2D4-4

n-gram モデルとトピックモデルと係り受け解析の統合による自然文サンプリング法

natural sentence sampling method using integration of n-gram model, topic model and dependency parsing

山田 優樹 Yuuki Yamada　谷口 忠大 Tadahiro Taniguchi　高野 敏明 Toshiaki Takano

立命館大学
Ritsumeikan University

In this paper, we propose a method that generates natural sentence as a reply to a given input sentence. For replying natural sentence, we need to consider two important points: syntactic structure is correct and its topic is same as input sentence. Our propose method considers that (1) word transition, (2) topics of input sentence, (3) syntactic structure. (1) and (3) are related to grammar, simultaneously by using n-gram model and dependency parsing. For realizing (2), we use latent Dirichlet allocation which allows a computer agent estimating latent topics in given input data on the basis of training data given to the system previously. In the evaluation experiment, our system made a reply sentence by 4 different methods including our proposal method. Through the experiment, we showed that system could generate sentences related to topics of input sentence and having more natural structure than other methods.

1. はじめに

近年，人間と会話を目的とするシステムの研究や開発が盛んに行われている．人間と会話をするシステムには，特定の質問と応答を行う観光案内システムや，自由な内容のやりとりを行うチャット bot などがある．質問応答システムではユーザからの入力文に対する返答文のルールをあらかじめ決めておく手法が多く用いられている [1]．しかし，私たちの会話には数多くの話題が現れ，それに対応する返答文を生成する為に膨大な量のルールを決定することは困難である．大規模なテキストデータから統計的に文を生成する手法では，多くの返答文ルールを定めることなく，多様な話題に対応することが期待できる．本研究では会話における文の生成過程を確率モデルのみによって表現するアプローチを採る．ユーザから与えられた文に対して，その文の内容に沿う文を生成するシステムの構築を目指す．

本論文で構築するシステムは，ユーザから与えられた入力文に対して，あらかじめルールで決められた文型ではなく，まず文を構成する単語の生起確率に基づいて文を生成する．その際，ユーザの発話内容のトピックとシステムが生成する文の単語間の係り受け関係を考慮して返答文を生成する．そして，そのシステムによって生成された発話がユーザにとって自然な発話か検証する．

2. 発話文生成手法

会話において人間が発話を生成する過程を考える．ただし，本モデルでは，発話を単語の系列で表す．発話を構成する単語の連続的な生起において，単語はその単語より前に生起する単語が影響していると考えられる．私たちは会話において，聞き手は話し手の発話内容に沿った返答をするために，現在の会話のトピックを推定しなければならない．ここでトピックとは，会話中の話題となる事柄や分野を指す [2].「人工知能」のトピックについて話すときに,「ロボット」や「コンピュータ」といった単語はよく用いられるが,「マグロ」はほぼ用いられることはないだろう．実際の単語の生起は，N-gram モデルで表されるような直前の単語の影響だけでなく，会話のトピックによる影響も受けていると考えることができる．しかし，単語をランダムに並べるだけでは，発話内容を聞き手に正しく伝えることは難しい．日本語では文節間の依存関係が文の係り受け構造を決めており，依存関係が尤もらしい文が意味がわかりやすい文となる．本研究では，これらの要素を以下の 3 つの確率モデルとして扱う．

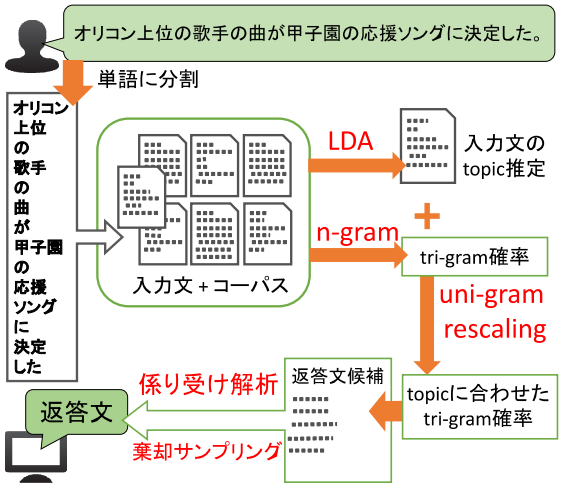

図 1: システム概要図

- N-gram モデルによる単語の生起確率
- トピックによる単語の生起確率
- 係り受けによる文の統語構造に基づく確率

このモデルに基づいて文を生成し，ユーザに出力するシステムを構築する．図 1 にシステムの概要を示す．

連絡先: 高野敏明，立命館大学情報理工学部，takano@em.ci.ritsumei.ac.jp

2.1 N-gram モデルによる単語の生起確率の影響

文を構成する単語の生起確率には N-gram モデルを用いる．N-gram モデルでは，ある単語 w_i が生起する確率 $P(w_i)$ は w_i より前に生起した単語に影響されると考える．

