単語ラティス形式の音声認識結果を対象とした発話意図の認識

浦谷 則好* 谷田 泰郎** 田中 英輝***
*ＡＴＲ音声言語通信研究所 **国際電気通信基礎技術研究所 ***ＮＨＫ放送技術研究所
uratani@slt.atr.co.jp tanida@itl.atr.co.jp tanakah@strl.nhk.or.jp

１．はじめに

言語処理で何らかの形で文脈を利用するためには、一般的に発話を分割し、それに発話意図を示すタグを付与することが必要となる。例えば、音声認識や機械翻訳の精度向上を目的として確率的に発話意図を予測する研究が行われている[1,2]が、これらの研究では発話は適切な単位に前もって分割されていることを前提としている。音声翻訳を対象とした場合、入力発話が適切に分割されていることを前提にすることは現実的でない。そこで、一話者の連続した発声を入力として、その発話を最適に分割し、発話意図を示すタグを付与する研究も行われている[3]。この場合の入力は音声認識結果の第１候補としている。しかし、通常音声認識結果は単語ラティスとして得られるので、このラティスに対して発話の最適分割処理とタグ付与を実行すれば音声認識率の向上も可能かもしれない。本稿ではそういう観点から単語ラティスに対する分割処理とタグ付与を同時に行ないつつ音声認識結果を得る手法を提案する。また、実験結果についても報告する。

２．課題の設定

入力としては一人の話者が継続して発声した一連の発声（ターン）とする。この音声認識結果は図１のような単語ラティスで得られるものとする。この単語ラティスを対象に、本稿の課題は（１）最適な音声認識結果を求めること、（２）最適な数の発話意図(SA:speech act type)単位に分割すること、（３）各SA単位に最適なSAタグを付与することである。ただし、SAタグは前もって２３個に設定されている。（表１）

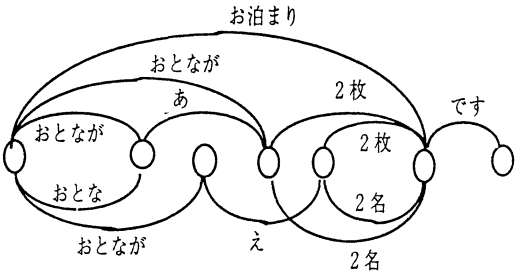

図１． 単語ラティスの例

表１． SAタグ

ACCEPT	承諾
ACKNOWLEDGE	聞いていることの通知
ACTION-REQUEST	動作の要求
APOLOGY	謝罪
BELIEVE	信念
CONF-QUESTION	確認の質問
DESIRE	話者の希望
EXPRESSIVE	最初に発声する間投詞
FAREWELL	別れの挨拶
GREET	会話の最初の挨拶
INFORM	内容の伝達
INFORM-REQUEST	情報の要求
NO	否定的回答
OFFER	申し出
OFFER-FOLLOW-UP	申し出の追加
PERMISSION-REQUEST	許可を求める発話
REJECT	拒絶
TEMPORIZER	途中に発声する間投詞
THANK	感謝
VOCATIVE	呼び掛け（
WH-QUESTION	いつ、何をなどの質問
YES	肯定
YN-QUESTION	内容の真偽に関する質問

文献［3］では、音声認識は正しい認識結果を出力することを前提に（2）（3）を課題としている。しかし、後で示すように音声認識結果が88%程度でも分割精度やタグ付与精度はかなり劣化する。したがって、音声認識スコアだけを用いて、まず音声認識結果を求めてから、（2）（3）を実行したのでは精度のよい分割やタグ付与が期待できない。また、分割可能性確率やタグ付与確率の利用は音声認識精度そのものを向上させる可能性がある。つまり、課題は（1）（2）（3）を同時に満足する手法を発見することである。

3．確率モデルと計算法

2章で提起した問題を解くために導入した確率モデルは基本的に文献[3]と同じものである。ただし、入力は形態素列Wではなく、候補単語ラティスVである。また、最適な形態素列W（つまり音声認識結果）、最適な発話意図単位列U、最適なSAタグ列Tを同時に求めることになるので、式で書くと、

$$(\hat{U},\hat{T},\hat{W}) = \arg\max_{U,T,W} P(U,T,W \mid V)$$
$$= \arg\max_{U,T,W} \prod_{j=1}^{k} P\left(u_j, w_j, t_j \mid h_j, V\right)$$

となる。ここで、u_jはそのSA単位のj番目を示す。h_jはu_jとj番目のt_jの履歴、すなわち$\left(u_1^{j-1}, t_1^{j-1}\right)$を示す。実際には1つ前、すなわち$\left(u_{j-1}, t_{j-1}\right)$だけで履歴を代表させている。SA単位の存在確率とタグ付与確率の計算法は文献[3]と同じものを採用している。

