2D3-OS-12b-5

発話における応答部・主導部の推定とそれらを構成する単語の推定 ベイズ階層言語モデルを用いて

辻　勇一朗 *1　岡　夏樹 *1　尾関　基行 *1　荒木　雅弘 *1　深田　智 *1　長井 隆行 *2
中村 友昭 *2　大森 隆司 *3

*1 京都工芸繊維大学　大学院工芸科学研究科　*2 電気通信大学　情報理工学研究科
*3 玉川大学工学部

Dialogue consists response part and led part. For the agent to learn how natural mixed initiative interaction, it is necessary to estimate the response unit-driven unit. In this study, it is possible to apply the unsupervised morphological analysis technology, to estimate the break of the response part and led part in the speech. Assuming the dependence on the response part, led unit, by utilizing their bigram Through the Gibbs sampling, to estimate the correct cut.

1. 導入

1.1 研究背景

近年，より自然な音声対話システムを構築する研究が盛んに行われている．Apple の iPhone に搭載されている Siri や，NTT ドコモのスマートフォンのサービスである「しゃべってコンシェル」などは，その性能や有用性などが広く一般に認識されるものとなっている．これらの音声対話システムは，違和感の無い自然な対話を実現しているが，それらは以下のような手法を組み合わせることによって実現されていると考えられる．

1. 手作業に依る対話ルールの逐次記述
2. タグ付けされたデータを使用する教師あり機械学習

当然ではあるが，これらの手法はどちらも非常に時間や人手を要する．また複数人で作業を行う場合は統一した “基準” を設ける必要があるがそれを 100%順守することは，人が作業を行う限り不可能であり，何を基準にするかさえ定まっていないことが多い．

1 については，自動化するのは難しい．2 についても，対話におけるタグ（何をタグとするか定まったものはない）を自動で付加する試みはなされていない．

しかし，例えば形態素解析の分野では，教師なしで形態素のタグ付けを行う研究がなされている．[1]

そこで本研究では，自然な対話における発話は「応答部」と「主導部」から構成されていると仮定し，それらを教師なし学習によって「応答部」「主導部」のタグ付けを推定することを考えた．

2. 応答部と主導部

2.1 応答部と主導部

人が人と対話をする時，少なくとも一方の人間は，もう一方の相手に何らかの意図を伝達したいという目的を持っている．例えば “窓を開けて欲しい” という要求を伝達する場合には「窓を開けて頂けますか」などと発話する．これを発話における “主導” と呼ぶ．しゃべってコンシェルなどの音声対話システムは，利用者の主導の意図を如何にして抽出し，それに適合する主導部を生成する手法に焦点を当てて研究がなされている．

しかし，自然な対話とは主導のやり取りだけで構成されているわけではない．“相手の主導の意図を正しく認識した” という意図を伝える “応答” が存在する．例えば,「うんうん」,「そうですね」,「確かに」，などの言葉や，定型句・挨拶などがこれに相当する．人と人の円滑な対話は応答と主導を適切に組み合わせることによって実現されていると考えられる．これらの組み合わせを学習することが出来れば，システムと人との対話をより円滑に進められることが期待される．

ところで，発話を構成する単語は応答か主導のどちらか一方のみに属しているわけではないことに注意されたい．たとえば対話の最初の「おはようございます」は何かに対する応答ではないため主導であるが，それに対する「おはようございます」は応答部である，と考えられる．応答・主導とはあくまで発話の構成要素の分類という意味で使用し，単語の分類ではない．

また以降，発話における応答の役割を果たす単語群を “応答部”，主導の役割を果たす単語群を “主導部” と呼ぶ．

2.2 依存関係

対話中の発話における応答部と主導部には，複雑な依存関係が存在する．例えば「おはようございます」という主導部に対して「こんばんは」という応答部は適切ではない．また，天気を尋ねる主導部に対して自分の年齢を答える主導部は不適切である．このように，一つの対話に対しても文脈や時勢など様々な依存関係が複雑に絡み合っている．

3. 目的

本論文では，問題を簡略化するため次のような仮定を置いた．

仮定 1 応答部と主導部は形態素単位で構成される

仮定 2 一発話は，応答部のみ，主導部のみ，
または（応答部+主導部）の形で構成される

仮定 3 応答部は直前の発話の主導部のみに依存する

この仮定のもと，与えられた対話コーパス中の各発話に対して応答部と主導部を教師なし学習によって抽出することを目的とする．

Contact: 辻勇一朗，京都市伏見区淀木津町 254-601，080-3858-0197，Y.tsuji.1729@gmail.com

4. 手法

対話データの構造を，図 1 に示す．

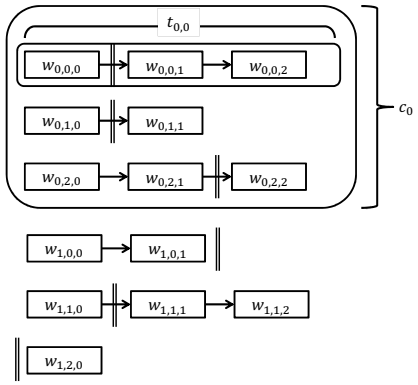

Figure 1: 対話データの構造

対話データは複数の対話（conversation）から成り立つ．一つの対話は二人の人物による発話（talk）が交互に並ぶものとし，一つの発話は複数の単語（word）から成る．