$$P(w_n|w_1^{n-1}) = P(w_n|w_{n-N+1}^{n-1}) \tag{1}$$

また，本研究の N-gram モデルには Kneser-Ney スムージングを用いる [3]．bi-gram 確率における Kneser-Ney スムージングは以下の式で表される．

$$P_{KN}(w_i|w_{i-1}) = \frac{\max(c(w_{i-1}, w_i - d, 0))}{c(w_{i-1})} + \lambda(w_{i-1})P_C(w_i) \tag{2}$$

ただし，

$$\lambda(w_{i-1}) = \frac{d}{c(w_{i-1})}|w : c(w_{i-1}, w) > 0| \tag{3}$$

d は割引率で $0 < d < 1$ の実数値である．式 (2) の $P_C(w_i)$ は w_i の直前に単語が生起する確率で，以下のように与えられる．

$$P_{C(w_i)} = \frac{|w_{i-1} : c(w_{i-1}, w_i) > 0|}{\sum_{w'_i} |w'_{i-1} : c(w'_{i-1}, w'_i) > 0|} \tag{4}$$

$|w_{i-1} : c(w_{i-1}, w_i) > 0|$ は w_i の直前に現れる単語の種類数を表す．Kneser-Ney スムージングでは，ある単語の直前に生起する単語の種類数を用いて確率値の補正を行っている．

2.2 トピックモデルによる単語の生起確率の影響

本論文では文のトピック推定には LDA(Latent Dirichlet Allocation) を用いる [4]．LDA では，文書には複数のトピックが存在し，それぞれのトピックから単語が生起し，文書が構成されるという考えのもと，トピック，単語の確率分布を推定するモデルである．

本システムではユーザからの入力文を一つの文書として扱い，ユーザからの入力文のトピックの推定を行う．入力された文のトピックに則した文を生成するため，推定されたトピックを用いて N-gram モデルの単語の生起確率を調整する．推定されたトピック z が与えられたときの N-gram モデルによる単語の生起確率 $P(w_i|w_{i-1}^{i-n+1}, z)$ を uni-gram rescaling[5] により近似することで，生起確率を調整する．

$$P(w_i|w_{i-1}^{i-n+1}, z) \propto \frac{P(w_i|w_{i-1}^{i-n+1})P(w_i|z)}{P(w_i)} \tag{5}$$

この式によって，直前 $(N-1)$ 単語のみの情報によって生成される N-gram モデルを，入力文のトピックを考慮した確率モデルに変更している．

2.3 係り受けによる文の統語構造

日本語の係り受け解析では以下のモデルがよく用いられている [6]．m 個の文節からなる文節列 $\{b_1, b_2, \ldots, b_m\}$ を B と定義する．$Dep(i)$ が文節 b_i の係り先を表すとし，係り受けパターン列 $\{Dep(1), Dep(2), \ldots, Dep(m-1)\}$ を D と定義する．なお，D は以下の 4 つの係り受けの制約を満たすものとする [7].

特徴 (1) 係り元から係り先は前方から後方に向かっている（後方修飾）

特徴 (2) 係り受け関係は交差しない.（非交差条件）

特徴 (3) 係り要素は受け要素を一つだけ持つ.

特徴 (4) ほとんどの場合，係り先決定には前方の文脈を必要としない.

統計的係り受け解析では，この係り受けの制約のもと，入力文節列 B に対する条件付き確率 $P(D|B)$ を最大にする係り受けパターン列 D を求めることになる．

$$D_{best} = \underset{D}{\operatorname{argmax}}\, P(D|B) \tag{6}$$

ここで各文節の係り受けの関係は独立であると仮定したとき，

$$P(D|B) = \prod_{i=1}^{m-1} P(Dep(i) = j|f_{ij}) \tag{7}$$

$$f_{ij} = \{f_1, \ldots, f_n\} \in R^n \tag{8}$$

と変形できる．ここで $P(Dep(i) = j|f_{ij})$ は，文節 b_i と文節 b_j が言語的素性集合 f_{ij} を持つときに文節 b_i が文節 b_j に係る確率を表す．言語的素性集合は，単語の品詞，語形，文節間の距離，括弧の有無などが挙げられる [8]．本研究では文節 b_i が文節 b_j に係るか係らないかの判定には日本語係り受け解析器 CaboCha[*1] を利用し，係り受け解析を行う．CaboCha はある単語が直後の単語に係るかどうかについて推定する際に，二値分類問題として扱い，SVM(Support Vector Machine) で判定している．SVM で係るか係らないかを識別関数で判別するために，d 次の多項式 kernel 関数が用いられている．d 次の多項式 kernel は，d 個の言語的素性の組み合わせを考慮した学習モデルとして使われている．素性や学習の詳細は参考文献 [6] を確認していただきたい．識別関数内で用いられる値によって，SVM は文節 b_i が文節 b_j が係るかどうか判定を行う．本論文では，この識別関数内の値を確率化させるために，式 (9) で表されるロジスティック関数を用いる．

