LE-008

音声認識結果とコンセプトへの重みづけによる WFST に基づく音声言語理解の高精度化

Improving WFST-based Language Understanding Accuracy by Weighting for ASR Results and Concepts

福林 雄一朗 † 駒谷 和範 † 中野 幹生 ‡
Yuichiro Fukubayashi Kazunori Komatani Mikio Nakano
船越 孝太郎 ‡ 辻野 広司 ‡ 尾形 哲也 † 奥乃 博 †
Kotaro Funakoshi Hiroshi Tsujino Tetsuya Ogata Hiroshi G. Okuno

1. はじめに

音声対話システムの言語理解部は，多様な言語表現を受理可能であることに加えて，新たなドメインでも迅速に構築できることが望ましい．これまで，言語理解部としていくつかの方法が提案されてきた．ユーザの発話をキーワードスポッティングやヒューリスティックなルールで分類する手法 [1] では，複雑なルールの準備には時間や手間がかかり，システム制作者への負担が大きい．また，コーパスを利用してコンセプトの出現確率を学習する手法 [2] や Weighted Finite State Transducer (WFST) を利用した手法 [3] が提案されてきた．しかし，これらの手法は大量の学習データを必要とし，新たなドメイン向けの言語理解部を構築するのは容易ではない．

我々は，WFST に基づく言語理解の新しい手法を開発した．我々の手法では，WFST に対する重みづけを，認識された単語と言語理解結果であるコンセプトの 2 つのレベルで行う．評価実験では，対象とするドメインで適切なパラメータを選択することで言語理解精度が向上することを確認した．このパラメータは，音声認識率に依存して変化するため，我々の手法では当該ドメインでの音声認識率が予測できれば，それに応じて適切なパラメータを選択することで言語理解精度を向上させることができる．

2. WFST に基づく言語理解

WFST に基づく言語理解部では，音声認識結果を入力し，出力として言語理解結果を得る．図 1 はビデオ予約システムの言語理解部の WFST の例である．入力の ε は，入力なしでの遷移が可能であることを表す．この例では，「開始時間 は 10 時 30 分 です」という入力列に対して，“開始時間 は $ 10 時 hour=10 $ 30 分 minute=30 です” という出力列が得られる ($は何も出力されなかった場合を考慮したダミー記号である)．また，“10 時” と “30 分” の遷移時にそれぞれ 1.0 が累積重みに加算され累積重みとして 2.0 が得られる．最終的に言語理解結果として，[hour=10, minute=30] を得る．さらに，任意の入力を受け付ける FILLER 遷移 (図 1 の’F’) を各フレーズ間に挿入することで，「えーと 開始時間 は 10 時 30 分 です」という入力に対しても正しい言語理解結果が得られる．しかしながら，FILLER 遷移を導入すると，WFST 上での遷移が何通りもあるので，多くの出力列が結果として得られ，最適な結果を選ぶ枠組みが必要となる．WFST に基づく言語理解では，多くの出力列から累積重み w が最も大きいものを言語理解結果として採用する．

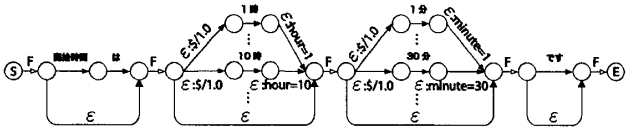

図 1: WFST による言語理解の例

3. 音声認識結果とコンセプトへの重みづけ

我々は WFST に対する重みづけを 2 つのレベルで定義する．音声認識結果に対する重みづけとコンセプトに対する重みづけである．

3.1 受理単語に対する重みづけ

WFST に入力し受理された単語に対して重みづけを行う．この重みづけでは，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．この重みづけ w_w を以下のように設計した．

1. **word (const.)**: $w_w = 1.0$
2. **word (#phone)**: $w_w = l(W)$
3. **word (CM)**: $w_w = CM(W) - \theta_w$

1 は受理された全ての単語に対して一定の重みを加える．この重みづけは，受理単語の数が多い出力を優先するための設計である．2 は，各受理単語の長さを考慮に入れた重みづけである．$l(W)$ $(0 < l(W) \leq 1)$ は，単語 W の音素数に比例した値とする．2 は，入力列の長さを 1 よりも詳細に表現していると言える．さらに，受理単語の信頼度 $CM(W)$ を考慮に入れた 3 も提案する．単語の受理・棄却の閾値として θ_w を設定することで，信頼できない認識結果を棄却できる．この重みづけは，内容語の数が多く，かつ信頼できる出力列を優先するための設計と言える．

