それぞれの見出し語との共起頻度と共起の偏りを基に定める．ある見出し語 w に対する連想語 w_i の重み $Weight(w_i, w)$ は w_i と w が共に出現する文の数 $nsco(w_i, w)$ と w が出現する文の数 $nso(w)$，連想辞書に含まれる全ての見出し語の数 nw，見出し語のうち w_i を連想語に持つ語の数 $nwh(w_i)$ を用いて式 (1) のように定義する．

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

$\frac{nsco(w_i,w)}{nso(w)}$ は文に w が含まれていた場合に w_i が含まれている条件付確率であり，高いほど w が出現する文に w_i も出現することが期待できる．これを共起しやすさの尺度とする．また，$\frac{nw}{nwh(w_i)}$ は辞書に含まれる見出し語のうち，w_i を連想語に持つ見出し語の割合の逆数であり，高いほど共起する語が限定される．これを共起の偏りとする．w に対して重みの大きい w_i は頻繁に共起し，また w が含まれる文に特有の語であると考えられることから w から連想しやすい語であると言える．また，“する ”，“ある ”などの高頻度でかつ多くの語と共起する語を共起語として持っていたとしても似た文に出現するとは言えないため，後述する適合度の計算において，高頻度で一般的な語によって適合度が高くなることを避ける効果も期待できる．

3.4 関連度に基づいた応答文パターンの置換

応答文パターンは形態素解析されているものとする．応答文に用いる語と応答文パターンに含まれる語との適合度に基づいて語の置換を行い，文を生成する．語は適切な語が与えられる前提のため，文生成の際には非文が生じないこと，また文生成に用いる語の選択が重要である．形態素解析以外の処理は施す必要がない

ため，容易に大量の応答文パターンが収集可能である．また，話の流れや内容などに沿った語が選択されることを想定しているため，対話文コーパス以外からの応答文パターン獲得が可能となっている．

与えられた応答文に用いる語の集合に対し，応答文に用いる語（与えられた語の部分集合）と用いる応答文パターン，及び応答文パターンに含まれる語との対応付けを決定する手法を説明する．応答文パターンに含まれる置換の対象となる語を置換対象語，与えられた語を応答候補語，応答候補語のうちある応答文パターンと組み合わされた語をその応答文パターンの置換語とする．置換対象語と応答候補語の組み合わせは以下の手法で決定する．

1. 最大適合度を 0 とする．
2. 置換対象語と応答候補語を組み合わせる．
3. 適合度を計算する．適合度が現在の最大適合度よりも大きいならば最大適合度を更新する．
4. 全ての組み合わせの計算が終了していないならば 2 へ戻る．そうでないならば適合度の最大値を持つ組み合わせによって置換を行う．

応答候補語の集合 S と応答文パターン R の適合度 $fit(S,R)$ を式（3）で定義する．

$$fit(S,R) = \max_{\varphi\in\Phi} \prod_{j=1}^{N} rel(\varphi(w_j^R), w_j^R) \tag{2}$$

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．応答候補語の和より置換対象語の和が多い場合，応答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 (3) で与えられる．

$$rel(w_i^S, w_j^R) = sim(w_i^S, w_j^R) \cdot MatchW(w_i^S, w_j^R) \tag{3}$$

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 [7] の出力する“ 名詞-一般 ”のように，品詞のハイフンで区切られた各階層である．また，$MatchW(w_i^S, w_j^R)$ は渡部らの手法を用いて式 (4) で定義する [8]．

$$MatchW(w_i^S, w_j^R) = \frac{\frac{\sum_{w_{i_l}^S = w_{j_k}^R} W_{i_l}^S}{\sum_{l=1}^{L} W_{i_l}^S} + \frac{\sum_{w_{j_k}^R = w_{i_l}^S} W_{j_k}^R}{\sum_{k=1}^{K} W_{j_k}^R}}{2} \tag{4}$$