$$f(x) = \frac{\exp(x)}{1 + \exp(x)} \tag{9}$$

単語の生起に関する知識 ϕ とトピックに関する知識 θ，トピック z を用いて，n-gram モデルと uni-gram rescaling から，単語列 S が生成されるとし，式 (10) で表す．

$$g(S) = P(S|z, \phi, \theta) \tag{10}$$

本研究で文が生成される確率は，式 (10) に係り受け構造に関する知識 ψ を加えて，定数を C とし，次式で表すものとする．

$$f(S) = P(S, D|z, \phi, \theta, \psi) \tag{11}$$

$$= C \cdot P(D|B, \psi) \cdot P(S|z, \phi, \theta) \tag{12}$$

式 (10) を用いて式 (11) からサンプリングを行うために，棄却サンプリングを行う．

本手法では目的分布を式 (11)，提案分布を式 (10) とし，以下のように式を変形させる．

$$\frac{f(S)}{C \cdot g(S)} = P(D|B, \psi) \tag{13}$$

係り受け構造の知識と，文節列から CaboCha を用いて係り受け関係を決定し，ロジスティック関数により確率化する．[0,1] の一様分布から値を出し，$P(D|B, \psi)$ よりも下回った場合，出力文とする．

*1 http://taku910.github.io/cabocha/

2.4 提案手法による文生成の流れ

以上の確率モデル，手法を用いて，システムに次のように文を生成させる．

ユーザはシステムに文を入力する．入力文を形態素解析器 MeCab[9] を用いて単語に分割する．単語単位に分割された入力文と，あらかじめ単語単位に分割した文書群に LDA を用いて各単語のトピック，入力文のトピックを推定する．LDA は MALLET[*2]を用いて行う．入力文と文書群から，Kneser-Ney スムージングを用いて各単語の tri-gram 確率を算出する．推定された入力文のトピックから，単語の tri-gram 確率を uni-gram rescaling によって調整する．その確率値を用いて，返答文の候補を複数生成する．返答文候補の文節間が係り受け関係になりうる確率を CaboCha とロジスティック関数によって算出し，総積をその文のスコアとして求める．棄却サンプリングによって，[0,1] の一様分布から出した値が，スコア値よりも下回った場合，その文を返答文とし，出力する．

3. 感性評価実験

本モデルで提案する手法と，その他の手法で返答文を生成，比較し，適切な返答文が生成できたか確認をした．出力文の自然性を評価するためにユーザにアンケートを行い，検証した．

3.1 実験条件

1 つの入力文に対して，以下の 4 つの手法で返答文を生成し，それぞれの手法により生成された返答文を，返答文 1，返答文 2，返答文 3，返答文 4 と呼ぶこととする．

手法 1 N-gram モデルを用いて文を出力

手法 2 N-gram モデル+係り受け解析による棄却サンプリングを用いて文を出力

手法 3 N-gram モデル+uni-gram rescaling を用いて文を出力

提案法 N-gram モデル+uni-gram rescaling+係り受け解析による棄却サンプリングを用いて文を出力

アンケート項目は Grice の会話の公準 [10] を元に 3 つを作成した．

質問 1 返答文の文量は適切だ

質問 2 入力文と関連性がある

質問 3 わかりやすい表現だ

この各質問に対して，ユーザに 7 択の評定を行ってもらった．7 択の回答尺度にはリッカート尺度を用いた．また，4 つの返答文のなかから，最も自然だと感じた文を 1 つ選択してもらった．実験参加者は日本語を母語とする日本人大学生 11 人である．本研究で生成する返答文は，会話内で発話される文を想定するため，使用するコーパスは新聞のような文体ではなく日常的な文が用いられている文書を用いる．コーパスにはエッセイ，大学生の日記などが書かれた web 小説を利用した．システムは tri-gram 確率で単語を生起させて文を構成した．句点，疑問符，感嘆符を単語生起の終了条件とした．また，生起する単語の最大数は 100 単語までとした．使用した web 小説は 52 本，総単語数 58713 単語，総異なり語数 6908 単語である．LDA を用いて推定した各 topic の単語の一部を表 1 に示す．

