音声言語情報処理　4 － 6
(1994. 12. 15)

フレーズスポッティングに基づく頑健な音声理解

河原 達也　北岡 教英　堂下 修司
京都大学 工学部 情報工学教室
〒606-01 京都市 左京区 吉田本町

あらまし

頑健な会話音声理解を実現するために、フレーズスポッティングに基づくアプローチを提案する。本アプローチは、(1) 単語連鎖、(2) フレーズ文法、(3) フレーズ間制約 の順に、徐々に強い制約を適用していく段階的探索に基づいており、各段階の処理は、前段階の結果をヒューリスティックスとする A* 探索として実現する。フレーズをスポッティングの単位とすることにより、単語スポッティングに比較してかなり高い抽出率を得た。また、本スポッティングアルゴリズムは、best-first 探索であるので、フレーズ候補を正しくスコア順に得ることができる。このスポッティングに基づく文認識・理解の探索戦略に関する検討を行ない、right-to-left 探索と island-driven 探索のアルゴリズムを提示する。最適な探索戦略を実現することで、認識精度を低下させることなく、自由発話に対する頑健性を実現した。

和文キーワード　音声言語理解, フレーズ, スポッティング, ヒューリスティック探索

Robust Spoken Language Understanding based on Phrase Spotting

Tatsuya Kawahara　Norihide Kitaoka
Shuji Doshita
Department of Information Science, Kyoto University
Sakyo-ku, Kyoto 606-01, Japan
e-mail: kawahara@kuis.kyoto-u.ac.jp

Abstract

For robust spoken language understanding, we propose a phrase spotting-based approach. It is based on progressive search strategy, which applies the constraint of (1) word concatenation (2) phrase syntax (3) inter-phrase semantics, in this order. The process of each stage is realized as an optimal (A*) search that uses the intermediate result of the preceding stage as heuristics. Use of the phrase as a spotting unit significantly improves the detection rate, compared with simple word spotting. The phrase spotting algorithm, which is best-first search, obtains the phrase candidates in the order of their scores. We also discuss the search strategies for sentence understanding based on the phrase spotting, and present algorithms of right-to-left search and island-driven search. With the optimal search strategy, the spotting-based approach can realize robustness as well as high accuracy.

英文 keywords　spoken language understanding, phrase, spotting, heuristic search

1 はじめに

音声言語理解システムにおいては、ill-formedな文を含む自然で自由な発話を扱えることが重要である。

自由発話に現れる言語現象は多種多様であり、それらをすべて記述するのが困難である。一方、限られたタスクにおいてはキーワードだけで意味が構成できることが多い。また、特に音声対話システムにおいては、認識・理解できないところは、対話を通して徐々に理解するということが可能である。したがって、認識不可能な部分をスキップしながら、意味理解に必要な部分のみを抽出していくスポッティングに基づくアプローチが有望である。

スポッティングに基づく音声理解はいくつかの成功例[1]はあるものの、特に中語彙以上のタスクでは成果をおさめていない[2]。これは、スポッティング自体が十分な抽出精度を得られず、湧き出し誤り(False Alarm)を多数生じるためである。

我々は、この問題を本質的に解決するアプローチとして、フレーズスポッティングに基づく音声言語理解の研究を進めているが、本稿では、フレーズスポッティング[3]自体の評価を行なうとともに、それに基づく文音声理解のための探索戦略について考察を行なう。

2 段階的探索 (Progressive Search)

従来のスポッティングに基づく方式における問題点としては、以下が考えられる。

1. 言語モデルを用いていない

   任意の時点(始終端フリー)で、任意の単語が均等に出現し得るというモデルは、実質的なパープレキシティが膨大なタスクに相当する。

2. 単語を単位としている

   マッチングのテンプレートが小さいと、安定した照合が困難になり、局所的な類似性やノイズの影響を受けやすい。

3. 決定的なセグメンテーションを行なう

   スポッティングの段階で各単語の決定的なセグメンテーションを行なうと、それらを組み合わせて文仮説を構成しても、その正確なViterbiスコアが計算できない。

上記の問題点を解決するために我々が用いる基本的な戦略は、段階的探索(Progressive Search)である。これは、以下のように、徐々に強い言語的制約を順次適用していくものである。

