ユーザ発話の揺れに頑健な音声対話システムのための条件付確率場を用いた発話意図の推定

大田健紘
諏訪東京理科大学システム工学部電子システム工学科
〒 391-0292 長野県茅野市豊平 5000 番地 1
E-mail：otakenko@rs.suwa.tus.ac.jp

1 はじめに

現在，インターネットの普及により，膨大な量の情報にアクセスすることができる．ユーザが明確な目的をもって検索を行う場合は，情報量が膨大であったとしても，google や yahoo のようなキーワードをベースとした検索システムが有効に働く．しかし，漠然とした問題を解決する際（例えば，近所のお勧めの観光地を知りたいなど），キーワードでは表現することが困難であり，また個人の嗜好にも依存するためキーワードをベースとした検索システムでは，目的の情報にたどり着くために多大な時間を要する．

漠然とした問題を解決する際は，対話をしながら要望を聞き出し，問題を明確にする手順を踏む必要がある．実際，我々がこのような問題に直面した場合，専門家と話し合いをしながら問題を解決しようとする．コンピュータにも人間のエージェントと同じ機能を持たせることができれば，便利な情報検索が実現できる．

音声認識率の向上により，音声対話による検索システムも実用化されつつある [1–5]．しかし，キーワードによる関係データベースからの検索が主流であり，自然言語による大規模なテキストからの検索システムはまだ少ない．キーワードによる検索の場合，決定的な文法を基にした音声認識が主に利用され，システムが理解できる発話が大きく制限される．自然言語による検索の場合は，多様な表現を認識できる必要があり，統計的言語モデルに基づく大語彙連続音声認識を用いるほうがよい．しかし，大語彙連続音声認識を用いる場合，決定的な文法に基づく音声認識と比較して，音声認識率が低下する問題がある．さらには自然言語による発話には，省略，倒置，あいまいさが存在する．そのため，検索システムはそれらに対して頑健に動作する必要がある．

これまでに音声対話システムとして，たけまるくん [1]，レストラン検索システム [2] や京都版ダイアログナビ [3] などがある．たけまるくんは，一問一答型のシステムであり，想定質問とそれに対する応答が用意されている．音声認識結果と想定質問とのマッチングには形態素の一致数を用いている．

レストラン検索システムも想定質問を用いているが，レストラン名や場所などがクラスとして表現されており，想定質問とのマッチングには構文情報も考慮して行われている．

京都版ダイアログナビは，音声認識結果の N-best を用いて検索を行い，検索されたテキストの重なりの程度を計測し，重なりが小さい部分は音声認識誤りの可能性が高いと考え，確認対話を生成し，あいまい性の解消を図っている．

本稿でも，想定質問をあらかじめ用意し，音声認識結果とマッチングをとる手法を採用する．ただし，想定質問はユーザの発話意図を推定する手がかりとして用いるだけであり，想定質問をそのまま検索には用いない．想定質問と音声認識結果のマッチングは，想定質問から学習した条件付確率場 (CRF：Conditional Random Field)[6] を用いて音声認識結果の各形態素に発話意図を表わすラベルを付ける問題として考える．発話中の各形態素に付与されたラベルから，ユーザの発話意図の候補を生成し，音声認識誤りの検出やあいまい性の解消のために確認対話を生成する．これにより，ユーザ発話のあいまい性，表現の揺れなどに頑健に動作する音声対話システムが期待できる．

2 条件付確率場について

条件付確率場 (CRFs： Conditional Random Fields) は，生成モデルに基づく隠れマルコフモデル (HMMs： Hidden Markov Models) とは異なり，識別モデルに基づくモデルである．生成モデルに基づく HMM は，観測系列 $\mathbf{x}$ とラベル系列 $\mathbf{y}$ の同時確率 $p(\mathbf{x},\mathbf{y})$ を，ある状態から観測値が出力される確率と，状態の遷移確率に分解してモデル化している．

$$\begin{aligned} p(\mathbf{x},\mathbf{y}) &= p(\mathbf{x}|\mathbf{y})p(\mathbf{y}) \\ &= \Pi_{t=1}^{T} p(x_t|y_t)p(y_t|y_{t-1}) \end{aligned}$$