3.2 コンセプトに対する重みづけ

単語レベルでの重みに加えて，コンセプトレベルの重み w_c も設計した．コンセプトは，複数の単語から成り，音声認識結果を WFST に入力することで得られる．コンセプトに対する重みは，それぞれのコンセプトに含まれる単語の信頼度などを用いて計算する．

1. **cpt (const.)**: $w_c = 1.0$
2. **cpt (avg)**: $w_c = \dfrac{\sum_{\boldsymbol{W}} (CM(W) - \theta_c)}{\#\boldsymbol{W}}$
3. **cpt (min)**: $w_c = \min_{\boldsymbol{W}} (CM(W) - \theta_c)$
4. **cpt (#pCM(avg))**: $w_c = \dfrac{\sum_{\boldsymbol{W}} (CM(W) \cdot l(W) - \theta_c)}{\#\boldsymbol{W}}$
5. **cpt (#pCM(min))**: $w_c = \min_{\boldsymbol{W}} (CM(W) \cdot l(W) - \theta_c)$

$\boldsymbol{W}$ は当該コンセプトに含まれる単語の集合で，W は $\boldsymbol{W}$ に含まれる単語である．また，$\#\boldsymbol{W}$ は $\boldsymbol{W}$ に含まれる単語の数である．1 は，1 発話から得られるコンセプト

† 京都大学 大学院 情報学研究科 知能情報学専攻
‡(株) ホンダ・リサーチ・インスティチュート・ジャパン

表 2: 最適な重みづけの組み合わせとそのときの言語理解精度 (() 内はベースライン)

	ビデオ予約					レンタカー予約				
	#utt.	w_w	α_c	w_c	LU Acc.	#utt.	w_w	α_c	w_c	LU Acc.
全発話	4186	**const.**	0	n/a	86.4 (86.4)	3364	**CM**-0.0	5.0	**#pCM(avg)**-0.2	78.7 (77.0)
認識率 100% 未満のみ	1297	**CM**-0.1	4.0	**#pCM(avg)**-0.4	51.3 (49.9)	1375	**CM**-0.1	5.0	**#pCM(avg)**-0.2	52.5 (47.4)

表 3: 音声認識率ごとの最適時の言語理解精度 (() 内はベースライン)

	ビデオ予約					レンタカー予約				
ASR Acc.	#utt.	w_w	α_c	w_c	LU Acc.	#utt.	w_w	α_c	w_c	LU Acc.
$-\infty$-10	645	**const.**	5.0	**#pCM(avg)**-0.7	17.2 (8.2)	625	**CM**-0.6	0.5	**#pCM(min)**-0.7	16.8 (1.8)
10-40	74	**const.**	3.0	**#pCM(min)**-0.8	65.8 (35.8)	130	**#phone**	3.0	**#pCM(avg)**-0.1	41.0 (34.7)
40-70	314	**const.**	2.0	**#pCM(min)**-0.7	60.0 (59.5)	395	**CM**-0.1	1.0	**avg**-0.7	68.7 (63.1)
70-100	264	**const.**	0	n/a	79.3 (79.3)	225	**#phone**	3.0	**const.**	82.2 (80.4)
100	2889	**const.**	0	n/a	100.0 (100.0)	1989	**const.**	0	n/a	100.0 (100.0)

表 1: 入力と出力の例

入力	いいえ	2 月	22	日	です
出力	*FILLER*	2 月	22	日	です
$CM(W)$	0.3	0.9	1.0	0.9	0.7
$l(W)$	0.3	0.9	0.9	0.6	0.6
コンセプト	-	month=2	day=22		-

が多くなるようにするための重みづけである．また，2 や 3 はコンセプトを構成する単語の認識結果が信頼できないものを棄却するための設定である．4 や 5 は，コンセプトに含まれる単語の信頼度の他にそれらの長さも考慮に入れた重みづけである．どちらもコンセプト部分が長くかつ信頼できる発話を優先するための設定である．受理単語に対する重みと同様に θ_c は，コンセプトの受理・棄却の閾値である．

3.3 累積重みの計算

言語理解部は，以上で示した 2 種類の重み w_w, w_c の重みつき和である累積重み $w = \alpha_w \sum w_w + \alpha_c \sum w_c$ が最も大きい出力列を言語理解結果として選ぶ．