*2 http://mallet.cs.umass.edu/

3.2 感性評価実験の結果と考察

入力文に対する各返答文についてのアンケート結果を示す．アンケート結果全てに対して，分散分析と多重比較による検定を行った．ここでは，入力文「節約しないと今月の電気代がヤバイ」に対する各返答文のセットを表 2 に示す．uni-gram rescaling では，表 1 の topic2 の単語がより出現しやすいように調整された．

「返答文の文量は適切である」という質問についての得点を図 2 に示す．アンケート結果では，返答文 1 と返答文 2 の間，返答文 3 と返答文 4 の間に 5%の有意差が見られた．返答文 2 と返答文 4 は係り受け解析によって選ばれた文である．この結果から，係り受け解析による非文ではない文の選択は，係り受け解析を行わない場合に比べユーザにとって適切な文量の文を出力するできることがわかる．文の長さを決める際に，係り受け解析による文の選定が有効な手段の一つと考えることができる．

「入力文と関連性がある」という質問についての得点を図 3 に示す．このアンケート結果では，返答文 4 とその他の返答文で，有意差を確認した．返答文 1 と返答文 2 では，uni-gram rescal-

表 1: 推定された各トピックにおいて出力確率の高い単語

topic	単語
0	甘い 選手 食べ 好き ケーキ 思う 物 オールドファッション よりしっとり 派 日本人 昔 戦 さん だ お菓子 相手 あなた
1	の は に て が た で と も し な か だ ない から い する 私 う
2	発電 エコ 袋 事 所 電力 電気 レジ バック 万 生活 ｋｗ 原子力 買っ ガス 会社 何故 あり 不足
3	洗濯 十 だ 年 時間 物 出来 壊れ 続け 家電 コーヒー 健康 人間 ゲーム 叩い 体 ボク 手 回
4	大学 目標 こと 私 雨 という 行動 生活 など 大学生 描写 天気 サークル 勉強 情景 思い バイト 内容 意見
5	車 夢 音 下ア 私 ながら 買い物 非常 我が家 あと 大丈夫 モノ 見る だから まして 家族 こう 現実 高級
6	を いる ない ある 人 という こと もの それ だろ ば う 的 たい 思う など たり 自分 まで
7	流星 観測 群 座 日食 夜空 皆既 時間 地球 瞬間 場所 太陽 天体 空 です 年 私 獅子 かなり
8	企業 月 就活 選考 インターン 応募 こと みたい セミナー 活動 なる 大学 面接 とか 時期 僕 思い まあ 学生
9	曲 パフューム ライブ 音楽 彼女 達 声 私 ちゃん イメージ 男性 かけ 女性 cm 中田 すぎる とにかく 妄想 あ
10	右翼 左翼 的 や 思う 立場 経済 自由 言葉 に関して と 均質 れる マスコミ 原発 求め 使わ 定義 意味
11	携帯 日本 今 電話 iphone 世界 ある 機能 しかし コミュニケーション 普及 さらに 関係 気付か 絵文字 化 若者 国際 いずれ
⋮	⋮

表 2: 入力文「節約しないと今月の電気代がヤバイ」に対する各返答文

手法	返答文の内容
手法 1	電気系統の復活できない人の島てなくて，大暴れでレモンは初めて子どもとりあえず，徹底的に水分を飛ばすから，わざわざ来てる
手法 2	電気使いまくっていない車を運転している人で飲んだり
手法 3	電気使いまくって製品扉を開け所へ行っ発電所へ行っ電線製作金臭いカミサマ電力会社自身野菜ジュース飲んだり
提案法	電気使い体感省エネ原発問題ありません