ただし，HMM は，変数間の独立性を仮定しており，独立でない素性をうまく扱うことができない．

一方で，識別モデルに基づく CRF は，観測系列 $\mathbf{x}$ により条件付けされたラベル系列 $\mathbf{y}$ の条件付確率 $p(\mathbf{y}|\mathbf{x})$ を，ロジスティック回帰を基にローカルな変数

間の関係をラベル間の遷移（遷移素性）によりモデル化している．

$$p(\mathbf{y}|\mathbf{x}) = \frac{\exp(<\mathbf{\Theta}, \mathbf{\Phi}(\mathbf{x}, \mathbf{y})>)}{\Sigma_{\tilde{\mathbf{y}}} \exp(<\mathbf{\Theta}, \mathbf{\Phi}(\mathbf{x}, \tilde{\mathbf{y}})>)} = \frac{\exp(\Sigma_{t=1}^{T} <\mathbf{\Theta}, \mathbf{\Phi}(\mathbf{x}, y_t, y_{t+1})>)}{\Sigma_{\tilde{\mathbf{y}}} \exp(\Sigma_{\tau=1}^{T} <\mathbf{\Theta}, \mathbf{\Phi}(\mathbf{x}, \tilde{y}_\tau, \tilde{y}_{\tau+1})>)}$$

ただし，$\mathbf{\Phi}(\cdot)$ は素性関数，$\mathbf{\Theta}$ は素性に対する重み，そして $<\cdot,\cdot>$ は内積を表す．

CRF は HMM のように素性間の独立性を仮定しておらず，柔軟な素性設計が可能であるため，素性間の独立性を仮定できない問題（自然言語処理の多くの問題）に対して HMM よりも有効である．

CRF と同じく識別モデルに基づくモデルとして最大エントロピーマルコフモデル (MEMM : Maximum Entropy Markov Model)[7] がある．MEMM は入力の先頭から順に問題を解いていくため，CRF とは異なり入力全体の情報を利用することができない．そのため，MEMM はラベルバイアス問題 [6] が発生する．一方で，CRF はラベルバイアス問題を解消できるため，MEMM よりも有効である．

識別パラメータの学習に関して，SVM やパーセプトロンからのアプローチもある．これらのメリットは，CRF がラティスの全パスの和を計算するのに対して，最大のコストを与えるパスの探索により近似している点である．そのため，CRF では計算時間の問題で解くことができない問題（異なる言語間でのフレーズの対応付け問題 [8]）などを解くことができる．

3 提案法概要

3.1 学習素性

素性は Cabocha[9] により解析された結果を用いて付与した．具体的には，出現形，品詞，文節境界および主辞，機能語である．さらには，固有名（例えば，諏訪東京理科大学のように場所を表わす単語には<location>など）や「どこ<where>」,「何<what>」などの疑問詞も素性として用いている．ただし，これらの素性を「Bag of Words モデル」として扱うのではなく，前後に出現するものも素性として含める．次節で述べるが，提案法では 2 段階の CRF を用いている．それぞれの段階で用いた素性数は，24765 および 50730 である．

3.2 CRF による発話意図の推定方法

想定するシステムの構成を Fig. 1 に示す．システムではあらかじめ想定質問を用意しておき，想定質問を CRF によりモデル化したもの基に，検索要求からユーザの発話意図を大雑把に推定する．さらには，マッチした想定質問と検索要求から，音声認識誤り箇