1. 入力発話全体に、単純な言語モデル(単語・音節の連鎖)
2. スポッティングには、局所的なフレーズ構文
3. 発話の意味理解には、フレーズ間の意味的制約

ここで、各段階の中間結果(HMMトレリス)は保存しておき、それを次の段階の探索におけるヒューリスティックスとして用いる。これにより、徐々に探索空間を狭め、よいヒューリスティックスを得ながら、効率のより探索が可能になる。また、中間結果を保存しておくことで、大局的な最適スコア(Viterbiスコア)を正しく計算できる。さらに、各段階で用いる制約を以降の段階で用いる制約の部分集合とすることにより、得られるヒューリスティックスがA*適格となり、次の段階での探索が最適解が保証されるA*探索として実現できる。具体的には、第1段階でViterbi探索、第2・第3段階でA*探索を実現する。

3 フレーズスポッティング

スポッティングに基づくアプローチでは、主としてキーワードが単位として用いられてきた。しかし、単語のテンプレートは小さく、局所的な類似性やノイズの影響を受けやすい。単語より長い単位を用いる方が、より多くの情報、すなわち言語的知識を導入できるので望ましい[4]。そこで、フレーズをスポッティングの単位とする。

フレーズは、'東京から'や'午後3時に'などのように、意味表現における格要素に対する記述として定義する[5]。これは、いくつかのキーワードと付属語から構成されるが、通常一呼吸で話され、自由発話においてもその構文が比較的保持される。しかも、フレーズは意味表現に直結するので、ill-formedな発話の頑健な理解に適している。

我々は、単語スポッティングにおいて、単純な言語モデルをバックグラウンドモデルとしてヒューリスティックに利用することを提案し、その効果を示した[6],[7]。本手法は、単語wが時刻t_1~t_2に存在するスコアを、ボトムアップ的なマッチングスコア$g(w)$と、残りの区間の言語らしさ$h(w)$の合計で評価するものである。

$$f(w,t_1,t_2) = h_l(w,1,t_1) + g(w,t_1,t_2) + h_r(w,t_2,T) \quad (1)$$

ここで、$h_l(w,1,t_1)$を前向きヒューリスティックス、$h_r(w,t_2,T)$を後向きヒューリスティックスとよぶ。この様子を図1に示す。

入力音声全体を評価することにより、長さの異なるスコアの比較が不要となり、大局的に最適化した

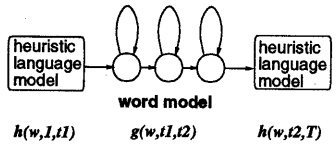

図 1: 言語モデルを用いた単語スポッティング

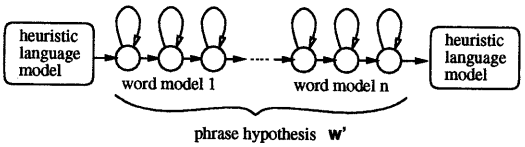

図 2: フレーズスポッティング

上で区分化ができる。また、言語モデルを用いることにより、実質的に入力のパープレキシティを減らすことになり、不自然な湧き出し誤りを抑えられる。さらに、このモデルは、スポッティング単語のモデルと同じく音素モデルに基づいて構築するので、単語部分のスコアと一貫性のある評価値が算出され、ヒューリスティックのスコアに基づいて、スポッティングのしきい値を動的に設定することも可能になる。

単語スポッティングと異なり、語彙数の組合せとなるフレーズをすべて評価するのは現実的でなく、探索的手法が不可欠である。会話音声理解を考えると、フレーズの上位 N 候補が正しく得られることが望ましい。しかし、スポッティングのように N の値が大きくなると、ビームサーチによる N-best 探索は難しい。そこで、best-first 探索を考える。best-first 探索は、候補をスコア順に出力するので、必要な数だけ候補を抽出するのも容易となる。また、island-driven 型の意味解析部との密結合も可能になる。

各フレーズ $\mathbf{w} = (w_1, \ldots, w_l)$ 候補について最良のもののみを求めればよい場合は、最適パスのみを計算する Viterbi アルゴリズムを適用することができる。もし、発話中に同一のフレーズが複数個出現することも考慮する必要があれば、N-best アルゴリズムに拡張すればよい。

$$f(\mathbf{w}) = \max_{1 \le t_1 < t_2 \le T} f(\mathbf{w}, t_1, t_2) \qquad (2)$$

