音声言語情報処理 34－40
(2000. 12. 22)

# 複合的言語制約に基づくキーフレーズスポッティングによる対話音声理解

**鹿島 博晶　河原 達也**

**京都大学 情報学研究科**
**〒606-8501 京都市左京区吉田本町**
**e-mail: {kashima,kawahara}@kuis.kyoto-u.ac.jp**

**あらまし** キーフレーズスポッティングに基づく頑健な対話音声理解において、統計的言語モデルと記述文法を組み合わせて適用する方式を提案する。本研究では、キーフレーズ内にはタスクに関する記述文法、キーフレーズ外には類似タスクの対話コーパスによる単語 2-gram を適用する。これにより、キーフレーズ外に対しても比較的強い言語的制約を与えることができ、タスクに関する意味理解に直結するキーフレーズを高精度に抽出することができる。さらに抽出された複数のキーフレーズ候補に対し、フレーズ間の記述文法を適用することにより、頑健な対話音声理解を実現する。ホテル検索対話システムを用いて収集したデータに対して、文単位の記述文法による方式と比較し、文法内、文法外いずれの発話においても理解率の向上が確認された。

**キーワード** 音声対話システム, 音声理解, フレーズスポッティング, 記述文法, 単語 N-gram

# Speech Understanding Based on Key-Phrase Spotting and Combined Language Models

Hiroaki Kashima　Tatsuya Kawahara

School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan

**Abstract** We propose combined N-gram models and descriptive grammars used in key-phrases spotting for robust speech understanding. We apply task dependent descriptive grammars to inside of key-phrases and word bigram models trained on similar task dialogue corpus to others. The combined language models for whole of sentence improve the accuracy in detecting key-phrases. Furthermore connecting key-phrase hypotheses based on inter key-phrases grammars realize robust speech understanding. The proposed approach was tested on data collected from realworld dialogue system on hotel retrieval task. The speech understanding strategy improves the accuracy in handling both in-grammar and out-of-grammar utterances over the conventional decoding approaches.

**key words** dialogue system, speech understanding, key-phrase spotting, descriptive grammar, word N-gram

# 1 はじめに

現在、音声認識を用いた情報検索や案内を行うシステムは、単語認識ベースのものがいくつか実用化されているものの、自然言語を用いたものはまだ十分に動作していない。これは、多様な言い回しや間投詞、言い淀みなどに十分に対処できていないためである。

これに対して検索や予約などのタスクを遂行するのに必要なキーフレーズ[1] に着目し、これらを認識して発話意図を理解する方式が有望である。

しかし、スポッティングに基づくアプローチは一般に言語制約が緩くなるため、特に文法内発話に対する精度が低下する問題があある。これに対して本研究では、統計的言語モデルを併用することにより改善を図る。本稿では、まずこの複合的言語制約を説明し、この言語制約を用いたキーフレーズスポッティング手法とそれに基づく対話音声理解について述べ、音声理解率を用いた実験的評価によりその有効性を検証する。

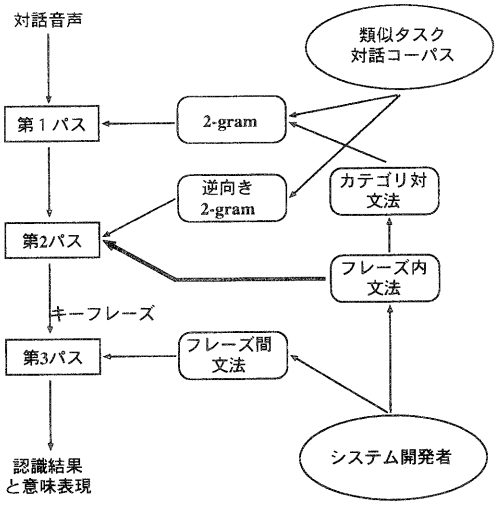


図 1: 音声理解アルゴリズムの全体構成

# 2 複合的言語制約の基本的概念

近年の音声対話システムでは、そのタスクに特化した対話コーパスが大量に得られる場合、N-gram モデルが用いられている [1][2]。この言語制約により、非定型な発話に対しても柔軟に対処できるが、個々のタスク毎に大量の対話コーパスを収集するのは困難である。