また各発話には応答部と主導部の区切れ目（sp,Separate point）が存在し，sp の直前までの発話を応答部，sp 以降の部分を主導部とする．例えば，sp=0 は応答部が存在しないことを意味する．各発話から応答部と主導部を抽出することは，この sp の位置を推定することと同値である．

以下に sp の推定方法を示す．

1. 各発話の sp をランダムに決定する
2. 各発話の主導部から次の発話の応答部への bigram をつくる
3. ランダムに発話を一つ選ぶ
4. bigram を用いてその発話の sp の尤度を計算し P_o とする
5. 同じ発話について，新しい sp をランダムに決める
6. 新しい sp の尤度を計算し P_n とする
7. P_o と P_n の比に基づいて sp を決定する
8. bigram を更新する

手順 2 について詳しく説明する．発話における主導部が単語 $L_1, \cdots, L_m$，次の発話における応答部が単語 $R_1, \cdots, R_n$ から構成されている時，全ての組み合わせ $(L_1 \rightarrow R_1), (L_1 \rightarrow R_2), \cdots, (L_m R_{n-1}), (L_m \rightarrow R_n)$ についてのカウントを bigram は保持している．例えば，主導部「 おはよう ございます」に対して応答部「 おはよう ございます」を bigram に追加すると，「おはよう → おはよう」，「ます → ございます」などのカウントが追加される．

手順 3 から手順 8 までを十分な回数繰り返すことで，全ての発話に対して尤もらしい sp を推定することが出来る．c 番目の対話の t 番目の発話における sp の尤度の算出方法を式（1）に示す．

$$P(sp=k) = P_{LR} * P_{LR+} * f(k,\lambda) \quad (1)$$

$$P_{LR} = \frac{\sum_{m \in R_{t-1}} \sum_{n=0}^{k-1} c(w_{c,t,n} \mid w_{c,t-1,m})}{\sum_{l \in t_{c,t}} c(w_{c,t,l})} \quad (2)$$

$$P_{LR+} = \frac{\sum_{m \in L_t} \sum_{n=0}^{k-1} c(w_{c,t+1,n} \mid w_{c,t,m})}{\sum_{l \in t_{c,t+1}} c(w_{c,t+1,l})} \quad (3)$$

$$f(k,\lambda) = \lambda e^{-\lambda k} (k \leq 0) \quad (4)$$

式 (2) は，作成した bigram に基づく最尤推定の式である．分子は，注目する sp の属する発話の応答部とその直前の発話の主導部の，全ての単語の組み合わせのカウントの合計である．また分母は応答部の全ての単語のカウントの合計である．式 (3) は式 (2) と同様のことを一つ先の対話の組で行っている．

式 (4) は指数分布の式である．これは sp が発話の前に位置することが多いために付加した補正項で，その平均値 λ は 1 とした

5. 実験

5.1 実験データ

本実験では対話例として，韓国 SBS 放送の日本語講座用テキストに収録されている日本語対話を用いた．[2] ただし，本実験は音声対話を想定しており，その音声認識が正しく行われたものを入力で用いることを想定し，データは全て平仮名で記してある．総対話数 59，総発話数 354，総単語数 4031 である．図 2 に用いたデータの一例を挙げる．“_” は単語の境界，“|” は応答部と主導部の境界である．“|” より左側が応答部，右側が主導部を表す．単語の境界は，既存の形態素解析器である “MeCab”[3] を使用した．応答部と主導部の境界は筆者が手作業で区分した．

|せんしゅう_みすてりー_つあー_に_いっ_た_ん_です_へえ_|みすてりー_つあー_って_どこ_に_いく_か_わから_ない_ん_だ_ろう_いきさき_も_わから_ない_の_に_よく_いく_ねえ_
|おべんとう_が_つい_て_にせんきゅうひゃくはちじゅう_えん_は_やすい_でしょ_う_だから_
そう_だ_ね_|それ_に_どこ_へ_いっ_て_も_じょん_さん_に_とって_は_はじめて_の_ところ_だよ_ね_
そう_なん_です_|すごく_けしき_が_よかっ_た_です_よ_
|どこ_へ_いっ_た_の_
|はるなこ_を_はじめ_おてら_や_はなばたけ_など_に_いっ_た_ん_です_
それ_は_よかっ_た_です_ね_|