部分フレーズ仮説 $\mathbf{w}' = (w_1, \ldots, w_n)$ に対する評価値は、それがフレーズを完成する期待値として、以下のように定義する。

$$\hat{f}(\mathbf{w}') = g(\mathbf{w}') + \hat{h}(\mathbf{w}') \qquad (3)$$

この場合、$\hat{h}(\mathbf{w}')$ は、バックグラウンドの言語モデルのスコアのみでなく、フレーズの未完成部分の期待値も示す。この様子を図 2に示す。

ここで、ヒューリスティックス $\hat{h}(\mathbf{w}')$ が、部分フレーズ仮説 $\mathbf{w}'$ のその後の展開で得られるスコアの上限となっているとき、この探索は A* 探索となる。このためには、ヒューリスティックス $\hat{h}(\mathbf{w}')$ に用いられる言語的制約が、フレーズの構文制約の部分集合であればよい。このとき、スポッティング候補がスコア順に正しく得られることが保証される。

3.1 ヒューリスティック言語モデル

バックグラウンドモデルとしてヒューリスティックに用いる言語モデルについては、単語スポッティングにおいて比較検討を行なった [7]。その結果に基づいて、ここでは、単語・音節連接モデルを用いる。これは、タスク内単語と日本語音節の任意の繰り返しを許すハイブリッドなモデルである。タスク内単語には、キーワードだけでなく、付属語なども含める。入力音声の既知語の部分は単語モデルにマッチングし、未知語や間投語の部分は音節モデルにマッチングすることが期待される。すなわち、語彙の知識が制約として働くと共に、未知語・間投語にも対処できる。このモデルを図 3に示す。

音節連接モデルは、A* 適格性を満たす。また単語連接モデルも、フレーズに出現するすべての語彙を含む場合、A* 適格性を満たす。

一般に、同一の音素モデルに基づいて構成した場合、音節連接モデルのスコアは単語モデルのスコアより高くなるので、単語モデルの部分が意味を持つためには、音節モデルのスコアを相対的に低くする必要がある。そこで、語彙制約に対する違反に関するペナルティ値を、音節モデルに課す。これは、音節モデルに関する A* 適格性を損なうが、単語連接モデルと併用した全体としては問題ない。

3.2 スポッティングアルゴリズム

フレーズの構文は、オートマトンもしくは LR 文法で記述し、現在の部分フレーズ仮説から次単語を予測できるようにする。

与えられた入力に対しては、まずヒューリスティック言語モデルを適用する。前向きヒューリスティックスと後向きヒューリスティックスに用いる制約は同一であるが、Viterbi 計算をそれぞれ left-to-right 及び right-to-left に行なう必要があり、そのトレリスを保存する。

そして、フレーズ仮説を best-first に展開する探索を行なう。このために、スタックデコーダを用いる。各仮説の評価値 $\hat{f}(\mathbf{w}')$ は、展開されたフレーズ部分と両方向のヒューリスティックスのトレリスを結

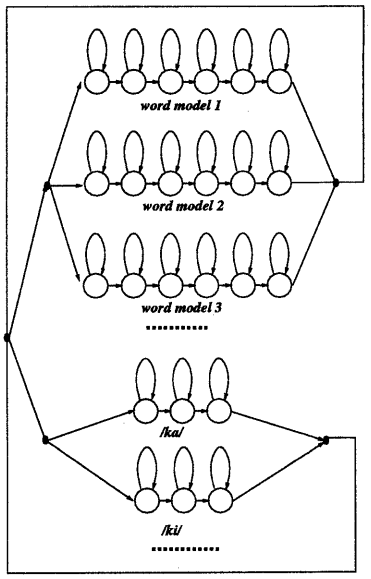

図 3: 単語・音節連接モデル

合することにより計算するが、単語連接モデルを用いる場合、この計算の多くはヒューリスティックス計算の結果をそのまま利用できる。

ヒューリスティックスが音節モデルのみの場合は、ヒューリスティックス計算自体が少なくなるが、探索の際にすべての仮説のトレリス計算を行なう必要がある。

スポッティングのしきい値、すなわち探索を終了する条件は、ヒューリスティックスによる最適スコアに基づいて決定する。

4 スポッティング実験

不特定話者の連続音声を対象にフレーズスポッティングの評価実験を行なった。発話サンプルとして、定型発話 50 文及び自由発話 25 文の 2 種の発話セットを用いた。自由発話 25 文は、間投語や言い淀みを含むものである。それぞれ 8 名の話者により発声された。各発話の例を以下に示す。

