3U-7

音声対話システムにおける WFST に基づく文法検証を利用した動的ヘルプ生成

福林 雄一朗　　駒谷 和範　　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

我々はこれまで，初心者でも使いやすい音声対話システムを目指して動的なヘルプを生成する手法を研究してきた [1]．動的ヘルプ生成における課題は，ユーザの発話からユーザのシステムに対する知識状態を正しく推定することである．しかし，ヘルプが必要な初心者がタスク遂行に適切な表現を知っているとは限らず，システムは想定外発話を誤認識してユーザの知識状態を誤って推定してしまう可能性がある．我々はこれまでに，システムの音声認識器と別の言語モデルを持つ音声認識器との音響尤度差を利用して，システムの想定する文法との近さを判定する手法を提案した [2]．しかし，尤度差だけでは特徴量として不十分であり本方法での判別の精度は高くなかった．さらに，認識結果は考慮していないので文法構造を反映した推定もできなかった．

本研究では，Weighted Finite State Transducer (WFST) を利用してシステムの想定する文法とユーザ発話の近さを計算する手法を開発した．本手法では，WFST を利用することで文法構造まで考慮して想定文法との近さを計算できる．評価実験では，本手法がユーザ知識推定の高精度化に利用できることを示した．

2. WFST に基づく文法検証

本研究では，音声認識結果を入力すると，システムが想定しているすべての文法との近さを累積重みの形で計算し，最終的に最も近い (=累積重みが大きい) 文法カテゴリを出力する WFST を目的とする．さらに，目的とする WFST ではその文法と音声認識結果のマッチングの結果を出力する．本研究では，重みづけを工夫することでそうした WFST を実現する．WFST に対する重みづけは，**受理単語に対する重みづけ**，**文法間違いに対する重みづけ**，**コンセプトに対する重みづけ**の 3 種に分類される．以下で述べる重みづけでは，想定文法に近ければ正の重みを，遠ければ負の重みを与えるように設計することで，発話が文法内か文法外であるかを累積重みの正負で判別できる．文法検証は，音声認識結果の N-best 候補それぞれに対して累積重みを計算し，それらの結果を統合して行う．以下ではそれぞれの重みづけ手法を説明した後，重みの計算例と文法検証の方法を説明する．

2.1 受理単語に対する重みづけ

受理単語に対する重みづけは，WFST により受理された単語に対する報酬であると考え，音声認識結果の単語レベルで信頼できる単語に対してより大きな重みを与える．

$$w_w = l(w_{asr})CM(w_{asr})$$

ここで，w_{asr} は音声認識結果として得られた単語，$l(w_{asr})$ は w_{asr} の長さ，$CM(w_{asr})$ は w_{asr} に対する信頼度 [3] である．$l(w_{asr})$ はモーラ数に比例する値で，語彙中で最も長い単語の長さで正規化するので，$0 < l(w_{asr}) \leq 1$ である．この重みづけは，長くかつ信頼できる出力列を優先するための設計である．

2.2 文法間違いに対する重みづけ

文法間違いは，想定文法中の単語が別の単語に置き換わる**置換 (SUB)**，必要な単語が抜け落ちる**脱落 (DEL)**，余分な単語が入り込む**挿入 (INS)** の 3 つに分類される．文法間違いに対する重みは，ペナルティと考え負の値を設定する．一般的には，WFST の出力列に含まれる文法間違いの部分が短かくなるように設定する．

$$\begin{aligned} w_{sub} &= -(CM(w_{asr})l(w_{asr}) + l(w_g))/2 \\ w_{del} &= -(\overline{l(w)} + l(w_g))/2 \\ w_{ins} &= -(CM(w_{asr})l(w_{asr}) + \overline{l(w)})/2 \end{aligned}$$

ここで，w_g は想定文法の対応する単語 (以下，想定単語)，$\overline{l(w)}$ は全語彙の $l(w)$ の平均である．w_{sub} は，入力単語長と信頼度の積と想定単語の長さの平均によるペナルティである．このペナルティでは，入力単語と想定単語のそれぞれの長さで想定文法との遠さを表す．w_{del} では w_{sub} における入力単語 w_{asr} に対応するものがないので，$CM(w_{asr})l(w_{asr})$ の代わりに $\overline{l(w)}$ を利用する．一方，w_{ins} では w_{sub} における想定単語 w_g に対応するものがないので，$l(w_g)$ の代わりに $\overline{l(w)}$ を利用する．それぞれの重みづけは，置換・脱落・挿入が起きることで，想定文法とどれだけ遠くなるかを表している．