Figure 2: 対話データの例

5.2 評価方法

過去に発話を応答部と主導部に分ける研究はされていないので，今回は次の 3 つの手法で sp を決定し，その精度を比較

する．なお，精度の算出方法は，各発話に対して現在の sp と正解の sp が一致していれば 1 点，不一致なら 0 点として，合計得点を総発話数で割ったものとする．

1. sp=0
2. sp=1
3. sp=2
4. sp=3
5. 初期状態
6. 提案手法

上から順に説明する．$sp = k(k = 0, 1, 2, 3)$ は sp の値を k であると決め打ちした時の精度である．これは正解データにおいて sp がどのように分布しているかを示している．初期状態は，サンプリングを実行する前の状態で精度を算出したものである．ただし，この実験における初期状態は，sp を各発話においてランダムに配置したものとする．提案手法は，サンプリングを行った後に精度を算出したものである．

また初期状態・提案手法の精度は，試行を 10 回繰り返したものの平均値とする．

6. 結果

各手法による精度を，表 1 に示す．

Table 1: sp の推定法とその精度の比較

手法	精度
sp=0	0.7203
sp=1	0.1441
sp=2	0.0311
sp=3	0.0847
初期状態	0.1879
提案手法	0.5071

初期状態の精度の分散は 0.05，提案手法では 0.001 であった．初期状態と提案手法の精度に対し両側 t 検定を行った結果，有意水準 0.5%で有意な差がみられた．(t=-4.27,p<0.005)

また，10 回の試行で応答部におけるカウントが高かった上位 10 個の単語を表 2 に挙げる．

7. 考察

実験結果から，初期状態の 19%と比べて提案手法では 50%と精度の向上が見られた．しかし，これだけでは提案手法が有効な手法であるとは考えられない．sp=0 での精度が 70%を超えていることを考慮すると，sp を 0 に近づけるだけで精度の向上を見込めるからである．

また，表 2 に挙げられた単語は，そのほとんどが助動詞や副詞であり，応答部として想定していた「 いい です ね」，「 ええっ」,「 うーん」などの単語は含まれていなかった．この原因として，次の二つが考えられる．

- bigram の作成方法
- sp の尤度の計算方法

Table 2: 応答部としてのカウントが高かった単語とそのカウント

の	639
も	629
だ	587
も	550
ない	507
ん	475
み	455
ひと	403
で	325
だけ	309

上から順に説明する．4. で述べたとおり，bigram にはある発話の応答部とその直前の発話の主導部の，全ての組み合わせが考慮される．

しかし，表 1 を見ても sp=0 である，つまり発話に応答部が存在しない確率は非常に高い．ある発話に応答部が無かった場合，その発話に存在する全ての単語が主導部となり，次の発話の応答部と依存関係を持つ．つまりこれは，ある単語の応答部らしさは直前の発話の主導部の長さに依存してしまうということである．

また対話データにおける発話長は 1 単語から最長 39 単語と振れ幅が大きい．即ち，こちらの意図しない形で単語の応答部らしさが影響を受けていたために，表 2 のような結果になったものと考えられる．

また，今回の実験では簡略のため sp の尤度を最尤推定にて求めている．しかし現在では Kneser-Ney スムージング [4] や NPYLM[1] など，より精度の高いスムージング手法が数多く提案されており，それらを採用することでより精度の向上が望めるのではないかと考えられる．

8. まとめ

対話は主導部とそれに対する応答部から構成される。エージェントが自然な混合主導対話のしかたを学習するためには、応答部・主導部の推定が必要となる。本研究では、教師なし形態素解析技術を応用することで、発話における応答部・主導部の切れ目を推定する。応答部・主導部に対して依存関係を仮定し、それらの bigram を利用してギブスサンプリングすることで、適切な切れ目を推定した．その結果，初期状態より精度を向上させることは出来たが，それにより提案手法が有効であるとの結論は得られなかった．今後は別の依存関係を考慮したり，最尤推定に対してスムージングを行うことによって，精度の向上を行っていきたい．

References

[1] 持橋大地, 山田武士, 上田修功. ベイズ階層言語モデルによる教師なし形態素解析. 情報処理学会研究報告 2009-NL-190, 2009.

[2] “日本事情で学ぶ日本語（Learn Japanese）” http://www.japanese-nihongo.com/monthly/

[3] Taku Kudo. MeCab: Yet Another Part-of-Speech and Morphological Analyzer http://mecab.sourceforge.net/.

[4] Reinhard Kneser and Hermann Ney. Improved backingoff for m-gram language modeling. In Proceedings of ICASSP, volume 1, pages 181?184, 1995.