(定型発話の例)
「明日の 2 時から 3 時まで研究会を開きたい。」
(自由発話の例)
「えーと、来週の火曜日から、えー、木曜日まで出張します。」

フレーズ文法の語彙数は 230 で、単語パープレキシティは 24.6 である。ヒューリスティック言語モデルには、単語・音節連接モデルと、比較のために音節連接モデルを用いた。いずれの場合も、音節部分にはペナルティ値を課している。

スポッティング結果は、正解の抽出率と湧き出し誤り率の両者で評価される。単語スポッティングとの性能を比較するために、ここでは 219 キーワードの検出率を調べた。

スポッティング対象のキーワード総数 (W) に対して、サンプルの継続時間 (H) 当たりの湧き出し誤り (FA) 率が 0 ～ 10(FA/W/H) における、正解 (A) の抽出率 (A/W) の平均値を、FOM(Figure Of Merit) と定義する。なお本実験では、発話サンプルの継続時間の総和 (H) は、定型発話で 0.350 時間、自由発話で 0.224 時間であった。定型発話と自由発話に対する結果を表 1に示す。各発話サンプルについて、正解の抽出率 (A/W) と湧き出し誤り率 (FA/W/H) の関係である ROC(Receiver Operating Characteristic) 曲線を、図 4と図 5に示す。平均抽出率 FOM は、この曲線において、FA/W/H が 0 ～ 10 の区間の平均値である。

単純な単語スポッティングと比較して、フレーズをスポッティングの単位とすることにより、同一のヒューリスティックスを用いても、FOM が 10% 程度向上している。これより、フレーズスポッティングの有効性が示された。

ヒューリスティック言語モデルとしては、単語・音節モデルを用いる方が、音節のみのモデルよりかなり認識精度が高く、語彙レベルの知識の導入が効果的であることも確認された。

表 1: キーワード抽出率 (219 単語)

言語モデル spotting + heuristics	定型発話 FOM	自由発話 FOM
フレーズ + 単語・音節	78.7	68.2
フレーズ + 音節	73.7	61.5
単語 + 単語・音節	72.5	61.8
単語 + 音節	50.2	48.3

また、フレーズスポッティング自体の絶対的な評価のため、スポッティングされた全候補におけるフレーズ抽出率 (A/P) と湧き出し誤り率 (FA/P/H) を表 2に示す。

表 2: フレーズスポッティングの結果

	定型発話		自由発話	
	A/P	FA/P/H	A/P	FA/P/H
入力長に依存 単語・音節ヒューリスティックス	94.9	75.1	84.7	87.5

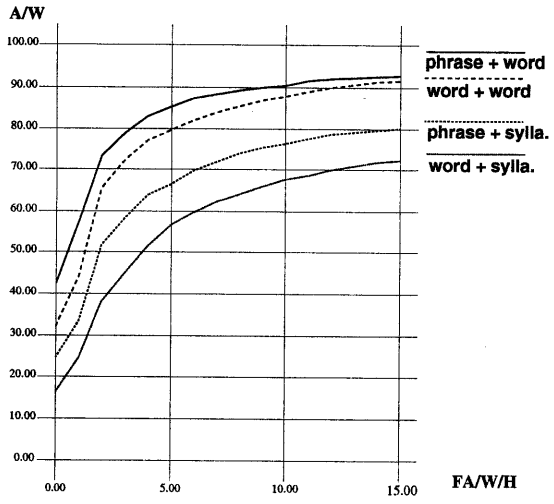

図 4: 定型発話に対するフレーズスポッティング結果 (219 キーワード検出率)

