Internal Use Only (非公開)

002

TR-IT-0326

音声認識過程での発話分割
Utterance Splitting in Speech Recognition

中嶋 秀治 Hideharu Nakajima
山本 博史 Hirofumi Yamamoto

2000.1.11

自然な話し言葉を使った対話では，１回の発声の中に複数の文が含まれる場合がある．音声理解や翻訳の前には，このような発話を文に分割することが必要となる．本稿では発話分割を音声認識と同時に行なう方法と評価結果について報告する．なお、内容は文献 [7] の説明に若干の修正を加えたものである．

©ＡＴＲ音声翻訳通信研究所

©ATR Interpreting Telecommunications Research Laboratories

もくじ

1 はじめに

ATR 音声翻訳通信研究所では，自然で自発的に発声された会話音声（以後，自然な会話と呼ぶ）を対象とする音声翻訳システムの近未来での実現に資することを目的として，音声言語処理技術の研究を行なっている．現在の音声翻訳システムにおいては，発話ごとに，すなわち音声の認識単位ごとに翻訳を行なう．処理対象である自然な会話においては，１回の発話の中に複数の文が含まれる場合がある．翻訳では，文を単位とした従来の翻訳に関する多くの知見を利用できることから，処理単位を文にするほうが望ましい．そのため，翻訳の前段階において，発話をそれよりも小さな文などの単位に分割することが必要となる [1][2]．従来は，そのような発話の分割処理が音声認識結果の第１位候補のみへの後処理として実現されていた [1][2][3].

本稿では，発話の分割を音声認識と同時に行い，発話の分割位置としての句点を含んだ単語グラフを出力する方式について述べ，旅行会話音声の分割と認識に適用した結果を報告する．

2 複数の文を含む発話

自然な会話では，1回の発話の中に，複数の文が含まれている場合があり，分割が必要となる [2]．ここではそのような発話の例を示す．

例えば，ホテルの予約やサービスの問い合わせに関するホテルの従業員と客との会話を想定して集められたATR自然発話音声言語データベース [4] には，次のような発話がある．この例

例1：複数の文を含む発話
宿泊客：もしもし交通手段についてちょっと教えて頂きたいんですが
ホテル：はいかしこまりましたどちらへお出かけでしょうか
宿泊客：延暦寺にはどう行ったらよろしいでしょうか

図 1: 複数の文を含む発話

では，ホテル側の発話が分割の必要な発話である．このように発話は文という単位にはなっていない．高い翻訳性能を得るためには，ホテル側の発話の文への分割が望ましい．しかし，従来の音声認識では「かしこまりました」のあとに来る筈の句点の認識は考慮されていなかった．

上の例では，「はいかしこまりました。どちらへお出かけでしょうか」または，「はい。かしこまりました。どちらへお出かけでしょうか」のような発話中の句点の位置での発話の分割が可能である．今2通りの分割例を挙げたように，「はい」のうしろに句点を打つ場合もあれば，そうでない場合もあり，話し言葉中の句点の打ち方についての明確な規定はない．また，話し言葉における文の定義も難しい．また，文間の無音区間の長さは様々であり，無音区間に関する物理量のみに基づいて文を定義し，発話を分割することは難しい [2]．そのため，本研究では会話の書き起しデータであるATR自然発話音声言語データベースにおいて，句点で区切られている単位を文と定義する．

従来は，この分割処理が音声認識の後処理 [2][3]，または，翻訳の前処理 [1] として行なわれていた．本研究では，句点を言語情報として扱うことによって，分割を音声認識と同時に行なう．

3 実現手法

ATR の音声認識では統計的言語モデルを用いている [5]．従来は句点の認識は考慮されていなかったため，句点を取り除いて言語モデルが作成されていた．本研究では，それらを学習データの中に残し，発話中の句点への遷移確率，および，発話中の句点からの遷移確率も推定させる．

本研究では，統計的言語モデルとして，多重クラス複合 N-gram[6] を用いる．このモデルはクラス N-gram が基本となっており，次式で単語の予測確率が計算される．

$$P(w_i|w_{i-1}) = P(w_i|C^t_{w_i})P(C^t_{w_i}|C^f_{w_{i-1}}) \tag{1}$$

ここで，w_* は単語，または複合語としての単語系列である．発話中の句点は発話末の句点とは別の単語として登録する．そして，発話中の句点は，C^t_* では発話終了記号と同じクラスとして登録し，C^f_* では発話開始記号と同じクラスとして登録する．

その他の点では多重クラス複合 N-gram[6] と同じ作成方法で，自動クラスタリング，単語系列抽出，パラメータ推定が行なわれる．

次に decoding であるが，発話中の句点が，発話開始記号や発話終了記号とは別の単語として登録されているので，従来の認識システムをそのまま用いることができる．

4 評価

4.1 実験用データ

音声翻訳研究の目的で集められたATR自然発話音声言語データベースを用いて評価実験を行なった．上記のデータベースのうち，分割実験の評価用データとして9会話（通常の2人による会話を，話者の役割（ホテル側／客）毎に区別してそれぞれを「片側会話」と呼ぶことにすると，18片側会話）を，音声認識の評価用データとして42片側会話を選んだ．本稿では，前者を「評価1」，後者を「評価2」と呼ぶことにする．評価1データは文献[2]で用いられたものと同一である．その他のデータと評価1データは言語モデルの学習用のデータとした．ここで，評価2データの話者は音声認識の目的から音響モデルの学習には含まれていない話者である．それぞれの片側会話数，のべ単語数，および，発話中の句点の総数を表1に示す．