2.3 コンセプトに対する重みづけ

コンセプトを構成する内容語は，文法において重要な役割を果たすので，受理単語に対する重みとは別に加算する．ただし，音声認識誤りにより得られる内容語は誤受理につながるので，信頼できない内容語にはペナルティが与えられるように設計する．

$$w_c = \sum_{\boldsymbol{W}} l(w_{asr})(CM(w_{asr}) - \theta_c)$$

ここで，$\boldsymbol{W}$ はコンセプトを構成する単語の集合，θ_c ($0 \leq \theta_c \leq 1$) は閾値を表す．この重みづけでは，信頼度が閾値未満の単語には負の重みを与えることで信頼できないコンセプトを棄却できる．

2.4 累積重みの計算と文法検証

文法検証は，以上で示した 3 種類の重みの和である累積重み w を音声認識結果の N-best 候補それぞれに対して計算して行う．

$$w = \sum w_w + \sum (w_{sub} + w_{del} + w_{ins}) + \sum w_c$$

N-best 候補の i 番目の文の累積重みを w^i，WFST の出力する文法カテゴリを g_i とすると，ある発話が文法内であるかどうかは，事後確率 p_i による重みづき和 w_s の正負で判定する．

$$\begin{cases} w_s = \sum_{i=1}^{N} p_i w^i \delta_{g_1,g_i} & > \quad 0 \quad (\text{in} - \text{grammar}) \\ & \leq \quad 0 \quad (\text{out} - \text{of} - \text{grammar}) \end{cases}$$

Dynamic Help Generation by using WFST-based Grammar Verification in Spoken Dialogue Systems: Yuichiro Fukubayashi, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno (Kyoto Univ.)

書き起こし		こーえつじ	の	れんらくさき (未知語) を		おしえて	ください	
音声認識結果	あ	こーえつじ		えと	さきょーく	おしえて	ください	
$CM(w_{asr})$	0.4	1.0	-	0.3	0.3	1.0	1.0	
$l(w_{asr})$	0.045	0.18	-	0.09	0.135	0.18	0.18	
$CM(w_{asr})l(w_{asr})$	0.018	0.18	-	0.027	0.0405	0.18	0.18	
想定文法	-	⟨ 寺社名 ⟩	の	⟨ 項目名 ⟩	を	おしえて	ください	
$l(w_g)$	-	0.24	0.045	0.24	0.045	0.18	0.18	
WFST 出力	INS	こーえつじ	DEL	SUB	SUB	おしえて	ください	
w_w	-	$+0.18$	-	-	-	$+0.18$	$+0.18$	$+0.54$
w_{sub}	-	-	-	-0.134	-0.0428	-	-	-0.176
w_{del}	-	-	-0.143	-	-	-	-	-0.143
w_{ins}	-0.129	-	-	-	-	-	-	-0.129
w_c	-	$+0.18\cdot(1.0-\theta_c)$	-	-	-	-	-	$+0.18-0.18\theta_c$

図 1: 重みの計算例

ここで，δ_{g_1,g_i} は N-best の 1 位と i 位の文の文法カテゴリが同じときに $\delta_{g_1,g_i}=1$，違うときに $\delta_{g_1,g_i}=0$ である．また，N-best 文の i 番目の文の事後確率 p_i は音声認識結果の対数尤度を利用して求める [4]．つまり，w_s は N-best の 1 位と同じ文法カテゴリの文の累積重みの和である．

図 1 に累積重み w の計算例を示す．この例では「れんらくさき (未知語) を」が「えと さきょーく」に誤認識されている．この認識結果を WFST に入力すると，さまざまな出力列が得られるが，累積重みが最も大きかった出力結果を想定文法に近いものとして採用する．その結果，一番近い想定文法は「⟨ 寺社名 ⟩ の ⟨ 項目名 ⟩ を おしえて ください」であると分かる．ただし，⟨⟩ はコンセプトに対応する内容語を表す．このとき WFST の出力を見ると，「⟨ 項目名 ⟩ →えと」，「を→さきょーく」が置換，「の」が脱落，「あ」が挿入，得られるコンセプトは「⟨ 寺社名 ⟩=こーえつじ」なので，それぞれの重みは，$w_w=+0.54, w_{sub}=-0.176, w_{del}=-0.143, w_{ins}=-0.129, w_c=+0.18-0.18\theta_c$ になる．ただし，$\overline{l(w)}=0.24$ とする．このとき，累積重み w は，$w=+0.54-0.176-0.143-0.129+0.18-0.18\theta_c=0.272-0.18\theta_c$ となる．この音声認識結果が N-best の第 1 位で，文法カテゴリが “info”(寺社の情報を調べる) であったとすると，$w_s=\sum_{i=1}^{N} p_i w^i \delta_{\mathrm{info},g_i}$ の値によって文法内・外の判別が行われる．