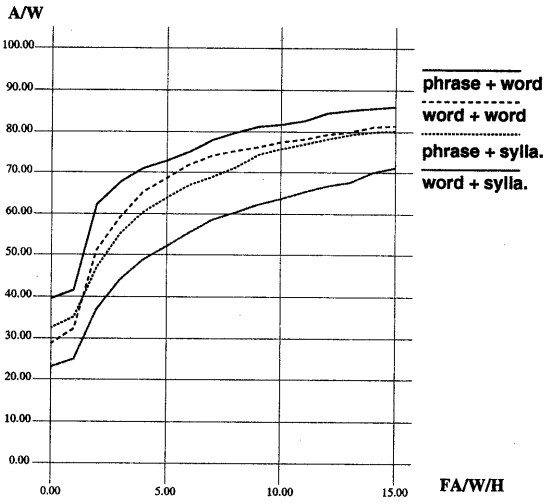

図 5: 自由発話に対するフレーズスポッティング結果 (219 キーワード検出率)

5 スポッティングに基づく文認識の探索戦略

従来のスポッティングに基づく認識では、得られた候補は、決定的にセグメンテーションされ、各候補の始終端とスコアからなる単語 / フレーズラティスの形式で表現されていた。この表現はコンパクトで、文認識の際の計算量を軽減する利点があった。しかしながら、候補の組み合わせに対する正しい Viterbi スコアを計算することは不可能になる。Viterbi 計算は、単語 / フレーズ間の任意の時点における接続について最大なものを選ぶことにより、大局的な最適化を行なう。ラティスを介することにより、最適化を行なえない情報損失は無視できないと考えられる。

したがって、Viterbi 計算を行なうのに必要な情報をスポッティングの中間結果とする。具体的には、各候補の最初と最後の状態のトレリスを保存しておく。意味レベルの言語的な制約に基づいて、フレーズ候補を組み合わせる際には、時間軸上でトレリスを接続して最適スコアを得る。この操作は、大局的な最適解を得る連続音声認識とほぼ等価になることが期待できる。しかも、結果的にフレーズスポッティングの際と同一のヒューリスティックスを用いる A* 探索が実現できる。

ただし、スポッティングを行なうことにより、仮説を構成する候補が限定される。これは情報の損失とも考えられるが、未知語や文法からの逸脱を含む発話に対して連続音声認識を適用すると、それらの部分で仮説が爆発することになるので、それらをスキップするスポッティングの利点である。

5.1 探索アルゴリズム

文仮説は、いくつかのスポッティングされたフレーズと認識されない部分からなる。フレーズ間のスキップされた部分は音節連接モデルで近似する。したがって、文仮説は図 6のようにモデル化される。ここで、全体が音節モデルにマッチングしないように、スキップに対するペナルティを音節モデルに課す。

探索におけるヒューリスティックスとして、フレーズスポッティングの際と同一の単語・音節連接モデルを用いる。これは、フレーズモデルと (ペナルティ付きの) 音節モデルのスコアの上限を与えるので、A* 適格である。探索アルゴリズムは、スタックデコーダを用いる A* 探索として実現できる。

アルゴリズムは以下の通りである。

1. 単一フレーズからなる初期文仮説を生成し、評価値と共にスタックへプッシュする。
2. スコアが最大の文仮説 n をポップする。
3. 仮説 n に終了フラッグが立っているならば、出力する。探索終了。
4. 仮説 n が文を完成しているならば、文モデルに対する Viterbi スコアを計算し、終了フラッグを立てて、スタックへプッシュする。
5. 仮説 n に接続するフレーズをフレーズ間知識から予測し、それぞれについて文仮説を生成し、評価値を計算して、スタックへプッシュ。ステップ 2 へ。

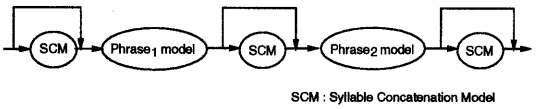

図 6: フレーズスポッティングに基づく文仮説モデル

ここで、仮説が文を完成するための条件は、文の意味制約を満たすことだけである。入力音声のカバー率は考慮しない。スコアの正規化は、探索途中では、未展開部分のヒューリスティックスコアで行なわれているが、このスコアはスキップした部分を音節モデルで近似する図6に必ずしも対応しない。したがって、ステップ4で、完成した文仮説の評価値計算を行なっている。すべての仮説に用いているヒューリスティックスはA* 適格であるので、もし文仮説の評価値が最大であれば、最適解である。