ここで，$w_{i_l}^S, w_{j_k}^R$ はそれぞれ w_i^S, w_j^R の連想語であり，$W_{i_l}^S, W_{j_k}^R$ はそれぞれ $w_{i_l}^S, w_{j_k}^R$ の連想語重みで，L , K はそれぞれ w_i^S, w_j^R の連想語の数である．この $MatchW(w_i^S, w_j^R)$ は互いの連想語が似通っているほど高い値となる．すなわち，共起する語が共通しているほど $MatchW(w_i^S, w_j^R)$ は高い値となる．また，共起頻度の高い語を共有しているほど $MatchW(w_i^S, w_j^R)$ は高い値となるが，出現頻度の高い一般的な語を共有していても $MatchW(w_i^S, w_j^R)$ は高い値とはならない．このことから $MatchW(w_i^S, w_j^R)$ が高い 2 語は多くの共起語を共有しており，さらにそれぞれの語は頻繁に共有する共起語と共に出現し，なおかつ共有する共起語はそれぞれの語に特有のものであるため，その 2 語は同じ語を多く含む文で用いられると考えられる．また，$sim(w_i^S, w_j^R)$ の値が高い 2 語は品詞が類似している．このことから，$rel(w_i^S, w_j^R)$ が高い 2 語は，含まれる語が似た文で用いられる語であり品詞も類似しているということになる．そのため，w_i^R を w_j^S で置換して生成される文は内容的なまとまりが損なわれないと考えられる．

4 応答文生成実験

4.1 実験方法

応答文パターンの置換対象語を応答候補語で置換することで適切な文が生成されることを示す．実験は予め応答候補語が与えられているという前提の下で行う．そのため，コーパス中の文に含まれている語を入力とする．出力された文の評価は被験者が行い，文の内容が理解できるかどうかを評価する．

応答文パターンの置換対象語それぞれに対して，入力されたすべての応答候補語との関連度を計算し，最も関連度の高い組み合わせから順に決定していく．生成文の適合度に $\frac{\text{置換語の数}}{\text{応答候補語の数}}$ をかけた値を文のスコアとする．

文生成に必要な応答候補語が与えられているという前提のため，コーパスの 1 文に含まれる内容語を 1 セットの応答候補語とする．ただし応答文パターンに含まれる語と同一の語が応答候補語に含まれている場合，応答文パターンに含まれる語は同一の語によって置換され変化しない．そのため応答文パターンとまったく同じ生成文は評価の対象とはしない．応答候補語に対して適合度が最も高くなる応答文パターンを探索し，応答文パターンに含まれる置換対象語を置換語によって置換し出力文とする．

全 33,408 文からなるコーパスを 10 分割し，分割されたコーパスの 9 個（30,067 文）を応答文パターン及び連想辞書の構築に用い，残りの 1 個のコーパス（3,341 文）に含まれる文から応答候補語を作成した．ただし応答文パターンには最低でも 2 語の置換対象語が含まれるため，1 セットに含まれる語が 2 語未満の応答候補語については文生成を行わなかった．連想辞書に収録されている語を少なくとも 2 語含む応答候補語は 2,054 セットあり，そのうち 2,053 セットの応答候補語に対して文が生成された．

生成文のうち応答文パターンと異なる文は 1,478 文であった．応答文パターンと異なる生成文から重複を削除し，スコアの上位 200 文を評価の対象とする．

被験者に生成された文を示し，対話の中に出現すれば内容を理解し得る文には良を，内容を理解し得るが，文の構造（語順，活用形，助詞・助動詞等）に誤りがある文には可を，内容を理解できない文には不可をつけることで評価する．