図1の各経路には音声認識スコア（音響スコアと言語スコア）、分割可能性確率（ただしスコア化）、SAタグの付与確率（ただしスコア化）が付与されている。これらは対数をとってあるのでスコアの総和が最大となる経路を見つければよいことになる。しかし、ラティスが大きいためそのままでは計算量が膨大となってしまう。そこで、音声認識スコアの大きいものから5つ（ビーム幅が5）だけを計算対象とした。さらに計算時間の節減のため分割可能性確率が高いノードでだけ計算を実施することとした。探索には前向きＤＰ＋後ろ向きA*のアルゴリズム[4]を用いた。

4．実験

実験に使用したのは９７４個の発声データであり、手作業により正解の形態素列（音声認識結果）の特定、SA単位への適性な分割、適正なSAタグの付与を実施した。正解の形態素列に対してSA単位への分割、SAタグ付与実験を実施したところ、SAタグ付与精度がかなり悪いことを発見した。原因を追求したところアルゴリズム（決定木など）の不備ではなく、手作業によるSAタグ付与のゆらぎであることが判明した。そこで、SAタグ付与基準の見直しを行い、元々３０個あったタグの中で定義があいまいなもの、他との区別が容易でないものを除いて２３個に縮小した。この結果、クローズドテストによる実験では、発話分割、ＳＡタグ付与に関して１００％の精度（適合率、再現率とも）が保たれることを確認した。

実験は用意した図1のようなワードグラフの入力データ９７４発声に対して、１０回の交差確認テストを繰り返すことによって実施した。また、評価対象としたのは出力される1-bestのみである。

実験結果を表2に示す。手法の覧で「1:01」などと書いてあるのは、「音声認識スコア：分割タグ付与スコア（分割スコア＋タグ付与スコア）」の重み（絶対値ではなく変位の大きさに対するもの）を示す。比較のために正解形態素列に対する音声認識精度、分割精度、ＳＡタグ付与精度の値を表2の最上行に示す。この場合は音声認識スコアを使わずにまず音声認識結果を確定し、その後分割タグ付与スコアを用いて分割点とタグを決めたことに相当する。音声認識精度は当然１００％である。分割精度は９５％くらいで、ＳＡタグ付与精度は８５％程度である。つまり、このアルゴリズ

ムではこれが上限（それ以上の精度を期待できない）であることが分かる。なお、交差数とは正解の分割点をまたがる誤分割が行われる回数を示す。正解分割点の個数は約２０００である。正解形態素列に対しての交差数は０であるから、分割ミスのすべては過剰な分割によることも分かる。なお、ＳＡタグ付与精度は「分割もタグも正しい」割合を示したものであるため、必ず分割精度より低い値となる。

表２の「1best認識結果」は音声認識スコアで１位となったものに対して、後で分割タグ付与スコアを用いて分割点とタグを決めたものである。提案する手法（音声認識と分割・タグ付与を同時に行なう方法）はこれより良い結果を期待したものである。表２を見ると残念ながら期待したほどの改善はみられない。「1:0.01」付近でわずかに音声認識精度とタグ付与精度の向上が見られるが、統計的に有意なレベルではない。予想されることではあるが、分割タグ付与スコアの重みを増していくと音声認識精度が下がり、これに伴って分割精度やタグ付与精度も低下している。

５．おわりに

音声認識の結果として得られる単語ラティスに対する分割処理とタグ付与を同時に行ないつつ音声認識結果を得る手法を提案した。単語ラティスを対象としても発話意図単位（SA単位）への分割、各SA単位へのSAタグ付与が可能であることを示すことができた。しかし、実験結果で示したように分割タグ付与精度の向上と音声認識精度の向上に関しては、期待した結果は得られなかった。しかし、まだ実験結果の検討は十分でない。例えばラティスのビーム幅を変えたらどうなるか、分割スコアとタグ付与スコアの寄与比率を変えればどうなるか、などの検討はこれからである。アルゴリズムの改良を進め、実験によって精度向上を図ることができれば改めて報告したい。

参考文献

[1]M. Nagata and T. Morimoto, "An Information-theoretic model of discourse for next utterance type prediction" Transactions of Information Processing Society of Japan, Vol.35, No.6, pp.1050-1061(1994)

[2]N. Reithinger and E. Maier, "Utilizing statistical dialogue act processing in verbmobil" Proceeding of the 33rd Annual Meeting of the ACL, pp.116-121(1995)

[3]田中英輝、横尾昭男：確率モデルによる発話の最適分割と意図認識，言語処理学会第５回年次大会pp.259-262,(1999))

[4]M. Nagata, "A stochastic Japanese morphological analyzer using a forward-DP and backward-A* N-best search algorithm" Proceedings of Coling94, pp.201-207(1994)

表２．　実験結果

	音声認識精度		分割精度		ＳＡタグ付与精度		
手法	再現率	適合率	再現率	適合率	再現率	適合率	交差数
正解形態素	100.00	100.00	94.61	96.17	84.63	86.03	0
1best認識結果	87.76	87.63	61.66	59.02	57.13	54.69	162
1：0.001	87.78	87.66	61.66	58.99	57.19	54.72	155
1：0.01	87.84	87.70	61.66	58.92	57.25	54.71	145
1：0.1	87.29	87.34	61.31	59.18	56.79	54.82	147
1：1	83.23	83.44	49.19	44.49	45.59	41.24	169