3. 評価実験

3.1 実験条件

実験では，発話検証により文法内か文法外かをどれだけ正しく判別できたかを調べる．また，WFST が出力する文法カテゴリがどれだけ正しいかも調べる．コンセプトに対する重みにおける閾値 θ_c は，0 から 0.9 まで 0.1 きざみで変化させ最も精度が高くなる値を採用した．

実験では，京都寺社案内システムで収集した 1520 発話を用いた．文法カテゴリは全部で 18 種類ある．文法検証で利用する音声認識器には Julius[‡] を用いた．言語モデルは認識文法から生成した例文 10000 文から作成した統計的言語モデルを利用した [5]．語彙サイズは 595 で平均の単語正解精度は 43.4%であった．

ベースライン手法として，文法内・外の判別には，2 つの音声認識器の音響尤度差を利用する方法 [2] を用いた．基準となる尤度差の閾値 θ は，-200 から 400 まで 10 きざみで調べ，最も精度が高くなる場合の値を採用した．また，文法カテゴリの判別には，文法カテゴリごとに文法ベースの音声認識器を用意し，その認識結果が最尤である文法カテゴリを採用する方法を用いた．文法ベースに認識器には Julian を用いた．

[‡]http://julius.sourceforge.jp/

表 1: 文法内・外の判別精度

	音響尤度差 ($\theta=50$)	本手法 ($\theta_c=0.6$)
判別精度	71.1	80.1

表 2: 文法カテゴリの判別精度

	Julian	本手法 ($\theta_c=0.7$)
判別精度	65.9	72.2

3.2 実験結果

まず，文法内・外の判別の結果を表 1 に示す．本手法では音響尤度差を利用する手法より判別精度が 9.0%向上することが分かる．この結果からユーザ知識の推定がより高精度になることが期待できる．これは，本手法では音声認識結果が想定している文法とどれだけ近いかをより正確に測定できているからだと考えられる．

次に，文法カテゴリの判別の結果を表 2 に示す．本手法では文法カテゴリの判別精度が Julian を用いた単純な手法より 6.3%向上する．判別精度に差が出たのは，ベースライン手法では文法外の発話を，音声認識誤りにより文法内であると判断してしまうからだと考えられる．また，文法内でも未知語などを含む発話に対しては，正しい認識ができないことが原因として考えられる．

以上の結果より，WFST に基づく文法検証がユーザ知識の推定の高精度化手法として期待できる．

4. おわりに

本報告では，音声対話システムの動的ヘルプ生成において，ユーザ知識推定の高精度化手法として，WFST に基づく文法検証を開発した．評価実験では，WFST に基づく文法検証がユーザ知識推定の高精度化手法として期待できる結果が得られた．

謝辞 本研究の一部は，科研費，グローバル COE，SCAT の支援を受けた．

参考文献

[1] Y. Fukubayashi *et al.* Dynamic Help Generation by Estimating User's Mental Model in Spoken Dialogue Systems. In *Proc. ICSLP*, pp. 1946–1949, 2006.

[2] K. Komatani *et al.* Introducing Utterance Verification in Spoken Dialogue System to Improve Dynamic Help Generation for Novice Users. In *SIGdial*, pp. 202–205, 2007.

[3] A. Lee *et al.* Real-time word confidence scoring using local posterior probabilities on tree trellis search. In *Proc. ICASSP*, Vol. 1, pp. 793–796, 2004.

[4] 駒谷, 河原. 音声認識結果の信頼度を用いた効率的な確認・誘導を行う対話管理. 情処学論, Vol. 43, No. 10, pp. 3078–3086, 2002.

[5] Y. Fukubayashi *et al.* Rapid Prototyping of Robust Language Understanding Modules for Spoken Dialogue Systems. In *Proc. IJCNLP*, pp. 210–216, 2008.