以下に、具体的な仮説の評価アルゴリズムを示す。ここでは、right-to-left 探索と island-driven 探索を考える。

Right-to-Left 探索

まず、時間軸方向に進む単純な一方向の探索を考える。探索の方向は、フレーズスポッティングと逆である必要がある。ここでは、left-to-right フレーズスポッティングを行ない、right-to-left 文認識を行なう。探索空間は、スポッティングされたフレーズの木となり、仮説の評価と管理が単純になる。

部分文仮説は、図7でモデル化され、ヒューリスティックスとあわせて評価される。このモデルの評価値計算は単純である。現在ポップされた仮説を n とし、フレーズ A を接続して新たな仮説を生成する場合を考える。この仮説の評価値は、n に音節モデルを加えた後向きトレリス $\beta_n(t)$ と、ヒューリスティックスを含むフレーズ A の前向きトレリス $\alpha_A(t)$ を接続することによって求められる。

$$\hat{f}(n+A) = \max_t \{\alpha_A(t) + \beta_n(t)\} \qquad (4)$$

ここで、A 自身のトレリス展開は、フレーズスポッティングの結果をそのまま利用できる。このトレリスは、この仮説が次にスタックからポップされてはじめて、$\beta_n(t)$ として実際に展開される。

Island-Driven 探索

次に、最尤のフレーズから出発し、スコアの高いフレーズから順に接続していく island-driven 探索を考える。これは、時間軸よりもむしろスコアの軸に沿って進む探索と考えられる。ある部分文仮説のフレーズ間に新たにフレーズが挿入される場合もあるので、異なった仮説から同一の仮説が生成される可能性がある。例えば、仮説 $\{A,B,C\}$ は、仮説 $\{A,B\}$ と仮説 $\{A,C\}$ の両方から生成され得る。すなわち、木探索でなくグラフ探索になる。したがって、仮説の管理が複雑になる。

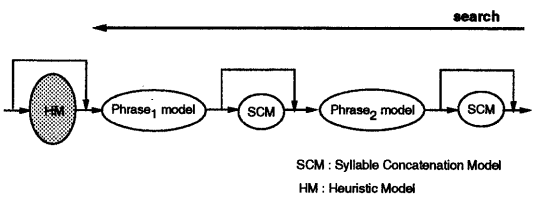

図 7: スポッティングに基づく right-to-left 探索

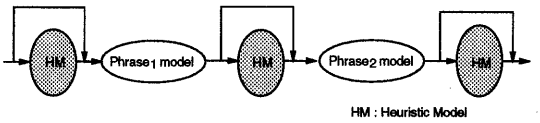

図 8: スポッティングに基づく island-driven 探索

部分文仮説は、図8でモデル化される。ここで、ヒューリスティックモデルは、未探索部分とスキップされる部分の両方を表現する。この場合、現在の部分文仮説とフレーズスポッティングのトレリスから Viterbi スコアを直接得ることは不可能である。そこで、近似的な評価関数を導入する。

現在ポップされた仮説を $n = \{A,C\}$ とし、フレーズ B を A と C の間に接続して新たな仮説を展開する場合を考える。新たな部分文仮説 $\{A,B,C\}$ の評価値は、上限であるヒューリスティックス $\hat{h}_0$ からのオフセットとして定義する。

$$\begin{aligned} \hat{f}(ABC) &= \hat{h}_0 - (\hat{h}_0 - \hat{f}_p(A)) - (\hat{h}_0 - \hat{f}_p(C)) \\ &\quad -(\hat{h}_0 - \hat{f}_p(B)) \qquad (5) \\ &= \hat{f}(AC) - (\hat{h}_0 - \hat{f}_p(B)) \qquad (6) \end{aligned}$$

$\hat{f}_p(B)$ は、フレーズスポッティングのスコアである。なお、初期値は $\hat{f}(null) = \hat{h}_0$ とする。

このアルゴリズムは、shortfall 法 [8] と同じ考え方である。新たにフレーズが付加される毎に、最適スコアからのオフセットを減じていく。文仮説が完成した段階 (アルゴリズムのステップ4) で、その正しい Viterbi スコアを改めて計算する。実際のトレリス計算は、この段階ではじめて行なわれる。したがって、実際の計算量は、最適解が得られるまでにいくつの文仮説が完成するか、しいては上記の近似評価値がどれほど精度がよいかに依存する。