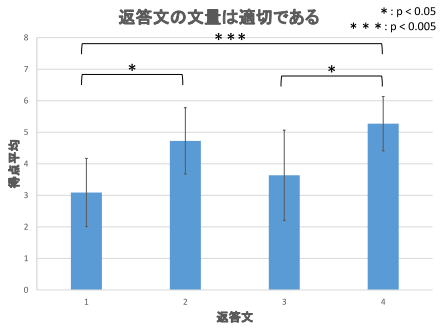

図 2: 質問 1 に対する各返答文の得点

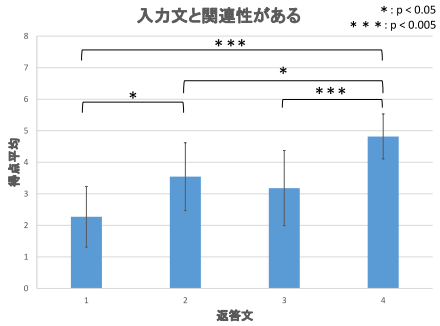

図 3: 質問 2 に対する各返答文の得点

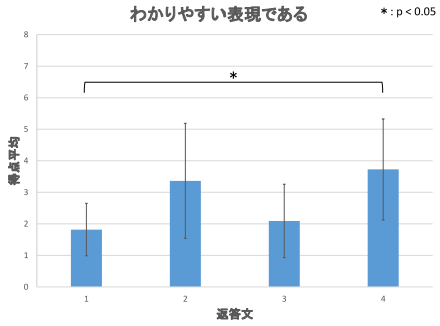

図 4: 質問 3 に対する各返答文の得点

ing を行っていない．ユーザにとって uni-gram rescaling を行った返答文 4 のほうが，より入力文の内容を反映していると感じられたことがわかる．しかし，同じ uni-gram rescaling を行ってる返答文 3 は，返答文 1，返答文 2 の間に有意差が見られなかった．返答文 3 には「電気」「発電所」「電線」「電力会社」といった,「電気」に関連する単語が多く見られるが，入力文の関連性の評価は高い平均値を出さなかった．

「わかりやすい表現である」という質問についての得点を図 4 に示す．このアンケート結果では，返答文 4 と返答文 1 の間に有意差が見られた．係り受け解析を用いた棄却サンプリングを行なっていない返答文 1 では,「4:どちらかというとそう思う」以上の回答は無かった．また，返答文 1 では，返答文 4 と比べ，文が長くなる傾向が見られた．また，入力文のトピックを反映せず，様々な単語が生起した文となっている．それに対し返答文 4 は，どういったことを表しているのかは一見するとわかりづらいが,「電気」「省エネ」「原発問題」といった関連するであろう単語と，適度な文量によって，返答文 1 と比べわかりやすい表現と判断されたのではないかと考えられる．また，被験者 11 人に，この入力文に対する返答文として自然であると感じた文を一つ選択してもらった．返答文 1，2 は 0 人，返答文 3 は 1 人，返答文 4 は 10 人といった結果になった．

4. おわりに

本稿では会話中の発話要素を n-gram モデル，トピックモデル，係り受け解析として，統合的に組み合わせ，入力文に対する返答文生成を行うシステムを構築した．また，システムが生成した文とその他の手法で生成した文をユーザに提示し，アンケートで感性評価実験を行った．入力文に対する返答文として他の手法と比較した結果，本研究の手法が多くの支持をアンケートで集めた．生成された文は，他の手法と比べ有意差は確認できたものの，高い評価値を得ることはできなかった．また，uni-gram rescaling を行った文は，助動詞の生起確率よりもそのトピックに属する単語の生起確率が上回ってしまい，「行っ発電所」など，動詞に不自然な箇所も多く見られた．助動詞の生起確率を調整する必要があり，この点については今後の課題である．

参考文献

[1] 奥村学, 磯崎秀樹, 東中竜一郎, 永田昌明, 加藤恒昭. 質問応答システム. コロナ社, 2009.

[2] 石崎雅人, 伝康晴. 言語と計算-3 談話と対話. 東京大学出版社, 2001.

[3] R Kneser and H Ney. Improved backing-off for m-gram language modeling. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, Vol. 1, pp. 181–184. IEEE, 1995.

[4] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, Vol. 3, pp. 993–1022, 2003.

[5] Daniel Gildea and Thomas Hofmann. Topic-based language models using EM. In *Proceedings of the 6th European Conference on Speech Communication and Technology (EUROSPEECH-99)*, pp. 2167–2170, 1999.

[6] 工藤拓, 松本裕治. チャンキングの段階適用による日本語係り受け解析. 情報処理学会論文誌, Vol. 43, No. 6, pp. 1834–1842, 2002.

[7] 関根聡, 内元清貴, 井佐原均. 文末から解析する統計的係り受けアルゴリズム. 自然言語処理, Vol. 6, No. 3, pp. 59–73, 1999.

[8] 内元清貴, 関根聡, 井佐原均. Me による日本語係り受け解析. 情報処理学会研究報告. 自然言語処理研究会報告, Vol. 98, No. 99, pp. 31–38, 1998.

[9] 工藤拓. Mecab: Yet another part-of-speech and morphological analyzer. *chasen.aist-nara.ac.jp*, 2011.

[10] P.Grice. *Studies in the way of wards*. Harvard University Press, 1989.