

バージン許容音声対話における LSM による許容発話範囲の拡張

松山 匡子 駒谷 和範 高橋 徹 尾形 哲也 奥乃 博

京都大学大学院 情報学研究科 知能情報学専攻

1. はじめに

音声認識結果が必ずしも信頼できないような雑音のある実環境下においても、頑健にユーザの意図を解釈するような新しい対話戦略が必要である。我々は、音声認識結果に加えてユーザのバージン（割り込み）タイミング情報を利用し、列挙型音声対話システムにおける頑健なユーザの指示対象理解を行っている [1]。

本稿は以下の2点について報告する。

1. LSM による解釈可能発話の拡張
2. タイミングの一般性の検証

列挙型対話では、ユーザは関連語を用いることがある。我々は、Latent Semantic Mapping (LSM) [2] を用いて、ユーザが用いる関連語に対処する。ここで関連語を、列挙項目に含まれない、列挙項目に関連する単語と定義する。関連語を含むユーザ発話に対しても列挙項目との距離を算出し、指示対象を同定する枠組が必要である。次に、列挙タスクが異なる場合の、ユーザのバージンタイミングの傾向を検証するため、新たに収集した発話データに対してタイミングの傾向を調査したので報告する。

2. LSM による解釈可能発話の拡張

2.1 タイミングと音声認識結果による解釈

我々はユーザ発話 U で指示された項目 T の同定問題を、確率 $P(T_i|U)$ を最大にする T_i を求める問題として定式化している [1]。この概略を図 1 に示す。

$$\begin{aligned} T &= \operatorname{argmax}_{T_i} P(T_i|U) = \operatorname{argmax}_{T_i} \frac{P(U|T_i)P(T_i)}{P(U)} \\ &= \operatorname{argmax}_{T_i} P(U|T_i) \end{aligned} \quad (1)$$

ここで事前確率 $P(T_i)$ は等確率であると仮定している。 $P(U|T_i)$ は、タイミングを用いて解釈する場合 C_1 と音声認識結果を用いて解釈する場合 C_2 の両方から算出されるとし、式 (2) で表す。 α は $P(U|T_i, C_1)$ と $P(U|T_i, C_2)$ のスケールの調整のために導入している。

$$P(U|T_i) = (1 - \alpha)P(U|T_i, C_1) + \alpha P(U|T_i, C_2) \quad (2)$$

タイミングによる解釈 $P(U|T_i, C_1)$ は、項目 T_i に対し、ユーザ発話 U がタイミング t_i で生起する確率で表し、式 (5) を用いて $P(U|T_i, C_1) = P(t_i|T_i, C_1) = f(t_i)$ とする。ユーザの発話タイミング t_i を、ユーザの発話開始時刻と各項目のシステム発話開始時刻との差と定義する。

音声認識結果による解釈 $P(U|T_i, C_2)$ は、ユーザ発話 U の音声認識結果 X と各列挙項目 T_i とのコサイン距離で

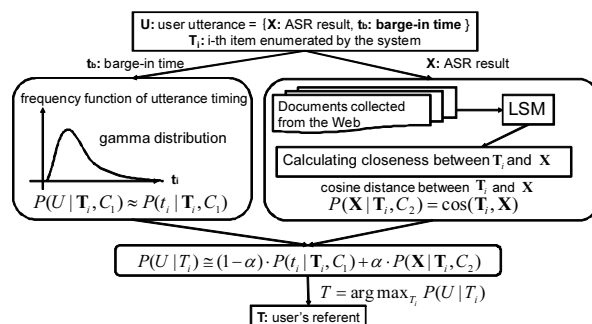


図 1: 指示対象同定手法の概略

表す。 T_i, X はそれぞれ M 次元ベクトルとする。 M はシステムの列挙項目中に含まれる単語数であり、 T_i の要素は単語の TF-IDF 値である。

$$P(U|T_i, C_2) = \cos(T_i, X) \quad (3)$$

2.2 LSM を用いた音声認識結果による解釈

[1] ではベクトル要素に関連語が含まれないため、関連語を用いて指示するユーザ発話と列挙項目との距離は算出できない。我々は、関連語に対して音声認識結果による正しい解釈を得るために、Wikipedia[‡] から関連文書を収集し、語彙を拡張する。次に収集文書に含まれるノイズの影響を取り除くために LSM を適用する。具体的な手順を以下に述べる。

