TR-I-0357

付属語の bigram、自立語の bigram を用いた音声認識 関連プログラム ユーザーズマニュアル

Speech Recognition Using Particle Bigrams and Content-Word Bigrams — User's Manual —

磯谷 亮輔
Ryosuke Isotani

1993.3

概 要

付属語 bigram、自立語 bigram を用いた音声認識に関するプログラムについて、コンパイルの方法、機能、使用法を簡単に説明する。

ATR 自動翻訳電話研究所
ATR Interpreting Telephony Research Laboratories

1 はじめに

付属語 bigram、自立語 bigram を用いた音声認識 [1] に関するプログラムについて説明する。プログラムは、

- テキストデータから付属語 bigram、自立語 bigram を学習するプログラム
- 学習した付属語 bigram、自立語 bigram を用いて、SSS-LR の出力する文節ラティスから文候補を選択するプログラム
- 周辺プログラム

からなる。サンプルのデータファイルも付属している。

2 動作環境

本ソフトウェアは、以下の環境でコンパイル、動作が確認されている。

マシン： DECstation 5000/200 (メインメモリ 32MB)
OS： ULTRIX V4.2A

また、一部のプログラムは jperl がインストールされ、jperl コマンドが /usr/local/bin に置かれていることを仮定している。jperl は perl を日本語化したもので、ftp サイトなどからフリーで入手できる。動作は jperl version 4.019 + 1.3(EUC) で確認した。

3 プログラムのコンパイル

ファイルはテープに `tar c .` で格納されている。これを展開し、`src/` ディレクトリで make すると、以下の 3 つのプログラムがコンパイルされる。

```
calc_ngram
ph_ngram
print_bi
```

この他、ディレクトリ `perl-scripts/` に以下のコマンドが置かれている。これらはすべて jperl で書かれており、/usr/local/bin に jperl がインストールされていればそのまま実行できる。

```
fuzoku
contents
mkwtab
score
```

4 プログラムの説明

4.1 文テキストからの付属語の抜きだし

コマンド名

```
fuzoku
```

形式

```
fuzoku [-a] [-w] [-p] input_file
```

機能

テキストファイル input_file から、文節末の単語を抜き出す。input_file の形式は ATR 対話データベース ADD の単語テーブルの形式で、EUC コードで書かれているとする。出力は後述の calc_ngram コマンドで用いる。出力の各行は、出現単語、読み、標準表現、音便コード、品詞コード、活用形、活用型の情報を“|”で区切ったもの (ATR 対話データベース ADD の単語テーブルのうち第 6 から第 12 フィールドのみを取り出したものに相当) である (-w オプションを指定しないとき)。

引数の説明

-a 入力テキスト中、間投詞、言い直し・言い淀み ([] 、 () に囲まれた部分) も対象とする。デフォルトでは、これらの語は読みとばす。

-w 出現単語のみ出力する。

-p 各文節の内容も出力する。

4.2 文テキストからの自立語の抜きだし

コマンド名

```
contents
```

形式

```
contents [-a] [-w] [-p] input_file
```

機能

テキストファイル input_file から、文節の最初の単語を抜き出す。出力は後述の calc_ngram コマンドで用いる。input_file および出力ファイルの形式は fuzoku と同じである。

引数の説明

-a 入力テキスト中、間投詞、言い直し・言い淀み ([] 、 () に囲まれた部分) も処理の対象とする。デフォルトでは、これらの語は読みとばす。

-w 出現単語のみ出力する。

-p 各文節の内容も出力する。

4.3 付属語 / 自立語 bigram 計算用単語テーブルファイル作成

コマンド名

```
mkwtab
```

形式

```
mkwtab dictionary word_list
```

機能

認識用単語辞書 dictionary と単語リストファイル word_list から、後述の calc_ngram で用いる単語テーブルファイルを作成する。辞書は SSS-LR の文法で用いている辞書ファイル、また、単語リストファイルは各行がスペースで区切られた 2 つのフィールドからなり、第 2 フィールドが fuzoku ある

いは contents の出力と同型式であるようなファイルである (後述のオペレーション例参照)。第 1 フィールドは、本プログラムでは無視される。

出力は単語リストファイル word_list の単語のうち単語辞書 dictionary に含まれるものだけを抜き出して、その単語と辞書に書かれている単語 ID との対応づけを行なったものである。ひとつの単語に複数の単語 ID が対応することを許す。出力の形式は、後述のオペレーション例を実行して得られるファイルを参照されたい。

一般には本プログラムの出力はそのまま calc_ngram で用いるのではなく、特定の品詞の単語を取り出すなど何らかの加工をすることを前提としている。また、そのまま使おうとすると、単語 ID の重複がおこりエラーとなることがある (後述の calc_ngram の説明参照) ため、あるフィールドを "*" に書き換える (calc_ngram の説明参照) ことにより同じ ID の単語をまとめるか、不要な行を削除するという修正が最低限必要となる。

