実時間音声対話システムTOSBURGの開発 (2)音声理解

6N−6

橋本 秀樹* 坪井 宏之** 竹林 洋一***
*東芝ソフトウェアエンジニアリング(株) **(株)東芝 関西研究所 ***(株)東芝 総合研究所

1. はじめに

自然な対話中の連続音声から発話の内容を理解する音声言語システムの研究が進められている[1][2]。このようなシステムでは自由な発話(spontaneous speech)中に出現する、不要語やポーズ、言い直し、省略、環境の雑音などの取扱いが問題となる。

我々は、不特定の利用者が自由に発声できるtask-orientedな実時間対話システムTOSBURG(Task-Oriented System Based on speech Understanding and Response Generation)を構築した。この実現のために、雑音免疫学習に基づくワードスポッティング法[3]を提案し、時間離散的なキーワードラティスの構文意味解析方法を検討してきた。本文では、実時間音声対話システムTOSBURG[4]における、キーワードスポッティングを利用した音声理解方式について報告する。

2. キーワードラティスを用いた対話音声理解

2.1 キーワードの利用

spontaneousな発話には、不要語、言い直し、省略、雑音などの現象が含まれるため、発話は多様となり、この様な現象すべてを文法等の知識に詳細に記述し、実時間で解析・理解する事は困難である。

我々はこの問題を克服するために、限定されたタスクの下で、発話から図1に示すようなキーワードを検出し、構文意味理解するアプローチを採用した。この方法はキーワードに依存するため、文法などの言語的知識が簡潔に表現できるという利点がある。反面、タスクによっては、個々の発話内容の詳細な理解は困難となるが、理解の詳細化および曖昧性の解消は対話を通じて行うことができる。また、単語照合が基本となるため、我々が先に提案した雑音免疫学習に基づくワードスポッティング法を用いることで、発話環境の雑音にロバストな実時間音声理解が可能である。

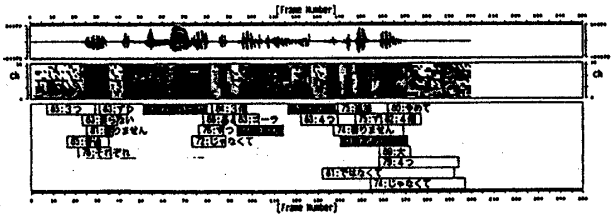

図1. キーワードラティスの例

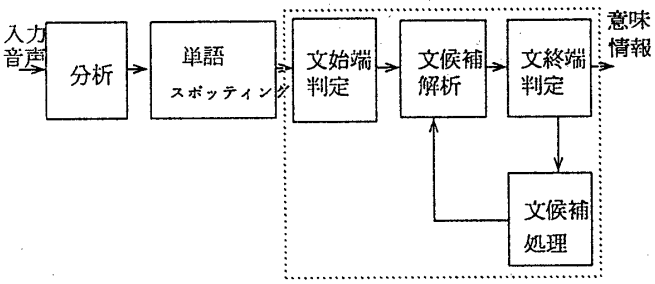

図2. 音声理解システムの構成

2.2 対話音声の理解

TOSRUBGの対話音声理解部は、音声認識部でキーワードがスポッティングされる毎にパーザを起動し、時間離散的なキーワード間の接続条件を考慮しながら、自由な発話を効率良く始終端フリーに解析、発話に対する複数候補の意味表現を出力する。発話の意味候補は対話処理部に渡され、対話の状況および履歴を考慮して、複数の意味候補からその場面で最も妥当な意味表現を評価・選択し、応答を生成して対話を進行させる。

3. 解析処理

対話音声理解部は、拡張LR法[5]をベースとし、時間離散的なキーワードラティスの解析を行なうための機能が補強されている。解析の制約として、文法(CFG)以外に、単語接続可能範囲、word pair grammarを参照する。単語接続可能範囲は、文候補中の単語間の時間的接続関係を規定する制約である。出力としての意味情報はフレーム形式で表現し、文法の拡張項にフレームやスロットの生成などの意味解析手続きを記述、構文解析と同時に意味表現を作成する。音声理解部は、文始端判定部、文候補解析部、文終端判定部、文候補処理部から構成される(図2)。ここでは、その各部の処理を、発声「うーんとハンバーガーとえーコーヒーを下さい」に対する解析例(図3)を用いて説明する(下線部は認識対象単語)。

3.1 文始端判定

入力単語が文の先頭として出現しうる単語であるか否かを、予め文法より作成されたLRテーブルを使って判定する。文頭と判定されたならば、その単語を先頭とした新しい部分文候補を作成し、文候補テーブルに登録する。例えばt1において、W1は文頭となりえず、新文候補は生成されないが、t4において、W4,W5が文頭になりうると判定され、部分文候

Development of Real-Time Speech Dialogue System TOSBURG (2) Speech Understanding
Hideki HASHIMOTO*, Hiroyuki TSUBOI** and Yoichi TAKEBAYASHI***
* Toshiba Software Engineering, ** Toshiba Kansai Research Lab., *** Toshiba R&D Center

