音声言語情報処理 19－16
(1997. 12. 12)

異種言語知識の併用による局所的信頼性向上

塚田 元　山本博史　竹澤寿幸　匂坂芳典
ATR 音声翻訳通信研究所

〒619-02　京都府相楽郡精華町光台 2-2

Tel: 0774-95-1374　E-mail: tsukada@itl.atr.co.jp

あらまし　本稿では，信頼性の高い発話断片を認識する手法を提案する．提案法は，n-gram に基づく統計的言語モデルと文法制約の両方を，多少の文法的逸脱を許容しながら適用することで，文法にかなった部分的な発話断片を求める．また，文法制約としては，文脈自由文法を想定するが，制約適用の際に，有限状態オートマトンに近似することにより，効率的な文法制約適用を実現している．自然発話音声データベースを用いた認識実験によって，従来の n-gram に基づく言語モデルで得られる認識結果と比べて，本手法により，格段に信頼性の高い発話断片が得られることを確認した．

キーワード　CFG, FSA, n-gram, 言語モデル，音声認識

Reliable utterance segment recognition by integrating grammar and statistical language constraints

Hajime Tsukada, Hirofumu Yamamoto, Toshiyuki Takezawa, Yoshinori Sagisaka
ATR Interpreting Telecommunications Research Laboratories

2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02

Tel: 0774-95-1374　E-mail: tsukada@itl.atr.co.jp

Abstract　This paper proposes a novel approach to recognize partial segments of utterance with high confidence instead of the complete utterance. The proposed method is based on a cooperative use of conventional n-gram constraints and additional grammatical constraints, applied to utterances considering insertions, deletions and substitutions. To apply the grammatical constraints efficiently, the constraints described by a context-free grammar are approximated in the form of a finite-state automaton. Through an experiment, it has been confirmed that the proposed method can recognize partial segments of an utterance with a higher reliability than conventional continuous speech recognition methods using only n-grams.

key words　CFG, FSA, n-gram, language model, speech recognition

1 はじめに

自然発話を対象とした音声認識では，文法を逸脱した発話を受理し，効率的に単語候補を削減するn-gramに基づく統計的言語モデルが広く用いられている．一方，音声翻訳をはじめとする，音声認識結果を用いて言語処理を行なう対話システムでは，通常，この統計的言語モデルとは独立に，発話理解のための解釈用文法が用いられている．総合的な音声対話システムの性能を最大限高めるためには，これら二つの言語制約を協調して音声認識に用いることが望ましい．

これまでにも，発話理解のための文法を音声認識部の言語モデルに反映させる研究が行われてきた．一例をあげると，n-gramに基づく統計的言語モデルを併用しながら，文法制約を適用する手法[11]や，翻訳部の文法を音声認識の制約として利用する音声翻訳システムの研究がある[7][8]．しかしこれらの手法は，文法制約を厳格に適用するため，自然発話のように，しばしば文法的逸脱の見られる発話を認識できないという限界があった．また，認識結果も完全に文法にかなっているものしか許容しないため，局所的な誤認識に対する柔軟性に欠けるといった問題もあった．

この問題を解決するため，本稿では，n-gramに基づく統計的言語モデルの制約を併用しながら，文法制約を文法的逸脱を許容しつつ適用する手法を提案する．ここで提案する手法では，主にn-gramによって，単語列全体としての確からしさをモデル化する．さらに，文法との整合の度合いによって，認識単語列中の局所的な確からしさをモデル化する．これら二つの異なった言語モデルの併用により，最尤な単語列のみならず，その中の含まれる信頼性の高い発話断片を求めることができる．