4.4 付属語 bigram、自立語 bigram の出現確率の計算

コマンド名

```
calc_ngram
```

形式

```
calc_ngram -2 -t table_file [-c] [-v] [-w outfile] file
```

機能

fuzoku あるいは contents の出力として得られる形式の入力テキストファイル file を読み込んで、table_file に指定された単語を対象として bigram の出現確率を計算する。table_file の形式は、mkwtab の出力と同じである。ただし、第 1 フィールド中の "|" で区切られた任意のフィールドに "*" を書くことができる (後述)。-w オプションが指定された場合、結果を binary data として outfile に書き込む。

処理の概要

calc_ngram は、まず単語テーブルファイル table_file を読み込んで、各行を 1 つの単語クラスとして扱う。単語クラスは、単語の情報と対応する単語 ID を持つ。1 つの単語クラスには 1 つ以上 (プログラム上は 10 個以下) の単語 ID が対応するが、1 つの単語 ID が複数の単語クラスに属することは許さない。単語クラスには内部で 2 から順に通し番号をふる。

つぎに、入力テキストファイルを最初の行から順に読み込み、各単語クラスの情報と照合し、すべてのフィールドが一致する単語クラスを探す。照合の際、単語テーブルファイルで "*" と書かれているフィールドは、常にマッチするものとして扱う。どの単語クラスにも一致しなければ、単語クラス 1 とする。文頭・文末 (入力テキストで "#" の行) は単語クラス 0 とする。

このように、入力テキストの各行に対して単語クラスの番号を割り当て、その unigram、bigram の出現回数をカウントする。入力をすべて読み終えた時点で、それぞれの出現回数をもとに、unigram、bigram の出現確率を計算する。-w オプションが指定されたときは、単語クラス bigram の出現確率を、対応する単語 ID の情報とともに出力ファイルに書き込む。また、bigram の出現確率をもとに、その perplexity を計算する。perplexity の値は -v オプションが指定されたときのみ表示する。

引数の説明

-2	bigram の出現確率を求める。
-t table_file	bigram を計算する対象の単語リストのファイル名を指定する。
-c	unigram、bigram の出現頻度を表示する。
-v	(verbose)
-w outfile	bigram のデータを出力するファイル名を指定する。

4.5 bigram データファイル表示

コマンド名

```
print_bi
```

形式

```
print_bi [-d dicfile] bigram_file
```

機能

calc_ngram の出力として得られる付属語 bigram あるいは自立語 bigram のデータファイルの内容を表示する。

-d オプションで辞書ファイル (SSS-LR の文法で用いているもの) が指定されたときは、単語 ID に対応する単語の情報も合わせて表示する。

4.6 付属語 bigram、自立語 bigram を用いた文節ラティスからの文候補の選択

コマンド名

```
ph_ngram
```

形式

```
ph_ngram [-lt fuzoku_bigram] [-lh jiritsu_bigram ] [-w weight] [-dt nt] [-dh nh]
```

機能

SSS-LR の文節認識結果を読み込み、各文節の候補の組合せとして得られる文認識結果のなかで、SSS-LR による音響的スコアと付属語 bigram、自立語 bigram による言語敵スコアの重みつきの和が最大になるものを動的計画法を用いて求める。付属語 bigram、自立語 bigram は、一方のみあるいは両方指定可能。音響モデルのスコア (SSS-LR の出力の $-1/10000$) に対する言語モデルによるスコアの重みは -w オプションで指定する (スコアは尤度あるいは確率の対数値)。出力は SSS-LR の出力 (ph_ngram の入力) と同形式。

処理の概要

標準入力から SSS-LR の文節認識結果を読み込み、付属語 bigram `fuzoku_bigram`、自立語 bigram `jiritsu_bigram` の確率値を加味した順位に候補を並べ替えて出力する。候補文節の最初の 1 単語を自立語、最後の 1 単語を付属語とみなす。1 単語のみからなる文節については、その単語を自立語兼付属語とみなす。文節認識結果の単語の情報は、入力の `#1` の行から取り出す。`#1` の行から得

られる単語 ID をもとに、付属語 bigram あるいは自立語 bigram のデータファイルを検索し、隣合う文節について候補間の bigram の確率を求める。詳細は、文献 [1] 参照。

本アルゴリズムの出力では最適な文節候補系列しか意味を持たないが、他の候補も残すため、各文節の 2 位以下の候補は最適候補に選ばれたものを除いてもとの順位を保つようにしている。また、各候補のスコアは、もとの同順位の候補のスコアをそのまま使用している。