5.2 知識源

探索を起動する知識源について考察を行なう。それらは、現在の部分文仮説を展開するフレーズを予測し、完成した文を受理する能力が要求される。

フレーズ間LR文法

フレーズ間文法は、フレーズを終端記号とする書き換え規則で可能な文のパターンを記述するものである。単語を終端記号とする文法より抽象度が高く、構文よりもむしろ意味やタスクの知識を表現している。

探索アルゴリズムは、通常のLRパージングとほぼ同じであるが、可能な終端記号がスポッティングされたフレーズに限定される。island-driven探索では、双方向の予測が必要となり、実現が複雑であるので、本研究ではright-to-left探索しか実現していない。この場合、認識不可能な部分をスキップする頑健なパージングの一種とみなせる。

意味ネットワーク

意味ネットワークは、タスクにおける単語と上位概念の関係を表現する。LR文法と異なり、単語やフレーズの表層的な語順を制約しないので、完全な意味主導理解を実現する。island-driven探索も、right-to-left探索とほぼ同様に実現できる。

意味ネットワークは、現在の部分文仮説を展開し得るフレーズを予測できる必要がある。ここでは、現在の仮説で活性化する概念ノードに不整合を生じない単語を含むフレーズを予測する機構を実現している[9]。

6 会話音声理解実験

以上の探索アルゴリズムと知識表現を、文音声認識・理解実験で評価を行なった。フレーズ間制約を用いて、スポッティングされたフレーズ候補を探索的に組み合わせながら、最尤の文仮説を得る。ここでは、単語・音節連接モデルをヒューリスティックスとしたフレーズスポッティングの結果を利用した。ただし、各発話におけるフレーズ候補の総数を最大150に制限している。

まず、最も単純な方式である、フレーズ間文法を制約としたright-to-left探索を実現した。A*探索のスタックサイズは100であり、生成する仮説の数は500に制限した。フレーズ間文法は、連続音声認識用の文単位の文法から作成した。

比較のために、文単位のLRパーザで文候補を得て意味解析する方式、単語対文法で文候補を得て意味解析する方式、単語スポッティングで単語ラティスを得て意味解析を行なう方式と比較した。文単位のLR文法の単語パープレキシティは15.6であり、文節間のポーズと文頭の一部の間投語にのみ対処している。単語対文法は、これから導出し、その単語パープレキシティは33.0である。単語スポッティングは、単語対制約をヒューリスティックスとする方式を用いている。いずれの場合も、同一の意味解析器を用いている。

文(意味)理解率及び単語認識率を表3に示す。各手法について、フレーズ間に任意の音節を許す場合と許さない場合の2通りを比較した。音節の挿入を許すと、間投語や未知語への対処が可能になるが、一方でパープレキシティが実質的に大きくなる。

定型発話に対しては、最も制約が強く、大局的な最適解が得られる文単位のLRパーザが最高の認識率を得た。ただし、フレーズスポッティングを介する方式も、5%程度の低下にとどまっている。定型発話は用意された語彙と文法に沿っているので、音節モデルを併用すると認識率の低下をもたらす。単純な単語対制約は、単語認識率はそれほど変わらないが、文認識率が極端に低下する。

自由発話に対しては、フレーズスポッティングに基づく方式が最高の認識率を得ており、本手法の頑健性が示された。文単位のLRパーザは、記述されていない非定型な現象に対して全く対処できないので、探索に失敗して解を得られない場合が多く、単語認識率が最低となった。音節モデルの併用は、頑健性の点から期待されたが、あまり効果がない。確かに間投語や未知語に対処できるが、一方で登録語にも一部マッチングするため、効果が相殺されるためである。単語対文法は、局所的な制約であるので、単語認識率はかなりよいものの、文レベルの頑健な理解は実現できていない。

フレーズスポッティングとフレーズ間文法を組み合わせると、文単位のLR文法と制約としてはほぼ等価になる。しかし探索戦略としては、定型発話については大局的な最適化を行なう連続音声認識が、自由発話については仮説が爆発しないフレーズスポッティングが有効である。

次に、スポッティングされたフレーズを組み合わせる知識源について検討した。フレーズ間文法と、意味ネットワークを比較した。right-to-left探索で行なった結果を表4に示す。フレーズ間に音節モデルを併用する場合としない場合の両方について示している。フレーズ間文法の方が、意味ネットワークより有効である。意味的な制約は、特に単方向探索においては弱い。