このことから記述文法[2] を言語制約に用いたシステムも多い [3][4]。[5] ではシステムの移行に際し、元のタスクの 2-gram を用いることができず、新たなシステムでは記述文法を用いている。記述文法の利点として、タスクに特化した知識を容易に導入でき、また語彙などの変更も簡単に行えるという点がある。しかし、受理可能な発話を完全に限定してしまうため、ユーザに自由な発話を許さないという問題がある。さらに、全く異なったタスクへの移行の際に、文単位の文法を作成し直すには、大きなコストがかかる。

これに対して我々は、対話音声など自由な発話を受理するため、キーフレーズスポッティングに基づく手法を提案した [6]。キーフレーズ内の文法は、比較的単純かつ定型的であるので文法作成のコストも小さいし、非定型な発話に対する頑健性の向上も示されている。しかし、キーフレーズとフィラーが任意に接続するような非常に緩い言語制約であるので、特に短いキーフレーズの湧き出し誤りが多く発生することを確認した。

[1] キーワード単体も含む

[2] ここでは、文脈自由文法や有限状態文法などのルールベースの文法を指す

以上の言語制約の利点と問題点を考慮した上で本研究では、タスクに特化したキーフレーズ部分に記述文法を、特にタスクに依存しないフィラー部分に対しては類似タスクのコーパスから得られる 2-gram を与える、複合的言語制約を提案する。フィラー部分に単語 N-gram を適用することにより、文全体に対する言語制約が強くなり、認識精度の向上が期待される。また、この単語 N-gram は類似タスクの大規模コーパスを利用できるので、タスクのポータビリティもよい。

さらに文全体を N-gram でデコードするのではなく、意味理解に必要なキーフレーズに着目してスポッティングすることにより、多数のキーフレーズ候補を効率よく求めることができる。

# 3 記述文法と類似タスクコーパスによる 2-gram の構築

本手法では、統計的言語モデル学習用のコーパスとして、類似タスクであることを前提としており、完全な合致を要求しない。この前提により、実行するタスクによっては、コーパスのタスクの特徴をより強く反映している 3-gram を用いることによる副作用が生じうる。したがって、ここでは、前向き後ろ向きともに 2-gram を用いる。以下、本手法における 2-gram 構築について説明する。

まず、キーフレーズ部分に直接作用しないフィラーどうしの遷移時に用いる 2-gram は、コーパスによる推定値

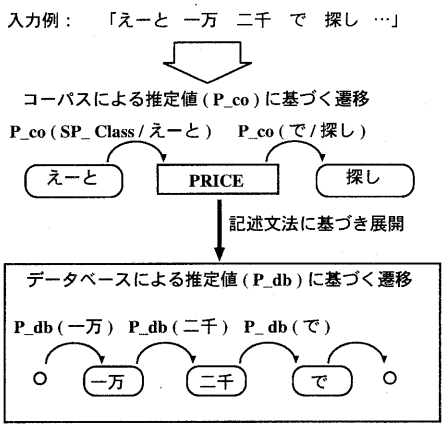


図 2: 構築した 2-gram の適用例

をそのまま適用する。これは、特に文末表現などのフィラー部分の言語制約として働く。

次に、フィラーとフレーズ間の遷移時に与える 2-gram には、コーパスによる推定値を近似的に用いる。フレーズ内単語は、システム開発者によりタスクに特化した語彙となるため、フレーズ内の全ての単語と同じ単語が類似タスクのコーパス中に存在することは期待できない。したがって、ここではコーパス中の名詞 (主に普通名詞) のクラス 2-gram を用いる。また、比較的タスク独立な概念 (値段、日付、地名など) をもつフレーズ内単語については、コーパス中の同概念のフレーズをクラス化し、そのクラス 2-gram を用いる。さらに、このクラス 2-gram を単語 2-gram に展開する際、システムで用いられるデータベース (あるいはそれに類似したもの) による 1-gram に基づき分配する。