これまで，音声認識における言語モデルの研究の多くは，最尤な単語列をモデル化することを目指してきた．しかし，音声対話システムのような複合的なシステムを考えたとき，音声認識の後段の処理で用いられる文法的な言語モデルとの整合性や，この文法的な言語モデルの観点で，確からしい発話断片を求めることが重要な課題となってくる．このような背景もあって，文脈自由文法(CFG)に基づく文法制約を，統計的な言語モデルと融合し，文法で完全に表現しきれない発話を認識する手法は，すでにいくつか研究されている[1][3][5][9]．また，文法制約の適用手法の研究としては，単語スキップの機能を取り入れたCFGに基づく効率的なロバスト・パージングの手法も提案されている[2]．本手法は，文法制約としてCFG形式を想定する点は，これらの研究と共通であるが，あらかじめCFGによる制約を有限状態オートマトンに近似することによって，より効率的かつ頑健な文法制約適用が可能となっている．

まず最初に，提案手法の概要を説明し，次に，文法制約の表現や適用方法について説明する．最後に，自然発話音声データベースを用いた認識実験によって，本手法の有効性を示す．

2 提案手法の概要

2.1 定式化

一般に，連続単語認識は，入力音声(O)が与えられたとき，$P(W|O)$を最大とするような単語列(W)を求める問題として定式化される．本稿で提案する認識手法は，Wと同時に言語的にタグづけされた単語列(W_T)を認識するというものであり，次のように定式化できる．

$$\begin{aligned} & argmax_{W_T,W}\ P(W_T,W|O) \\ = \ & argmax_{W_T,W}\ P(O|W_T,W)P(W_T,W) \end{aligned}$$

さらに，OはW_Tとは独立であるとみなすとともに，$P(W_T,W)$を展開すると次のようになる．

$$\begin{aligned} \approx \ & argmax_{W_T,W}\ P(O|W) \\ & P(W)P(W_T|W) \end{aligned} \qquad (1)$$

一般的な式と比べると，音響モデル$P(O|W)$，言語モデル$P(W)$の他に，本定式化では，単語列に対

する言語的なタグづけの尤らしさ $P(W_T|W)$ が新たに追加されている．

提案手法では，$P(W)$ を n-gram に基づく言語モデルで，$P(W_T|W)$ を文法によってモデル化する．さらに，言語的なタグとして，品詞および，文法逸脱を示すマーカ (Ins, Subst, Del) を用いる．例えば，文 "I(pron) saw(verb) a(det) girl(noun) with(prep) a(det) telescope(noun)" を生成する文法があるとき，"hi saw girl with a telescope" という単語列は，"hi(Subst⟨pron⟩) saw(verb) ϵ(Del⟨det⟩) girl(noun) with(prep) a(det) telescope(noun)" のようにタグづけされる．

文法が認識対象を適切にモデル化できている場合，認識結果に含まれる文法逸脱部は，信頼性が低いと考えられる．提案手法では，この文法逸脱部の単語を除くことによって，信頼性高く認識された発話断片を求める．

2.2 実現方法

本稿では，手法の有効性を確認することを目的に，簡易な実現方法を採用した．図1に，処理の流れを示す．

異なった言語制約を適用するにあたり，マルチパス探索手法を用いた．第一パスでは，n-gram に基づく言語モデルを用いて，最尤な単語列を求める．第二パスでは，文法を用いてこの単語列に対して，文法逸脱のタグづけを行う．これは，2.1節 (1) 式の W をもとめるに $argmax_W\ P(O|W)P(W)$ を，W_T を求めるのに $argmax_{W_T}\ P(W_T|W)$ を近似的値として使ったことに相当する．より正確に W，W_T を求めるためには，第一パスで得られた単語グラフを第二パスで再スコアづけするように今後拡張が必要である．

文法制約としては，文脈自由文法 (CFG) で表現されたものを想定する．これは，音声対話システムにおいて，音声認識部の後段の音声理解部は，構文構造を解析する必要があるため，CFG を拡張したものが広く用いられている理由による．この CFG で表現された文法制約を効率的に適用するために，あらかじめ有限状態オートマトン (FSA) に近似する．さらに，単語の挿入 / 脱落 / 置換を受理するように拡張するとともに，これら文法逸脱部にマーク付するため，有限状態トランスデューサ (FST) に変換する．この CFG から FST への一連の変換方法については，次の3節で詳述する．