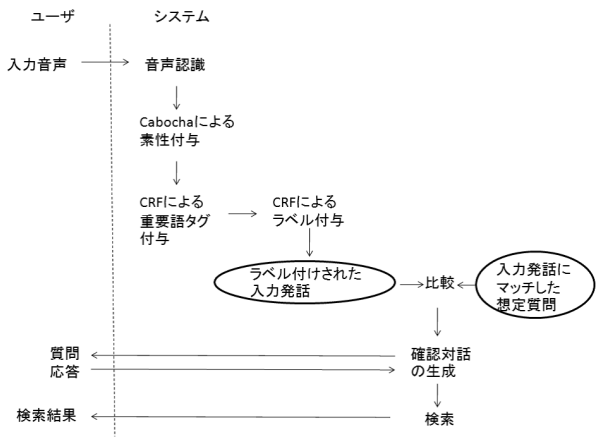

Fig. 1　システムの構成

所の推定および，検索のために不足している情報を推定することが期待される．これを，確認対話を生成するための手がかりとして用いる．

想定質問はまず，Cabocha により解析され，前節で述べた素性を付与する．例を Fig. 2 に示す．構文情報の “1” は文節境界を，“2” は主辞を表わしている．ただし，主辞と文節境界の位置が一致した場合は “2” を付与している．ラベルは似た意味を表わす発話には同じラベルが付与されている．例えば,「<location>への行き方」を尋ねる発話にはすべて同じラベルが付与されている．現在，ラベルは全部で 10 個である．想定質問の一部抜粋を Fig. 3 に示す．ただし，想定質問はクラスを展開している．

次に検索要求に対しても同様に Cabocha を用いて解析を行い，素性を付与する．ただし，固有名や「どこ<where>」などの疑問詞に付与する素性は Cabocha だけでは付与することができないため，2 段階に CRF を適用する．CRF にはフリーのツール CRF++[10] を用いた．まず，1 段目の CRF では，出現形，品詞，文節境界および主辞，固有名や「どこ<where>」などの疑問詞に対してラベルを付与する．そして 2 段目の CRF で，入力文がどういう意図で発話されたのかを想定質問から学習された CRF モデルにより推定する．学習データに対しては手作業でラベルを付与する必要があるが，テストデータに対しては完全に自動的にラベルを付与することが可能となる．

通常，CRF は形態素解析などのラベル付与問題に適用され，分類問題には SVM[11] やブースティング [12] が一般的に用いられる．しかし，本研究では発話意図の推定を各形態素へのラベル付与の問題と考え，CRF を用いた．これにより，発話途中での意図の転換や曖昧な表現へ対応が可能になると考えられる．

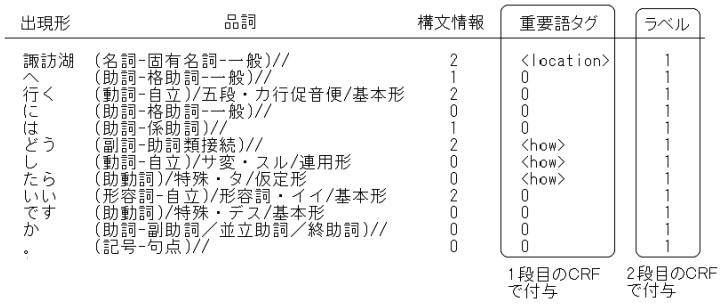

Fig. 2　想定質問に付与した素性の例

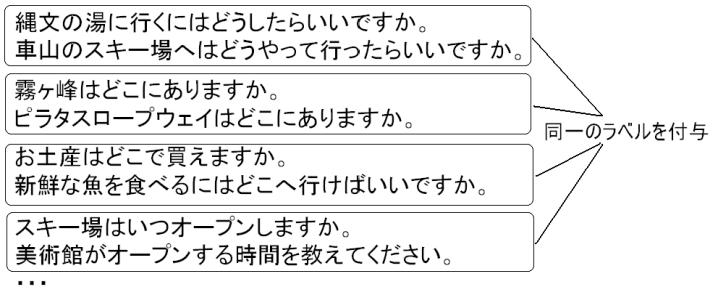

Fig. 3　想定質問の一部

4　評価

4.1　音声認識率について

評価に用いた言語モデルは，国立国語研究所により公開されている「日本語話し言葉コーパス」(CSJ : Corpus of Spontaneous Japanese)[13] をベースに，Web 上から収集した諏訪圏の観光情報のデータを追加することにより作成した．ただし，CSJ データからフィラーや言い直しを含む文および，低頻度の語彙は除外した．音響モデルは CSJ のものをそのまま用いた．音声認識エンジンには Julius[14] を用いた．発話単位での音声認識率は 27 発話をテストデータとして用いて，約 81.5%になった．ただし，テストに用いたデータには未登録語は含まれない．

4.2　重要語タグとラベルの推定結果について

段階的に用いた CRF のそれぞれの性能を調べた．1 段目で正しく推定できたのは形態素単位で約 94%，発話単位では約 70.4%であった．2 段目で正しく推定できたのは約 92.6%であった．推定結果の一例を Fig. 4 に示す．Figure 4 は，本来「白樺湖には何がありますか。」と認識されるべきであるが,「白樺湖には何か言いますか。」と誤った例である．音声認識誤りを含んでいるが，ラベル “4”（想定質問として,「車山にはどんなものがありますか。」や「茅野市には何があるか教えてください。」などを含むラベル）と正しく推定されている．