引数の説明

-lt fuzoku_bigram	付属語 bigram のデータファイル。calc_ngram の出力として得られる形式。
-lh jiritsu_bigram	自立語 bigram のデータファイル。calc_ngram の出力として得られる形式。
-w weight	音響モデルのスコアに対する言語モデルのスコアの重み。省略時 1.0。
-dt nt	付属語 bigram について、対象外の単語に対する出現確率を $1/nt$ 倍する。省略時 $nt = 1$。
-dh nh	自立語 bigram について、対象外の単語に対する出現確率を $1/nh$ 倍する。省略時 $nh = 1$。

4.7 認識率集計

コマンド名

`score`

形式

`score [-E] [-{k|j|r|o}] [-s] [-1] result [correct]`

機能

正解文字列ファイル correct を指定すると、SSS-LR の出力形式のファイル result を読み込んで、文節および文の 1 位認識率を計算して表示する。正解文字列ファイルが指定されない場合は、各文節の 1 位候補を連結した文認識結果を文節区切りつきで出力する。正解文字列ファイルの形式はサンプルを参照。いずれの場合も、result で "-" を指定すると標準入力から読む。

引数の説明

- -E 認識誤りを出力 (correct 指定時のみ)
- -k result として #k の情報を用いる
- -j result として #j の情報を用いる
- -r result として #r の情報を用いる
- -o result として #o の情報を用いる
  (default は -k)
- -s (doo)iu ↔ (doo)juu を区別する (default では区別しない)
- -1 文節 / 文 認識率を 1 行で表示

5 オペレーション例

付属のサンプルデータを用いたオペレーション例を説明する。なお、各コマンドはコマンドサーチパスに入っているものとする。また、サンプルデータはディレクトリ sample に入っているので、それを作業ディレクトリにコピーして使用する。ディレクトリ sample に含まれるファイルは以下の通りである。

text.sample	テキストファイル
dic.sample	認識用単語辞書
fwbi.sample	付属語 bigram データ
cwbi.sample	自立語 bigram データ
SSSout.sample	SSS による文節認識結果
correct.sample	認識の正解ローマ字文字列

なお、fwbi.sample, cwbi.sample は、text.sample とは別の大量のテキストデータから学習したものである。

1. 付属語の取り出し

```
% fuzoku text.sample > fuzoku.out
```

2. 付属語単語テーブルの作成

```
% grep -v '^#' fuzoku.out | sort | uniq -c > fuzoku.lst
% mkwtab dic.sample fuzoku.lst > fwtab
```

3. 付属語 bigram の計算

```
% calc_ngram -2 -t fwtab -w fwbi fuzoku.out
% print_bi -d dic.sample fwbi （データの内容の確認）
```

4. 同様の手順で、自立語 bigram を計算する

```
% contents text.sample > contents.out
% grep -v '^#' contents.out | sort | uniq -c > contents.lst
% mkwtab dic.sample contents.lst > cwtab
% calc_ngram -2 -t cwtab -w cwbi contents.out
% print_bi -d dic.sample cwbi （データの内容の確認）
```

5. 付属語 bigram、自立語 bigram を用いた文節ラティスからの文候補の選択
大量のテキストデータから学習した付属語の bigram fwbi.sample、自立語の bigram cwbi.sample を用いて、SSS-LR により文節発声の文を認識した結果の文節ラティス SSSout.sample から文候補の選択を行なう。

```
《付属語 bigram のみ》
% ph_ngram -lt fwbi.sample -w 1.0e-05 -n < SSSout.sample > res.f
《自立語 bigram のみ》
% ph_ngram -lh cwbi.sample -dh 1000 -w 1.0e-05 -n < SSSout.sample > res.c
```

《付属語 bigram + 自立語 bigram》

```
% ph_ngram -lt fwbi.sample -lh cwbi.sample -dh 1000 -w 1.0e-05 -n \
< SSSout.sample > res.fc
```

6. 認識結果の集計

正解ローマ字文字列ファイル correct.sample を用いて、文節認識率、文認識率を集計する。

《言語モデルを用いない場合》

```
% score -r SSSout.sample correct.sample
phrase correct:    42 /  49 (85.7%)
sentence correct:   9 /  16 (56.3%)
```

《付属語 bigram のみを用いた場合》

```
% score -r res.f correct.sample
phrase correct:    45 /  49 (91.8%)
sentence correct:  12 /  16 (75.0%)
```

《自立語 bigram のみを用いた場合》

```
% score -r res.c correct.sample
phrase correct:    46 /  49 (93.9%)
sentence correct:  13 /  16 (81.3%)
```

《付属語 bigram、自立語 bigram を併用した場合》

```
% score -r res.fc correct.sample
phrase correct:    49 /  49 (100.0%)
sentence correct:  16 /  16 (100.0%)
```

参考文献

[1] 磯谷亮輔, 嵯峨山茂樹, 粟津辰功: "付属語の N-gram、自立語の N-gram を用いた音声認識," 日本音響学会講演論文集, (1993.3 発表予定).