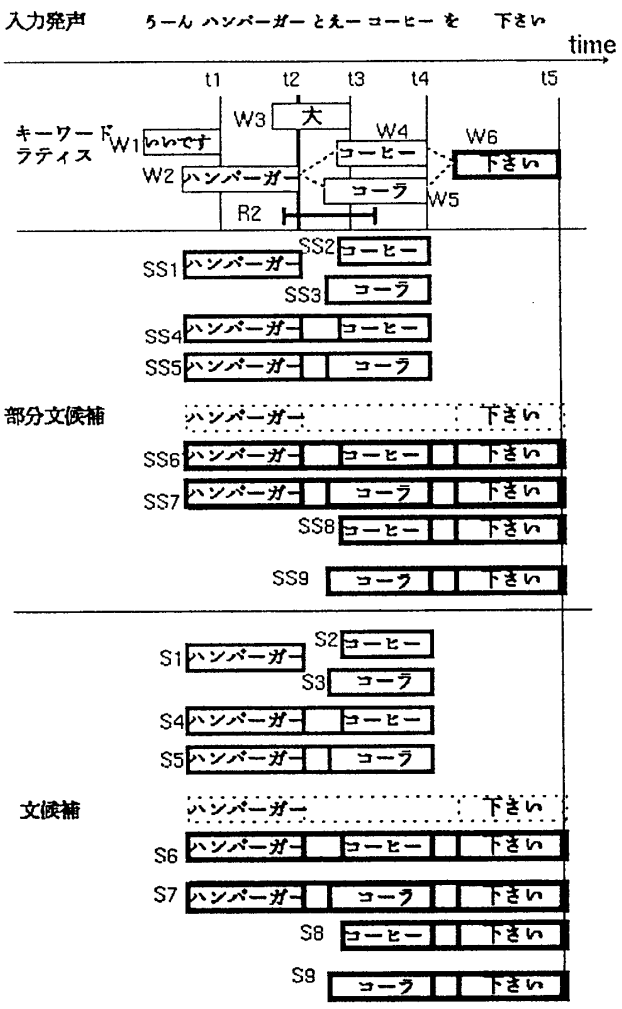

図3. 解析例

補、SS2とSS3として登録される。

3.2 文候補解析

ここでは先ず、部分文候補と入力単語の時間的な接続可能性を検定する。次いで、LRテーブルおよびword pair grammarを参照し、両者が構文的意味的制約を充足しているかを調べる。例えばt4では、入力単語W4,W5の始端点が、それに先行する部分文候補SS1の最後尾の単語の接続可能範囲R2の範囲内にあり時間的文法的に接続可能と判定できるため、SS1の複製とW4,W5が接続され、新たな部分文候補SS4,SS5として文候補テーブルに登録される。この接続範囲の設定により、不要語、雑音、認識対象外の単語、言い淀みなどを除いた部分文候補を生成して解析を進めることができる。スコアは単語の尤度と継続時間長から求めて、部分文候補の生成時にスコアに基づくビーム幅を設定して枝刈りを行う。

3.3 文終端判定と文候補処理

文始端判定部および文候補解析部で新たに生成された部分文候補を受け、文終端判定部は、それらの文としての完結性を見る。例えばt4においては、新部分文候補SS2,SS3,SS4,SS5の全ての部分文候補が文として成立するため、それら文法を完全に充足した文候補S2,S3,S4,S5として出力される。

また、部分文候補の増大により処理量が増加しない様に、全ての部分文候補のうち、最後に文候補解析を行ってからある時間以上経過したものを文候補テーブルから削除する。つまり、次の時点以降に認識される単語候補が時間的に接続し得ない部分文候補を取除く。

3.4 終端同期処理

上述した処理をスポッティングしたキーワードの終端に同期して行うことで、キーワードスポッティングと構文意味解析処理を同時にパイプライン的に進める事を可能とし、音声入力から解析結果を得るまでの時間的な遅れを押えることができる。

4.認識実験

4.1 実験条件

ハンバーガーショップでの注文をタスクとした認識実験を行った。語彙数は49である。単語辞書の訓練データは男女計60名の孤立単語発声と男性5名の文発声を用いた。評価用データは男性4名がそれぞれ発声した350文であり、平均5個のキーワードを含む。文法は81ルールからなる。学習データの文発声のキーワード接続範囲の論理和を求め、評価に用いる接続範囲とした。

4.2認識結果と検討

文認識率は43.4%、単語認識率は82.8%であった。認識誤りは音声中の継続時間の短い単語(ひとつ、ふたつなど)に多い。これは音声中における無声化の影響や発話の息けにより尤度が低いためである。

5. むすび

task-orientedな実時間不特定話者音声対話システムTOSBURGにおける、キーワードラティスに基づく対話音声理解の方法を述べた。本方式は、自由な発話からキーワードが検出される毎に構文意味解析を行い、複数の意味候補を対話処理部に送ることで、音声理解部と対話処理部の融合された処理を行うことができる。今後、文法の整備と並行して、音声認識辞書の改良と共に、音韻環境の影響を考慮したスコアリング方式を検討する予定である。

参考文献
[1]W.Ward, ICASSP91, pp.365-367(1991-11)
[2]R.De Mori他,ICASSP91,pp.797-800(1991-11)
[3]金沢他, 信学技報, SP91-22 (1991-6)
[4]竹林他, 本予稿集 (1992-3)
[5]M.Tomita,ICASSP86,pp.1569-1572,(1986-11)