まず、システムが列挙する項目ごとに関連文書を収集する。列挙項目に対応する収集文書 d_i は、Wikipedia から収集した文書に、列挙項目を 1000 文足しあわせた文書である。これは、Wikipedia から収集した文書中の単語に比べてもとの列挙項目に含まれる単語の出現頻度を大きくするためである。つぎに、収集文書に対する単語の頻度をもとに $M \times N$ 共起行列 W を作成する。ここで M は全文書に含まれる総単語数、 N はシステムが列挙する項目数である。共起行列 W の (m, n) 成分 $w_{m,n}$ は以下式で求める。

$$w_{m,n} = (1 - \varepsilon_m) \frac{\kappa_{m,n}}{\lambda_n} \quad (4)$$

ここで、 $\kappa_{m,n}$ は文書 d_n に現れる単語 r_m の出現回数、 λ_n は文書 d_n に含まれる単語数である。また ε_m は文書全体における単語 r_m のエントロピーである。列挙項目ベクトル T_i は共起行列 W の i 列目のベクトルに対応する。

次に、作成した共起行列に対して特異値分解と次元縮約を行ない、共起行列の階数を k に減じる。この次元縮約により、収集文書のノイズの影響を除去する。ユーザ発話の認識結果ベクトル X の要素は、認識結果に含まれる単語 r_m の単語信頼度とする。これは、音声認識結果が誤りである可能性を指示対象同定に反映させるためである。

[‡]<http://ja.wikipedia.org/>

Extending Acceptable Utterances by Using LSM in Barge-in-able Spoken Dialogue Systems: Kyoko Matsuyama, Kazunori Komatani, Toru Takahashi, Tetsuya Ogata, and Hiroshi G. Okuno (Kyoto Univ.)

表 1: 指示対象同定精度 [%]

条件	参照表現 (#:263)	内容表現 (#:137)	合計 (#:400)
(1) 従来手法	85.2	57.7	75.8
(2) LSM による拡張	85.9	29.9	66.8

3. 指示対象同定精度の評価実験

3.1 評価データ及び実験条件

評価には、被験者 20 名による対象を指定する発話 400 発話を用いた。システムは RSS フィード中のニュースタイトルを列挙する。発話データは、「それ」「今の」などタイミングによる解釈が必要な参照表現 263 発話、「サッカーのニュース教えて」「二番目の」など音声認識結果による解釈が必要な内容表現 137 発話である。

(1) 従来手法と (2) LSM による拡張後の指示対象同定精度を比較する。(1) は、システムの列挙項目に含まれる単語のみを用いて解釈を行う。音声認識時には各 RSS フィード中の列挙項目と CIAIR[§]の対話コーパスからそれぞれ作成した統計的言語モデルを用いた。各 RSS フィードに対して、語彙サイズは平均 5,835, $M = 173.5$ である。内容表現の単語正解精度は 45.3% である。(2) の認識には各 RSS フィードの列挙項目に対して収集した文書から作成した統計的言語モデルを用いた。この文書は Wikipedia から収集した文書と「そのニュースを聞かせてください」などのコマンド発話 115 文を足したものである。文書を収集する際の検索キーワードは、列挙項目に含まれる単語を用いて手動で設定した。各 RSS フィードに対して語彙サイズおよび M は平均 17,253 語, $N = 15.8$, また $k = N - 2$ である。内容表現の単語正解精度は 31.9% である。単語正解精度が低いのは、ユーザ発話の分離による歪みが原因である。いずれも音声認識器には Julius[¶]を用いた。また式 (2) の α は予備実験により 0.6 とした。

3.2 指示対象同定精度の評価と今後の課題

指示対象同定精度を表 1 に示す。特に (2) の内容表現の同定精度が低下しているが、これは語彙拡張による認識誤りが原因である。内容表現の単語正解精度は (1) と比べて 13.4 ポイント低下しており、誤認識された単語を含む項目が誤って指示対象と同定される場合が散見された。