また、キーフレーズ内の遷移は、記述文法から導出されるカテゴリ対制約に基づき設定される。カテゴリ対を単語対に展開する際も、データベースにより推定された 1-gram に基づき確率を分配する。以上の手順により構築した 2-gram の適用例を図 2 に示す。

# 4 キーフレーズスポッティングに基づく音声理解

従来のスポッティングを用いた手法では、その単位として主にキーワードを用いることが多い。しかし、単語のテンプレートでは局所的な類似性やノイズの影響を受けやすく、より長いフレーズを単位とした方が抽出精度が高いことが示されている [7]。したがって、ここではフレーズをスポッティングの単位とする。フレーズは、「所在が」や「一万円以下の」のようにキーワードとその付属語からなり、対話音声のような非定型な発話においても、その構文が保持される。さらに、フレーズの単位は意味表現にも直結するため、頑健な音声理解が期待できる。

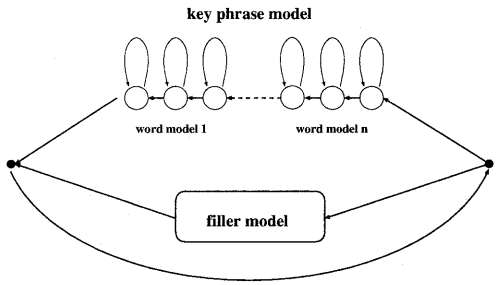


図 3: フレーズ・フィラー 連接モデル

本手法では、特にキーフレーズに着目した段階的探索を行う。これは、以下のようにキーフレーズ部分に対し、段階的に強い言語的制約を適用していくものである。

**第 1 パス** 単語 2-gram モデル 3

**第 2 パス** スポッティング時に局所的にフレーズ内文法

**第 3 パス** フレーズ接続時に意味制約としてフレーズ間文法

第 1, 第 2 パスにおいて、単語 2-gram と記述文法の複合的言語制約に基づきキーフレーズをスポッティングする。スポッティングされたフレーズは、第 3 パスにおいて、フレーズ間の記述文法に基づき接続され、文として理解する。

## 4.1 スポッティングアルゴリズム

まず第 1 パスでは、構築した 2-gram に基づき前向きにフレーム同期のビーム探索を行い、入力全体として単語候補の絞り込みを行う。

次に第 2 パスでは、第 1 パスの結果をヒューリスティックとし、スタックデコーディングによる後ろ向きの best-first 探索を行う。この探索過程においてキーフレーズのスポッティングを行う。ここで用いる言語的制約は、フレーズ・フィラ?の連接モデル (図 3) である。これは、キーフレーズとフィラ?の任意の繰り返しを許すモデルであるが、フレーズ内ではフレーズ内文法が完全に適用される。さらに、基本的に全ての遷移に対し、後向きの 2-gram が適用される。

```
S   : SENTLOOP
SENTLOOP   : SENTLOOP SENT
SENTLOOP   : SENT
SENT   : FIELD_NAME1 PRICE
SENT   : PRICE
SENT   : FIELD_NAME2 PLACE
SENT   : PLACE
```

図 4: フレーズ間の記述文法の例

スタックデコーディングを行うことにより、N-best の候補を求めることができるが、これは入力全体をデコードすることを前提としているため、一部の数単語 (主に機能語) が置換しただけの類似した単語列が出力されることが多い。一方、本アルゴリズムの目的は音声理解に重要なキーフレーズの候補を得ることである。このようなスポッティングを実現するために、仮説のマージを行う。すなわち既出のフレーズ列と同じフレーズ列を導く展開を行う経路は破棄する。具体的には、キーフレーズあるいはフィラーモデルとして受理された仮説にさらに 1 単語接続された仮説がスタックから取り出されたとき、その接続境界時刻を保存する。この時刻が既に保存されている時刻と一致すればその仮説の展開はそこで中止される。ここで受理された仮説がキーフレーズであればその始終端時刻やその仮説のスコアとともに出力する。