信頼性の高い発話断片は，文法逸脱部を除くことによって，抽出することができるが，その除去方法もいくつか考えられる．4節の実験では，文法逸脱単語のみを除去しする方法 (loose method) と，隣接単語を含めて除去する方法 (tight method) を検討した．誤認識している単語は，単語境界も間違っていることが多いため，隣接単語を含めて除去する方法の方が，より高い信頼性の発話断片が得られると予想される．

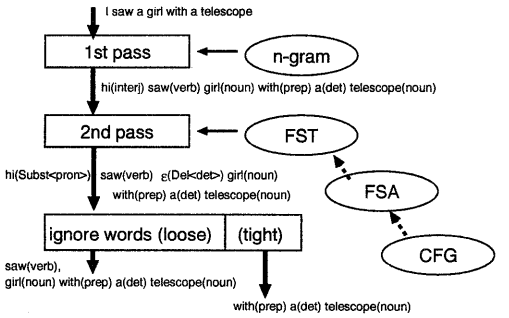

図 1: 処理の流れ

3 文法制約の表現

CFG で表現できる言語のクラスは，FSA で表現できるものよりも大きいため，一般的に，CFG を等価な FSA に自動変換することはできない．CFG を近似的に，FSA に変換する手法はいくつか存在するが，ここでは，Pereira らのアルゴリズム [7] を用いる．本アルゴリズムでは，基の CFG で受理される単語列は，変換後の FSA でも必ず受理されることが保証されている．図2の規則からなる CFG を，本アルゴリズムによって変換した FSA を図3に示す．CFG の終端記号は，品詞であるため，変換された FSA の

入力シンボルも品詞となっている．

SENT → NP, VP, NP.

SENT → SENT, PP.

NP → **det**, **noun**.

NP → **pron**.

NP → NP, PP.

PP → **prep**, NP.

VP → **verb**.

図 2: 英語の CFG 規則の例

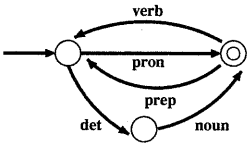

図 3: 近似された FSA

また，近似した FSA による文法制約の表現は，CFG と比べて，認識候補探索の観点からいくつかの利点を持っている．一つは，探索アルゴリズムが簡単であることであり，もう一つは，構文構造として曖昧な発話についても，近似変換された FSA では決定的なパスで受理可能という性質である．後者の性質は，認識タスクの規模が大きくなり，文法が複雑になるほどその効果を発揮する．これは，任意の FSA は決定的かつ状態数最小の FSA に等価変換できるという性質によるものである．例えば，“I(pron) saw(verb) a(det) girl(noun) with(prep) a(det) telescope(noun)” という文において，“with(prep) a(det) telescope(noun)” という前置詞句 (PP) は，図 2の CFG 規則「NP → NP, PP」の PP から生成されるか，それとも規則「SENT → SENT, PP」の PP から生成されるか曖昧である．しかし，図 3の FSA は，この例文を決定的なパスで受理する．

こうして，近似的に変換された決定的かつ最小の FSA を，文法的逸脱を受理するように拡張するとともに，逸脱した単語にタグづけできるように，入力シンボルに対応して出力シンボルを付与し，有限状態トランスデューサ (FST) に拡張する．図 3の FSA の任意の状態において，挿入 / 脱落 / 置換を想定した FST を，図 4に示す．