推定誤りについて分析を行ったところ，1 段目については重要語タグの付与の仕方に問題があったと考えられる．例えば，スキー場のように場所を表わす単語には<location>というタグが付与され，ATM のような物を表わすタグには<mono>というタグを付与することを想定している．しかし，これらの単語に対して形態素解析器により付与される品詞は共に「名詞-一般」である．そのため，これらのタグの付与が困難になったものと考えられる．

2 段目については，音声認識誤りと重要語タグの付与の誤りが同時に起こる場合に最終的なラベル推定を誤る傾向がみられた．ただ，このようにラベル推定に誤っている場合は，推定候補の第一候補であったとしても，CRF++に計算される周辺確率が極端に低くなるため，ラベル推定に失敗していることを検出できるものと考えられる．

5　考察

発話意図推定に対する音声認識誤りや未登録語の影響について述べる．Figure 5 に未登録語を含む発話に対してラベルを付与した例を示す．音声認識誤りにより，出現形が異なる単語に変化したとしても，品詞や係り受けの情報を利用できるため，正しくラベル付与ができた．このことは異なる問題ではあるが [15] においても，CRF の有効性として述べられている．

次に，音声認識誤りが情報検索に与える影響について述べる．この発話は「海遊館はどこにありますか。」というものであるが，音声認識誤りにより読み

出現形	品詞	構文情報	重要語タグ	ラベル
白樺湖	(名詞-固有名詞--一般)//	2	<location>	4/0.766293
に	(助詞-格助詞--一般)//	0	0	4/0.789750
は	(助詞-係助詞)//	1	0	4/0.788994
何	(名詞-代名詞--一般)//	2	<what>	4/0.769862
か	(助詞-副助詞／並立助詞／終助詞)//	1	0	4/0.736743
言い	(動詞-自立)/五段・ワ行促音便/連用形	2	0	4/0.692395
ます	(助動詞)/特殊・マス/基本形	0	0	4/0.639101
か	(助詞-副助詞／並立助詞／終助詞)//	0	0	4/0.573596
。	(記号-句点)//	0	0	4/0.532982

Fig. 4　ラベル推定の結果

出現形	品詞	構文情報	重要語タグ	ラベル
皆	(名詞-代名詞--一般)//	2	<what>	2/0.332239
言う	(動詞-自立)/五段・ワ行促音便/基本形	2	<what>	2/0.413390
館	(名詞--一般)//	2	<mono>	2/0.720865
は	(助詞-係助詞)//	1	0	2/0.867201
どこ	(名詞-代名詞--一般)//	2	<where>	2/0.945051
に	(助詞-格助詞--一般)//	1	<where>	2/0.963058
あり	(動詞-自立)/五段・ラ行/連用形	2	0	2/0.968133
ます	(助動詞)/特殊・マス/基本形	0	0	2/0.964782
か	(助詞-副助詞／並立助詞／終助詞)//	0	0	2/0.956576
。	(記号-句点)//	0	0	2/0.917364

Fig. 5　未登録語に対するラベル推定の一例

は同じであるが表記は異なるものになった．このような場合に，表記の情報だけでなく読みの情報も用いて検索を行うことにより，未登録語を含んでいたとしても検索が可能になるので，検索への影響は小さいものと考えられる．

6　まとめと今後の課題

本稿では，条件付確率場を用いた発話意図の推定について述べた．小規模なデータでの評価ではあるが，音声認識誤りや未登録語などに対して頑健に動作することが示された．今後は，大規模なコーパスを用意し，検証を進める．

また，素性やタグの付与の仕方により結果は変わるので，最適なものを探す必要がある．

条件付確率場については，セミマルコフ CRF[16] や隠れ CRF[17] などが提案されており，これらの導入も検討する．

さらには，用いている素性ベクトルの次元数が大きいため，確率的潜在構造解析 (Probabilistic Latent Semantic Analysis : PLSA)[18] などを用いて次元圧縮の効果も検討する．