なお、[7] では、厳密な A*探索を実現するために 3 パスでスポッティングを行っていたが、このように前向き、後向きの 2 パスでも実質的に同じ候補が得られることがわかった [6]。

### 4.2 フレーズ間文法を用いたキーフレーズ接続

第 3 パスでは、スポッティングされたフレーズを組み合わせて、文として認識・理解する。探索手法としてトレリスパージングも考えられるが、精度の向上の割に計算コストが大きいため、ここではラティスパージングを採用する [7]。この場合、フレーズ候補はそのスコアと意味制約にしたがって接続される。意味制約としてシステム開発者により記述されるフレーズ間文法を用いる。その例を図 4 に示す。

探索アルゴリズムとしては、第 2 パスと同様にスタックデコーダを用いた best-first 探索を採用する。フレーズを接続する際には、次のような近似的な評価関数を導入する。スタックのトップにある仮説を $q_0 = w_1, w_2$ とし、これにフレーズ $w_3$ を接続して生成する場合を考える。このとき文仮説 $q_1 = w_1, w_2, w_3$ の評価値は、上限であるヒューリスティックの最大スコア $h_0$ からのオフセットとして次のように定義する。

$$\begin{aligned}\hat{f}(w_1w_2w_3) &= h_0 - (h_0 - \hat{f}(w_1)) - (h_0 - \hat{f}(w_2)) \\ &- (h_0 - \hat{f}(w_3)) \\ &= \hat{f}(w_1w_2) - (h_0 - \hat{f}(w_3))\end{aligned}$$

ここで $\hat{f}(w_i)$ はスポッティング時に求められたスコアである。初期仮説は、$\hat{f}(null) = h_0$ とする。$h_0$ は、第 1 パスでの最尤パスのスコアである。

ただし、スポッティングにおいては、文全体のパーズを前提としていないため、このままでは単語数の少ない仮説が優先的に受理される。そこで、ここではフレーズ間のスキップする区間に比例した一定値をさらに減じた。

$$h_0 - \hat{f}(s) = p \times (t_2 - t_1)$$

## 5 対話音声理解実験

### 5.1 実験条件

本手法の評価を、対話音声理解実験により行った。本手法の実装は、大語彙音声認識エンジン Julius[8]、Julian[9] をもとに行った。比較のため、文単位の記述文法に基づき得られた文候補を意味解析した方式 [10]、言語制約としてフレーズ・フィラーの連接モデルのみ用い 2-gram を与えないでキーフレーズスポッティングを行う方式 (Julian をもとに実装) と比較した。文単位の記述文法では、文頭あるいは文末へのフィラー挿入を許している。

評価タスクは、ホテル検索である。評価用音声サンプルとして実際に音声対話システム [11] を用いて収集した 665 発話 [12] を用いた。話者は、音声対話システム初心者 24 名 (男性 19 名、女性 5 名) である。

音響モデルは、性別依存で学習された不特定話者向け triphone HMM[13] である。総状態数は 2032 であり、1 状態当り 16 混合分布を持つ。

表 1: 2-gram 学習用コーパスの仕様

| コーパス | テキストサイズ | 語彙サイズ |
|---|---|---|
| ATR-SDB | 136051 | 3430 |
| ATR-SLDB | 37552 | 2015 |
| RWC | 34762 | 2418 |
| total | 208365 | 5432 |

表 2: 対話音声理解を目的とした FA+SErr による従来方式との比較

| | 単語数 | 文法内発話 | 準文法内発話 | 文法外発話 | total |
|---|---|---|---|---|---|
| 正解数 | | 561 | 116 | 120 | 797 |
| 文単位<br>記述文法 | 942 | 14.8% | 51.4% | 175.2% | 45.8% |
| フレーズスポッティング<br>(連接モデルのみ) | 1290 | 16.8% | 34.2% | 154.2% | 37.9% |
| **フレーズスポッティング<br>(複合的制約)** | 6124 | **11.9%** | **30.5%** | **149.4%** | **33.2%** |