表1: 学習用と評価用のデータ

	片側会話数	総単語数	句点数
学習	7,202	1,385,130	32,096
評価1	18	2,437	73
評価2	42	4,990	89

以上のデータを用いて，発話中の句点を含む言語モデル(SPLT)とそれを含まない言語モデル(BASE)の2種類のモデルを作成する．

両モデルにおいて，語彙のサイズは約14,000，獲得された単語系列数はおよそ4,700であり，クラス数は C_*^t ， C_*^f ともに700とした．

4.2 分割の評価

SPLTと評価1データとを用いた音声認識実験を行なった．この認識結果には句点が含まれる．分割については，認識結果の第1位候補での句点の再現率と適合率の観点から評価する．結果は表2の通りであった．表2の中の「評価1’」は発話末の句点（216個）を評価に含めた

表2: 発話分割の再現率と適合率

	再現率	適合率	句点の総数
評価1	78.08	90.47	73
評価1’	94.46	97.84	289

場合の値である．

4.3 分割誤り

評価1データでの分割誤りの事例の幾つかを挙げる.

削除誤り（分割漏れ）には例2のような事例があった.「×」が分割位置であるにも関わらず正しく分割されなかった分割位置である.「申し訳ございません」のような感動詞の後ろ,

例2
削除誤：申し訳ございません × シングルは・・・
削除誤：東京シティーホテル御滞在 × 零三の・・・
削除誤：調べます × しばらくお待ち下さい

図 2: 削除誤

体言止めの後ろ，および，一部の終止形の後ろでの分割ができていない（句点を認識できていない）.

挿入誤り（過分割）には，例3のような事例があった.「※」が誤って挿入された分割位置を示す.

例3
挿入誤：そうですか ※ 料金はそれぞれおいくらなのですか
挿入誤：そうですか ※ じゃバス付の方でお願いしたいのですが

図 3: 挿入誤

終助詞のあとに句点が置かれることは多いが，データベース内の「そうですね」や「そうですか」の後の位置には，句点ではなく読点がおかれており，挿入誤りとなった.

4.4 単語認識率の比較

音声認識結果の第1位候補での単語認識率（%Accuracy）を表3に示す.表3の「SPLT（句点無）」は，BASEと比べるために，認識結果の第1位候補と正解との間で，句点以外の単語を対象としてDPマッチングを行なって得た値である.表3のように，SPLT（句点無）とBASEとの比較によれば，SPLTモデルはBASEと単語の認識性能においてほとんど違いが

表 3: 単語認識率

	評価 1	評価 2
SPLT(句点無)	92.90	85.57
BASE	93.07	85.27

無く，句点を学習することによる性能の劣化は無い．

5 考察

本稿の表2の結果は，統計的な情報だけに基づいて得られた分割の再現率と適合率である．一方，文献[2]では，数値的な分割処理の後に，Heuristicsを用いて評価1のデータに対する分割結果の補正を行なって評価している．また，当時と現在とでは，音声認識の条件も大きく異なる．そのため，文献[2]と本研究との直接の比較は行なえない．しかし，表2の「評価1'」の結果にもあるように，本研究の分割結果は，文献[2]でのテキスト入力（音声認識100%を想定）に対する分割結果とほぼ同等である．また，削除誤り（例2）や湧き出し誤り（例3）の事例は，文献[2]の誤りとほぼ同じであった．従って，本手法用のHeuristicsを作成すれば同等の性能が得られると予想される．

音声認識と言語処理とのインタフェースとしては情報を多く含んだ単語グラフが用いられ始めている．そのため，従来の認識結果の第1位候補のみに対して分割を行なう方法よりも，分割結果を含んだ単語グラフの方が言語処理に多くの情報が伝わる．また，本手法では，モデルのパラメータ推定がN-gramの枠組みで統一されるため，文献[2]での閾値探索が不要になり，モデル構築と管理が容易になる．

6 おわりに

本稿では，発話の分割を音声認識と同時に行い，文境界の記号としての句点を含んだ単語グラフを出力する手法を述べ，ATRデータでの性能の評価を行なった．その結果，句点以外の単語の認識性能を劣化させること無く，発話を分割できることを確認した．今後は分割誤りの扱いについての検討を行なう予定である．

参考文献

[1] O.Furuse et al., Splitting Long or Ill-formed Input for Robust Spoken-language Translation, Proc. of COLING-ACL'98,1998.

[2] 竹澤 他, 発話単位の分割または接合による言語処理単位への変換手法, 自然言語処理, Vol.6,No.2,1999.

[3] B.Reaves et al., ATR-MATRIX:Implementation of a Speech Translation System, 春季音講論,1-6-25,1998.

[4] T.Takezawa et al., Speech and Language Databases for Speech Translation Research in ATR, Proc. of the 1st International Workshop on East-Asian Language Resources and Evaluation(EALREW '98), 1998.

[5] 内藤 他, 旅行会話タスクにおける ATRSPREC の性能評価, 秋季音講論,1999.

[6] H. Yamamoto et al., Multi-class Composite N-gram Based on Connection Direction, Proc.of ICASSP,1999.

[7] 中嶋 他, "発話分割付き実時間音声認識", 秋季音講論,1999.