最後に、探索方向に関する考察を行なう。意味ネ

表 3: 会話音声理解のためのアプローチの比較

アプローチ	定型発話		自由発話	
	文	単語	文	単語
フレーズスポット	67.5	81.5	43.0	72.7
フレーズスポット + 音節	60.3	81.1	45.5	75.0
文単位 LR 文法	71.3	86.2	36.5	59.9
文単位 LR 文法 + 音節	62.5	82.3	37.0	61.3
単語対文法	58.3	82.9	23.5	72.3
単語対文法 + 音節	49.0	78.3	20.5	67.4
単語スポット	44.0	81.5	27.5	72.7

表 4: 会話音声理解のための知識源の比較

知識源	定型発話		自由発話	
	文	単語	文	単語
フレーズ間文法	66.8	81.5	43.0	72.7
フレーズ間文法 + 音節	59.8	81.1	45.5	75.0
意味ネットワーク	54.3	77.8	30.5	61.1
意味ネットワーク + 音節	48.0	74.1	31.5	67.3

right-to-left 探索

ットワークパーザに適していると考えられる island-driven 探索を実現した。right-to-left 探索との比較を表 5に示す。island-driven 探索は、right-to-left 探索よりも失敗する場合が多い。双方向に仮説を展開するので、仮説数が爆発するためである。意味主導型のパージングについては、さらに検討を行なう必要がある。

表 5: 会話音声理解のための探索方向の比較

探索方向	定型発話		自由発話	
	文	単語	文	単語
right-to-left	48.0	74.1	31.5	67.3
island-driven	44.0	56.9	27.0	43.0

意味ネットワーク + 音節モデル

7 おわりに

頑健な会話音声理解を実現するために、フレーズスポッティングに基づくアプローチを提案した。本アプローチは、(1) 単語連鎖、(2) フレーズ文法、(3) フレーズ間制約 の順に、段階的に強い制約を適用していくことにより、探索空間を徐々に狭めながら、各段階の処理を前段階の結果をヒューリスティックスとする A* 探索を実現する。

フレーズをスポッティングの単位とすることにより、単語スポッティングに比較してかなり高い抽出率を実現した。また、本スポッティングアルゴリズムは、best-first 探索であるので、フレーズ候補を正しくスコア順に得ることができる。単語連接モデルをバックグラウンドの言語モデルとして利用することの有効性も確認した。

スポッティングに基づく文認識の探索戦略に関する検討も行ない、right-to-left 探索と island-driven 探索のアルゴリズムを示した。従来のようにスポッティング時に決定的なセグメンテーションを行なわないことで、定型発話に対しても連続音声認識と同程度の認識率を得た上に、自由発話に対してはより高い文理解率を得た。

参考文献

[1] 坪井宏之, 竹林洋一, 橋本秀樹. キーワードスポッティングに基づく連続音声理解. 信学技報, SP91-95, 1991.

[2] 甲斐充彦, 間宮康之, 中川聖一. 自然発話の認識・理解のための解析・照合手法の比較. 情処学会研究会報告, SLP94-2-12, 1994.

[3] 北岡教英, 河原達也, 堂下修司. 自由発話認識・理解のためのフレーズスポッティング. 信学技報, SP93-116, NLC93-56, 1993.

[4] 伊藤慶明, 木山次郎, 岡隆一. 文スポッティングにおける部分文の認識. 信学技報, SP93-32, 1993.

[5] 伊藤彰則, 岡本東, 牧野正三. 対話音声認識のための文節構造モデル. 信学技報, SP94-27, 1994.

[6] 宗続敏彦, 河原達也, 荒木雅弘, 堂下修司. 自由発話理解のためのキーワードスポッティング法. 信学技報, SP92-116, 1993.

[7] 河原達也, 宗続敏彦, 三木清一, 堂下修司. 会話音声中の単語スポッティングのための言語モデルの検討. 信学技報, SP94-28, 1994.

[8] W.A.Woods. Optimal search strategies for speech understanding control. *Artificial Intelligence*, Vol. 18, pp. 295–326, 1982.

[9] 額賀信尾, 荒木雅弘, 河原達也, 堂下修司. 概念階層構造を持つネットワークを用いた漸進的音声言語理解. 信学技報, SP93-126, 1994.