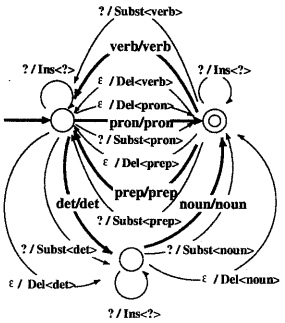

図 4: 文法的逸脱を受理する FST

スラッシュの左辺が入力シンボル，右辺が出力シンボルを表す．入力シンボル ϵ は特殊な記号で，入力を読むことなしに遷移が可能であることを表す．疑問符が含まれている状態遷移は，疑問符をすべての入力シンボルに置き換えた複数の状態遷移に相当する．出力シンボルの Ins⟨...⟩ は付加を，Del⟨...⟩ は削除を，Subst⟨...⟩ は置換を表す．

4節の実験では，文法制約の適用は，入力品詞列に対して文法逸脱単語数が最小となるような FST 上のパスを見つけ，対応する出力シンボル列を得る手続きとして実現した．この手続きによって，“hi(interj) saw(verb) girl(noun) with(prep) a(det) telescope(noun)” という文は，図 4の FST を用いて，“hi(Subst⟨pron⟩) saw(verb) ϵ(Del⟨det⟩) girl(noun) with(prep) a(det) telescope(noun)” のように再タグづけすることができる．以下の説明では，この文法制約の適用手続きを単にロバスト・パージングと呼ぶことにする．

本文法制約の適用手法を，CFG を直接用いる従来の手法 [1][3][9] [2] と比較すると，構文的な曖昧さにより探索空間が増大しないため，多くの場合に，効率よく文法制約の適用が行えるという利点をもつ．長い発話を認識する場合は，構文的な曖昧性も膨大なものとなるため，本手法は特に効果を発揮する．また，文法逸脱パタンとしても，単なる挿入だけでなく，脱落，置換を考慮にいれることで，より頑健な処理を実現している．

4 実験

提案手法が，信頼性の高い発話断片を認識する効果を確かめるために，次の二つの認識結果を比較した．

1. n-gram に基づく統計的言語モデルのみ用いてもとめた一位認識結果
2. 上記認識結果を文法制約でロバスト・パージングした後，文法を逸脱したとタグづけされた単語 (および隣接単語) を取り除いて残った発話断片

信頼性の評価尺度としては，認識結果中に含まれる正解単語の割合を表現するために，次の式で定義される適合率を用いた．

$$\text{適合率} = \frac{\text{正解と一致した単語数}}{\text{認識された単語数}} \times 100$$

また，以下の説明の都合上，文法のカバー率を次のように定義しておく．

$$\text{文法のカバー率} = \frac{W_1}{W_2} \times 100$$

W_1 : ロバスト・パージングの入力単語数

W_2 : loose/tight method で除去されないで残った単語数

4.1 実験条件

不特定話者の音声認識システム [10] を用い，ATR 自然発話音声データベース [6] 中のホテル予約 55 会話に含まれる日本語の 1,535 発話を対象に認識実験を行った．CFG としては，ポーズ句を単位とした部分木文法 [11] を用いた．本文法は 1,832 の規則からなり，1,132 語の辞書項目をもつ．文法は，認識対象として用いた 55 会話から代表的な 9 会話を選びそれを参考に作成したものであるが，認識対象発話を完全にカバーしているわけではない．タスク・カバー率は，89%(loose method)，71%(tight method) であった．これら文法逸脱の多くは，辞書項目が不足していることが原因となっている．また，併用する統計的言語モデルとしては，可変長 n-gram[4] を用いた．可変長 n-gram は，認識対象の 55 会話を含む 98 会話を用いて作成した．

4.2 実験結果

図 5に，実験結果を示す．長方形の中の区切られた部分の面積によって，可変長 n-gram を用いた一位認識結果に含まれる正解および誤った単語の割合を表す．斜めの分割線により，2.2節で説明した loose または tight method によって除去された単語および残った単語の割合を示す．この図でわかるように，loose または tight method によって，誤った単語は正解単語より多く棄却された．この結果，認識された発話断片に含まれる正解単語の割合が増し，表 1に示すように適合率を大幅に向上させることができた．