一方、LSM による拡張により関連語に対処できることが確認できた。システムの“ソフトバンク対西武、プロ野球アジアシリーズ...”, “岡田監督会見...”, ..., という列挙項目に対してユーザが前者を指して“(試合結果の)ダイジェスト”と関連語を用いた発話例がある。この発話は (1) では“ダイジェスト”という関連語を扱えないため同定できなかったが、(2) では同定可能になった。今後、LSM による同定精度を向上させるために、ユーザが関連語を用いた場合にのみ LSM による処理を導入する。例えば、従来手法と LSM による解釈を同時に行い、尤度差を比較してどちらかの解釈を選択する方法が考えられる。

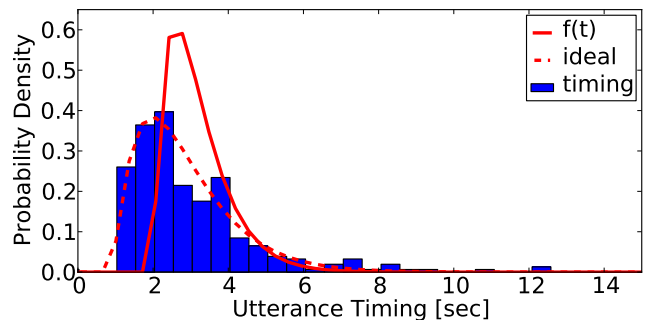


図 2: クイズタスクにおけるユーザの発話タイミング

4. ユーザのバージン発話タイミングの傾向

我々は新たにクイズ形式の列挙型対話においてユーザのバージン 1,184 発話を収集し、列挙タスクが異なる場合の発話タイミングに関して以下の二つを行う。システムが項目を列挙する際の項目間のポーズ長は 1.0 秒, 1.5 秒, 2.0 秒の 3 種類としてデータを集めた。

1. 発話タイミングの傾向の調査
2. [1] のガンマ分布のパラメータ設定時の仮定の検証

図 2 はシステムのポーズ長 1.0 秒, 列挙項目平均長 2.0 秒, 306 発話の指示項目に対する発話タイミングのヒストグラムである。ヒストグラムは 0.5 秒区間毎の発話回数を、全発話数と区間秒で割って正規化している。図 2 より、ユーザは指示項目に対して 2.0 秒付近で発話することが多く、その後発話が増加する傾向がみられる。ポーズ長が異なる残りの 878 発話についても同様の傾向がみられたため図は割愛する。列挙型対話においてはタスクにかかわらず、ピークと減衰が存在する傾向が確認できる。

発話タイミングを以下のガンマ分布で近似している。

$$f(t_i) = \frac{1}{(\rho - 1)! \sigma^\rho} (t_i - \mu)^{\rho-1} e^{-(t_i - \mu)/\sigma} \quad (5)$$

式 (2) の $P(U|T_i, C_1)$ を得るために、[1] では式 (5) のパラメータはあらかじめシステムの列挙項目長とポーズ区間長から決めている。具体的には $\sigma = 2.0$, μ は平均項目長、分布の減衰速度を示す ρ はポーズ長と平均項目長に比例する (比例定数 0.2) と仮定している。これらの仮定から図 2 のヒストグラムに相当するガンマ分布のパラメータは $\sigma = 2.0$, $\mu = 2.0$, $\rho = 0.6$ となり、図の赤線で示される。理想的なガンマ分布のパラメータは $\sigma = 2.3$, $\mu = 0.8$, $\rho = 0.8$ であり、赤点線で示している。 σ と ρ はあらかじめ設定した値と大差なく、おおよそのピーク位置と減衰速度が近似できている。 μ はユーザが項目を聞いてからそれを指定するまでの時間差であるので、各単語長や文節長で設定するなど動的な設計が必要である。これらのガンマ分布のパラメータの設定方法を再検討し、タイミングによる解釈の精度の向上を目指す。

謝辞 本研究の一部は、科研費、GCOE の支援を受けた。

参考文献

- [1] 松山他. バージン発話タイミングを導入した指示対象同定. 音声言語処理研究会, Vol. 2009-SLP-76, No. 4, pp. 1-7, 2009.
- [2] J. R. Bellegarda. Latent semantic mapping. *IEEE Signal Processing Magazine*, Vol. 22, No. 5, pp. 70-80, 2005.

[§] <http://db.ciair.coe.nagoya-u.ac.jp/>

[¶] <http://julius.sourceforge.jp/>