2-gram 学習用コーパスとして、ATR 自然発話音声データベース (ATR-SDB,ATR-SLDB) と RWC 音声対話データベース (RWC) より、客の発話部分を用いた。コーパスのサイズは、のべ 21 万単語 (=形態素)、語彙数は 5432 単語である (表 1)。タスク依存フレーズ内文法の単語数は 712 単語であり、2-gram 学習用コーパスの単語と組み合わせて、合計 6124 単語の単語辞書が構成された。2-gram 学習時にはバックオフ平滑化を行っており、バックオフ係数の推定には Witten Bell ディスカウンティングを用いている。カットオフは行っていない。

対話音声理解を目的とした評価基準として、内容語 (=スロット) の誤受理率 (False Acceptance; FA) と誤棄却率 (Slot Error; SE) の和を用いる。

$$FA = \frac{\text{受理した中で誤っていたスロット数}}{\text{受理したスロット数}}$$

$$SErr = 1 - \frac{\text{受理した正解スロット数}}{\text{実際の正解スロット数}}$$

評価用サンプルにおける全正解数は、797 である。

## 5.2 対話音声理解における従来方式との比較

従来方式との FA+SErr[3] の比較を表 2 に示す。評価用サンプルは、以下のように 3 タイプに分類されている。ただし、ここでの語彙と文法は、文単位の記述文法におけるものである。ホテル検索における各タイプの発話例を図 5 に示す。

- 文法内発話 : 用意された語彙と文法に完全に従っている
- 準文法内発話 : 助詞の省略、言い淀み、文頭末のフィラー挿入を含む
- 文法外発話 : 未知語、文法外表現、文中のフィラー挿入がある

[3] 置換誤りは FA と SErr の両方で計数されている

**文法内発話**
**所在が京都市の宿**
**ホテルタイプは旅館でお願いします**
**レストランとバーのあるホテル**

**準文法内発話**
**所在、京都市の宿**
**旅館、旅館で**
**えっとー、レストランとバーのあるホテル**

**文法外発話**
**所在が三条の宿　　（「三条」が未知語）**
**旅館タイプでお願いします（「旅館タイプ」が文法外表現）**
**レストランと、えーと、バーのあるホテル**

図 5: ホテル検索における発話例

文法内発話においては、複合的言語制約を用いた本手法が最高の理解率を示しており、特に言語制約の緩いフレーズ・フィラーの連接モデルを用いたフレーズスポッティングによる方式と比較し、大きな向上が確認され、フィラー部分に対する 2-gram 適用の有効性が示された。また、文単位の記述文法と比較しても高い理解率を示した。これは、文法内発話サンプルのいくつかが音響的には一部不明瞭であり、制約の強い文単位の記述文法では対処できなかったことに起因すると考えられる。

準文法内、文法外発話に対しても本手法が最高の理解率を示している。これらの発話においては、文単位の記述文法に基づく方式に対し、フレーズスポッティングに基づく方式が全体的に高い理解率を示しており、非定型な発話に対する頑健性が示された。

表 3: スポッティングにおける言語制約の効果

| 言語制約 | FA | SErr | total |
|---|---|---|---|
| 連接モデルのみ | 19.3% | 18.6% | 37.9% |
| 複合的言語制約 | 15.7% | 17.6% | 33.2% |

また、全発話に対する理解率において、フレーズ・フィラー連接モデルによるスポッティングと比較し、5%向上した (表 3)。このうち FA における向上が 4%を占めており、複合的言語制約を用いた本手法により、キーフレーズの沸き出し誤りが削減されていることが確認された。