表 1: 実験結果

評価	数（文）
適切な文	126
不適切でない文	8
不適切な文	66

応答候補語	子猫，里親
応答文パターン	猫の里親見つかりました
生成文	子猫の里親見つかりました

図 3: 文生成の成功例

4.2 実験用連想辞書・応答文パターン作成

実験に用いる連想辞書と応答文パターンは同じコーパスから作成する．コーパスは電子掲示板の文書を用いて作成した．電子掲示板の文書は，その内容が語りかけや質問，またはそれらに対する応答など，比較的対話文に近いと考えられ，また大量の文書を容易に収集できるため，本稿ではコーパスとして採用した．電子掲示板の文書から日付などを削除し，句点及び HTML の要素
を文の境界とした．連想辞書や応答文パターンはこの文を単位として作成した．連想辞書に収録された語は，名詞-サ変接続，名詞-ナイ形容詞語幹，名詞-一般，名詞-形容動詞語幹，名詞-固有名詞-一般，名詞-固有名詞-人名-一般，名詞-固有名詞-人名-姓，名詞-固有名詞-人名-名，名詞-固有名詞-組織，名詞-固有名詞-地域-一般，名詞-固有名詞-地域-国，形容詞-自立，動詞-自立の語で，コーパス中の 2 文以上の文で出現した語である．連想辞書に収録された語は 7,513 語である．

応答文パターンの置換対象語は連想辞書に収録された語から動詞を除いた語である．ただし品詞が未知語または記号-アルファベットの語を含むもの，1 語のみからなるもの，連想辞書に収録されていない自立語を含むもの，置換対象語が 2 語未満のものは応答文パターンから除外した．この処理は自動的に行う．得られた応答文パターンは 10,824 文であり，応答文パターン構築のためのコーパスに含まれる文全体の量の 36.0%である．

4.3 実験結果

評価対象文を 3 名の被験者に評価させ，2 名以上の被験者が良と評価した文を適切な文，2 名以上の被験者が不可と評価した文を不適切な文，どちらでもない文を不適切でない文とする．実験結果を表 1 に示す．適切な文と不適切でない文の割合は 67.0%である．実験結果より，応答文パターンに含まれる語を応答文に用いる内容語で置換することで適切な文生成が行えることが示された．

応答候補語	うち，目
応答文パターン	みなさんのキジトラちゃんの目は何色ですか??
生成文	みなさんのうちちゃんの目は何色ですか??

図 4: 文生成の失敗例

4.4 考察

文生成の成功例を図 3 に示す．応答文パターンの“ 猫 ”と候補語の“ 子猫 ”が適切に置換されており，里親は元々文に含まれているため変化しない．この例では入力された語から，内容のまとまりがあり文法的にも正しい文が生成されている．

また，文生成の失敗例を図 4 に示す．応答文パターンの“ 目 ”はそのままに，“ キジトラ ”が候補語の“ うち ”で置換され，内容が理解できない文が生成された．“ うち ”はコーパス中では“ うちの ”のような形で“ うちのもの ”の意味で使われる．特に“ うちの猫 ”という意味で頻繁に使われる語である．また，“ キジトラ ”は猫の柄であり，比ゆ的に“ キジトラ柄の猫 ”を指す語としてコーパスに出現する．“ うちの ”という表現から助詞“ の ”が脱落し，またルールに前に出現する語と切り離された接尾語“ ちゃん ”が含まれていたため，不自然な文が生成されたと考えられる．

5 おわりに

応答文パターンに含まれる語を品詞の類似性と共起に基づく関連度が高い語で置換することで，文法的にも内容的にも適切な文が生成されることを示した．文生成実験では適切な文と不適切でない文の割合は 67.0%の精度を得ることに成功した．このことから，応答文パターンに含まれる語の置換によって文を生成することで，自立語と既存の文から新たな文を生成できることが示された．

参考文献

[1] 荒牧英治，黒橋禎夫，柏岡英紀，“ 用例ベース翻訳確率的モデル化，”自然言語処理，Vol.13，No.3，pp.3–19，2006.

[2] 笠原要，松澤和光，石川勉，“ 国語辞書を利用した日常語の類似性判別，”情報処理学会論文誌，Vol. 38，No.7，pp.1272–1283，1997.

[3] 黒田道友，荒木健治，“ 雑談を対象とした SeGA-ILSD の多言語に対する汎用性の評価，”情報処理学会研究報告，2007-NL-177，pp.79–85，2007.

[4] 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.

[5] 佐藤和, 延澤志保, 太原育夫, “ 連想に基づいた応答文生成のための内容語選択, ” 情報処理学会第 68 回全国大会, Vol.2, pp.483–484, 2006.

[6] 長尾真，“ 自然言語処理，”岩波講座ソフトウェア科学，岩波書店，pp.343–345，1996.

[7] 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.

[8] 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.