現状では，音声対話システムとして完成していないが，京都版ダイアログナビのように情報推薦ができるシステムを目指している．今後我々が取り入れるべき機能を示す．

- 未登録語と音声認識誤りを含む検索要求に対する頑健な検索
- 曖昧な質問，例えば「お勧めのレストランを教えて」などを実現する検索
- 人間のエージェントが知識や経験を積むことにより成長していくのと同じように，対話履歴や，過去の推薦結果の採用，不採用の履歴から学習していくシステム

1 つ目の機能については，検索を行う際に出現形のみを用いて行うのではなく，読みや品詞などの情報を用いて検索を行い，検索されたテキスト群を用いて音声認識辞書および言語モデルを適応的に再構成し，音声認識を行うことで実現できるものと考えている．2 つ目の機能については，ブログなど Web 上の情報から評判情報を抽出する技術 [19] の応用により実現できるものと考えている．3 つ目の機能については，能動学習や半教師有り学習などのアプローチを取り入れる．

参考文献

[1] 西村 竜一, 西原 洋平, 鶴身 玲典, 李 晃伸, 猿渡 洋, 鹿野 清宏: “実環境研究プラットホームとしての音声情報案内システムの運用”, 電子情報通信学会論文誌, Vol.J87-D-II, No.3, pp.789-798, 2004.

[2] 駒谷 和範, 河原 達也, 清田 陽司, 黒橋 禎夫, Pascale Fung: “柔軟な言語モデルとマッチングを用いた音声によるレストラン検索システム”, 情報処理学会研究報告, SLP-39-30, 2001.

[3] 翠 輝久, 河原 達也, 正司 哲朗, 美濃 導彦: “質問応答・情報推薦機能を備えた音声による情報案内システム”, 情報処理学会論文誌, Vol.48, No.12, pp.3602-3611, 2007.

[4] Heather Pon-Barry, Fuliang Weng and Sebastian Varges: “Evaluation of Content Presentation, Strategies for an In-car Spoken Dialogue”, ICSLP2006, pp. 1930-1933, 2006.
[5] Antonio Roque, Anton Leuski, Vivek Rangarajan, Susan Robinson, Ashish Vaswani, Shri Narayanan, David Traum: “Radiobot-CFF: A Spoken Dialogue System for Military Training”, ICSLP2006, 2006.
[6] J. Lafferty, A. McCallum, and F. Pereira: “Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data”, Proc. of ICML, pp.282-289, 2001.
[7] Andrew Mccallum, Dayne Freitag, Fernando Pereira: “Maximum Entropy Markov Models for Information Extraction and Segmentation”, ICML 2000, pp591-598, 2000.
[8] Ben Tasker, Simon Lascoste-Julien and Dan Klein: “A Discriminative Matching Approach to Word Alignment”, Proc. of HLT-EMNLP, pp. 73-80, 2005.
[9] http://chasen.org/~taku/software/cabocha/
[10] http://crfpp.sourceforge.net/
[11] Vladimir N. Vapnik: “The Nature of Statistical Learning Theory”, Springer, 1995.
[12] Yoav Freund: “Boosting a weak learning algorithm by majority”, Information and Computation, 121(2):256-285, 1995.
[13] http://www.kokken.go.jp/katsudo/seika/corpus/
[14] http://julius.sourceforge.jp/index.php
[15] 齋藤 邦子, 鈴木 潤, 今村 賢治: “CRF を用いたブログからの固有表現抽出”, 言語処理学会第 13 回年次大会発表論文集, D1-3, 2007.
[16] Sunita Sarawagi and William W. Cohen: “Semi-Markov conditional random fields for information extraction”, In Advances in Neural Information Processing Systems 17, pp.1185-1192, 2004.
[17] S. B. Wang, A. Quattoni, L.-P. Morency and D. Demirdjian: “Hidden conditional random fields for gesture recognition”, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.1521-1527 2006.
[18] T. Hofmann: “Unsupervised learning by probabilistic latent semantic analysis”, In Proc. of Uncertainty in Artificial Intelligence, 1999.
[19] 杉木 健二, 松原 茂樹: “消費者の意見に基づく商品検索”, 情報処理学会論文誌, Vol.49, No.7, pp.2598-2603, 2008.