パラメータとして **word (CM), cpt (#pCM(avg))** を選択した時の累積重み w の計算方法を表 1 を用いて説明する．入力が「いいえ 2 月 22 日 です」である場合，受理単語に対する重みの総和は $\alpha_w(3.5-4\theta_w)$ である．また，"month=2" に対する重み $\alpha_c(0.9 \cdot 0.9 - \theta_c)/1 = \alpha_c(0.81 - \theta_c)$ と "day=22" に対する重み $\alpha_c(1.0 \cdot 0.9 - \theta_c + 0.9 \cdot 0.6 - \theta_c)/2 = \alpha_c(0.72 - \theta_c)$ により，コンセプトに対する重みの総和は $\alpha_c(1.53 - 2\theta_c)$ である．したがって，累積重み w は $\alpha_w(3.5-4\theta_w) + \alpha_c(1.53-2\theta_c)$ となる．

4. 評価実験

4.1 実験条件

3. 章で定義した重みづけを実験的に評価する．実験では，重みづけや各重みの係数 $\alpha_{w,c}$，閾値 $\theta_{w,c}$ を変化させ言語理解精度を比べた．係数 α_w は 1.0 または 0 に固定し，α_c を 0, 0.5, 1.0, 2.0, 3.0, 4.0, 5.0 と変化させた．$\alpha_{w,c} = 0$ は，対応する重みが利用されないことを意味する．また，$\theta_{w,c}$ は 0〜0.9 を 0.1 ごとに変化させた．

実験では，ビデオ予約ドメインの 4186 発話とレンタカー予約ドメインの 3281 発話を用いた．音声認識器は Julius を用いた．言語モデルは，各ドメインの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である．ビデオ予約ドメインの言語モデルの語彙サイズは 209 で，レンタカー予約ドメインの言語モデルの語彙サイズは 226 であった．平均単語認識率はビデオ予約ドメインで 86.3%，レンタカー予約ドメインで 72.3% であった．それぞれのドメインの言語理解の正解は，書き起こしを WFST に入力して作成した．

4.2 音声認識率と最適なパラメータの組み合わせ

入力に対して単純に文法との最長一致をとる言語理解をベースラインとする．これは，重みづけを w_w =**word (const.)** に，$\alpha_c = 0$ の場合，つまりコンセプトへの重みづけを利用しない場合に相当する．

全発話に対して最適な重みづけと係数の組み合わせを求めた結果と，認識率が 100% 未満の発話に対して同様に重みづけと係数を求めた結果を表 2 に示す．どちらのドメインでも全発話に対しては最適時とベースライン時の平均言語理解精度に大きな差はない．一方で，認識率が 100% 未満の発話では最適時とベースライン時の平均言語理解精度に差が生じている．これは，どちらのドメインでも認識率が高い発話が全体の半分以上を占めているからだと考えられる．つまり，音声認識結果が正しい場合は単純に最長一致をとればよいと言える．

そこで，発話データを音声認識率ごとに分類し，それぞれで同様に最適な組み合わせを求めた．そのときの言語理解精度を表 3 に示す．表 3 のクラス 10-40 は，音声認識率が 10% 以上 40% 未満であることを表す．この結果から，音声認識率に応じた適切な重みづけにより言語理解率が向上することが分かる．特に認識率の低い発話での向上が大きい．これは，閾値 θ_w, θ_c を設定することで，信頼できない単語やコンセプトを棄却し，湧き出し誤りを抑制しているからだと考えられる．この結果は，音声認識率など発話の状況に応じた重みづけにより言語理解精度が向上する可能性を示したと言える．

5. おわりに

我々は，音声対話システムにおける WFST を利用した言語理解部を開発した．WFST の重みは，音声認識結果中の単語の音素数や信頼度を利用して計算される．したがって，重みづけが比較的単純であり，新たなドメイン向けの言語理解部の構築が容易である．

謝辞 レンタカー予約のシステムの作成については，北海道大学情報学研究科，伊藤 敏彦氏，永野 由佳氏のご協力を得ました．

参考文献

[1] Stephanie Seneff. TINA: A natural language system for spoken language applications. *Compurational Linguistics*, Vol. 18, No. 1, pp. 61–86, 1992.

[2] Katsuhito Sudoh and Hajime Tsukada. Tightly integrated spoken language understanding using word-to-concept translation. In *Proc. EUROSPEECH*, pp. 429–432, 2005.

[3] Alexandros Potamianos and Hong-Kwang J. Kuo. Statistical recursive finite state machine parsing for speech understanding. In *Proc. ICSLP*, pp. 510–513, 2000.