適合率は向上させることができたが，文法制約が可変長 n-gram の認識結果に含まれる正しい単語を棄却しているという点は，好ましくない．可変長 n-gram の認識結果に含まれる正しい単語を棄却せず，誤った単語を完全に棄却するのが理想である．正解単語列は認識対象タスクの単語列の傾向を反映していると考えられるため，正解単語のカバー率とタスク・カバー率には，強い関係があると考えられる．表 2に，両カバー率を示す．この結果が示すように，両カバー率はほぼ一致した値となっている．この実験結果は，今後辞書項目を充実するなどして，文法のカバー率を向上させれば，正解単語の棄却を抑えることが期待できるということを示唆している．

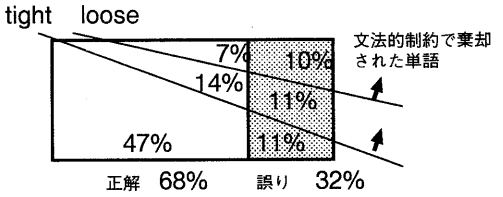

図 5: 実験結果

	適合率
可変長 n-gram のみ	68%
可変長 n-gram + FST(loose)	73%
可変長 n-gram + FST(tight)	81%

表 1: 適合率の向上

	loose	tight
正解単語カバー率	89%	69%
タスク・カバー率	89%	71%

表 2: 正解単語とタスクカバー率

5 むすび

n-gram に基づく統計的言語モデルと CFG の言語制約を併用することによって，信頼性の高い発話断片を認識する手法を提案した．我々の手法は，CFG に基づく制約を，文法的逸脱を許容する FST にあらかじめ変換しておくことによって，効率的な文法制約適用を実現している．自然発話音声データベースを用いた認識実験によって，従来の n-gram に基づく言語モデルで得られる認識結果と比べて，本手法により，格段に信頼性の高い発話断片が得られることを確認した．

音声対話システムを構築する際，本認識手法を用いることによって，信頼性高く認識された発話断片を頼りに，部分的な理解処理を行うことも可能となる．その結果，一般の認識手法を用いた場合には，不可能であった頑健な処理が実現できる．

謝辞

実験で用いた可変長 n-gram は，政瀧浩和氏から，音響モデルは，ハラルド・シンガー氏から提供頂いた．また，単語グラフに基づく音声認識システムは，清水徹氏らによって構築されたものである．これらの諸氏に感謝する．

参考文献

[1] W. Eckert et al., Combining stochastic and linguistic language models for recognition of spontaneous speech, ICASSP, 1996

[2] A. Lavie, GLR*: A robust parser for spontaneously spoken language, ESSLLI-96 Workshop on Robust Parsing, 1996

[3] H. Lloyd-Thomas et al., An integrated grammar/bigram language model using path scores, ICASSP, 1995

[4] H. Masataki et al., Variable-order n-gram generation by word-class splitting and consecutive word grouping, ICASSP, 1996

[5] M. Meteer et al., Statistical language modeling combining n-gram and context-free grammars, ICASSP, 1993

[6] T. Morimoto et al., Speech and language database for speech translation research, ICSLP, 1994

[7] F. Pereira et al., Finite-state approximation of phrase-structure grammars, ACL, 1991

[8] D. Roe, et al., A spoken language translator for restricted-domain context-free languages, Speech Communication, Vol. 11, 1992

[9] S. Seneff et al., Language modeling for recognition and understanding using layered bigrams, ICSLP, 1992

[10] 清水ほか，大語い連続音声認識のための単語仮説数削減，信学論 (D-II), Vol. J79-D-II, No. 12, 1996

[11] 竹澤ほか，部分木に基づく構文規則と前終端記号バイグラムを併用する対話音声認識手法，信学論 (D-II), Vol. J79-D-II, No. 12, 1996