## 6 まとめ

キーフレーズスポッティングにおける言語制約として、記述文法と単語 2-gram を複合的に適用することにより、頑健な対話音声理解を行う手法を提案した。本研究では、キーフレーズ内にタスクに関する記述文法、キーフレーズ外には類似タスクによる対話コーパスによる 2-gram を適用する。これにより文全体に比較的強い言語制約が与えられ、キーフレーズを高精度に抽出する。さらに、意味理解に直結したキーフレーズ候補に、フレーズ間文法を適用することにより音声理解を実現する。

本アプローチは、文全体の記述文法と比較し、文法内、文法外いずれの発話においても理解率が向上した。また、2-gram を用いずキーフレーズ・フィラー連接モデルを用いたキーフレーズスポッティングに基づく方式と比較し、全体として理解率の向上が確認され、特に誤受理率の大きな減少を確認した。

今後は、キーフレーズ毎に信頼度を求め、効果的に確認を行う対話戦略 [10] との統合を行う予定である。

## 参考文献


[1] L.F. Lamel, S. Rosset, J-L.S. Gauvain, and S.K. Bennacef. The LIMSI ARISE system for train travel information. In *icassp*, pp. 501 504, 1999.

[2] B. Pellom, W. Ward, and S. Pradhan. THE CU COMMUNICATOR: AN ARCHITECTURE FOR DIALOGUE SYSTEMS. In *Proc. ICSLP*, Vol. 2, 2000.

[3] 中野幹生, 堂坂浩二, 宮崎昇, 平沢純一, 田本真詞, 川森雅仁, 杉山聡, 川端豪. TV 番組の録画予約を受け付ける実時間音声対話システム. 情報処理学会研究報告, 98-SLP-22-8, 1998.

[4] 桐山伸也, 広瀬啓吉. 文献検索音声対話システムの機能拡張とその評価. 情報処理学会研究報告, 2000-SLP-30-10, 2000.

[5] 小暮悟, 伊藤敏彦, 中川聖一. 音声対話システムの移植性に関する考察 観光案内システムとデータベース検索システム . 情報処理学会研究報告, 99-SLP-25-3, 1999.

[6] T.Kawahara, C.-H.Lee, and B.-H.Juang. Flexible SpeechUnderstanding Based on Combined Key-Phrase Detection and Verification. In IEEE Trans. Speech Audio Processing, Vol.6,pp.558-568, 1998.

[7] 河原達也, 北岡教英, 堂下修司. A*探索に基づいたフレーズスポッティングによる頑健な音声理解. 電子情報通信学会論文誌, Vol. J79-DII, No. 7, pp. 1187-1194, 1996.

[8] 李晃伸, 河原達也, 堂下修司. 単語トレリスインデックスを用いた段階的探索による大語彙連続音声認識. 電子情報通信学会論文誌, Vol. J82-DII, No. 1, pp. 1 9, 1999.

[9] 李晃伸, 河原達也, 堂下修司. 文法カテゴリ対制約を用いた A*探索に基づく大語彙連続音声認識パーザ. 情報処理学会研究報告, 98-SLP-24-15, 1998.

[10] 駒谷和範, 河原達也. 音声認識結果の信頼度を用いた頑健な混合主導対話の実現法. 情報処理学会研究報告, 2000-SLP-30-9, 2000.

[11] 田中克明, 河原達也, 堂下修司. 汎用的な情報検索音声対話プラットフォーム. 電子情報通信学会技術研究報告, SP98-109, 1998.

[12] 安達史博, 駒谷和範, 河原達也. 音声対話情報検索システムにおける想定外の発話の分析とその対処. 人工知能学会研究会資料, SIG-SLUD-A001-2, 2000.

[13] 河原達也, 李晃伸, 小林哲則, 武田一哉, 峯松信明, 嵯峨山茂樹, 伊藤克亘, 伊藤彰則, 山本幹雄, 山田篤, 宇津呂武仁, 鹿野清宏. 日本語ディクテーション基本ソフトウェア (99 年度版) の性能評価. 情報処理学会研究報告, 2000-SLP-31-2, 2000.