2008－SLP－72　(7)
2008／7／18

音声認識のための言語処理：何が足りないか？

森 信介
京都大学学術情報メディアセンター
〒606-8501 京都市左京区吉田本町
forest@i.kyoto-u.ac.jp

あらまし
言語モデルとしてよく用いられる単語 n-gram モデルは、適切に単語に分割された対象分野のコーパスが大量に利用可能な状況では実用上十分な性能を実現できる。しかし、対象分野の言語資源としては、単語境界情報のない生コーパスだけが利用可能であったり、基準に一貫性のない単語 (表記) の一覧が与えらえるのみであるという状況が現実的にはほとんどである。本稿では、一般的な分野の単語分割済みコーパスとそこに出現する単語の発音辞書があり、ある程度の量の対象分野の生コーパスが利用可能であるとの前提の下、音声認識のための言語モデルを効率よく構築することを目的として、現在利用可能な言語処理技術について述べるとともに、望まれる言語処理技術について考察する。

キーワード　言語モデル　確率的単語分割　読み推定　部分修正コーパス

Language Processing for Speech Recognition: What is Missing?

Shinsuke MORI
Academic Center for Computing and Media Studies
Sakyo-ku, Kyoto 606-8501, JAPAN
forest@i.kyoto-u.ac.jp

Abstract
A word n-gram model, which is often used as a language model, functions sufficiently well for practical uses under the condition that a large corpus in the target domain with word boundary information is available. In many practial cases, however, only a raw corpus without word boundary information is available, or a list of spellings of words and word sequences is given. In this article, assuming that we have a general corpus with word boundary information, a word list with pronunciation appearing in that corpus, and a certain amount of a raw corpus in the target domain, we discuss current natural language processing technologies and missing ones.

Key Words　Language Modeling, Stochastic Segmentation, Pronunciation Estimation, Partially Annotated Corpus

1　はじめに

音声認識に代表される生成的な課題に用いられる言語モデルの役割は、対象分野の単語列の出現傾向を適切にモデル化することである。言語モデルとしてよく用いられる単語 n-gram モデルは、単語列の頻度に基づいており、適切に単語に分割された対象分野の例文 (コーパス) が大量に利用可能な状況下において実用上十分な性能を実現できる。さらに、音声認識に応用する場合は、コーパスの約 95%以上を被覆する単語に対して文脈に応じた適切な発音が付与されていれば、言語モデルとしては実用上十分である。

しかしながら、多くの場合には単語境界情報のない生コーパスと、単語列[1] とその発音のリストからなる対象分野の辞書のみが利用可能であったり、より悪い状況では、生コーパスのみが利用可能である。

本稿では、一般的な分野の単語分割済みコーパスとそこに出現する単語の発音辞書があり、ある程度の量の対象分野の生コーパスが最低限利用可能であるとの前提の下、音声認識のための言語モデルを効率よく構築することを目的として、現在利用可能な言語処理技術について述べるとともに、望まれる言語処理技術について考察する。

2　言語モデルの分野適応

前節で述べた前提で、ある対象分野の言語モデルを構築する一般的な手順は、以下の通りである (図 1 参照)。

1. 対象分野の例文の自動単語分割
2. 認識語彙の選択
3. 読みの付与
4. 単語 n-gram 頻度の計数

本節では、この手順における単語分割と読み付与の自動化について検討する。

2.1　自動単語分割

日本語は文中の単語区切りが自明ではないので、対象分野の例文における単語単位の確定が言語モデル構築の最初の処理である。これには、茶筌 [1] などのコーパスに基づく形態素解析器 [2, 3] が用いられることが多い。読みや品詞は必ずしも必要ではないので、本稿ではこの処理をより一般的に自動単語分割と呼ぶ。

自動単語分割の最大の問題は、多くの場合に対象分野の文に対する分割精度が高くないことである。この際に取り得る方法は概ね以下の 3 つである。

[1] 既存の単語分割済みコーパスや自動単語分割器の単語の定義では、単語でなく単語列 (複合語) が与えられることが多い。

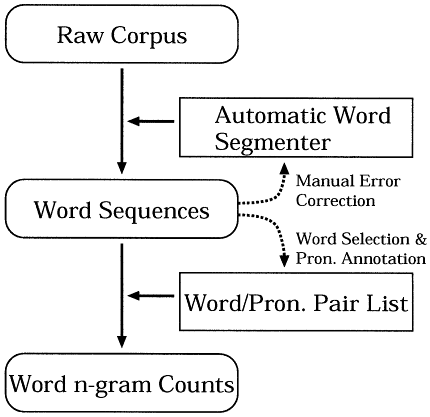


図 1: 言語モデルの分野適応の手順

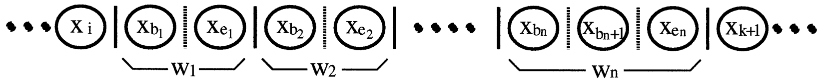
1. 自動単語分割の結果をそのまま用いる
2. 自動単語分割の結果から対象分野特有の単語を人手で抽出し自動単語分割器に追加する
3. 自動単語分割の結果を人手で修正し、自動単語分割器を再構築する

必要となるコストはこの順に大きくなるが、自動分割結果から推定される言語モデルの性能はこの順に高くなる。さらに、単語の追加や単語分割結果の人手による修正に関しては、作業の対象とする量に応じてコストが大きくなるとともに言語モデルの性能が高くなる。作業箇所は任意に決定できるので、低いコストでより大きい性能向上を得るための工夫の余地がある。

現実的によく行われる対処方法は自動単語分割器への単語の追加である。具体的には、自動分割の出力から既知語を除去することで未知語候補を集収し、これらから適切な単語を人手で選別する。このようにして得られる対象分野特有の単語を自動単語分割器に追加し、改めて自動単語分割を行う[2] 。

自動単語分割器をコーパスから構築することができる場合には、数千文程度であっても対象分野の文を人手で単語分割し、自動単語分割器を再構築することが望ましい。これは、言語モデルの性能向上に大きく寄与する (図 1 における Manual Error Correction)。この際に、対象分野特有の単語を含む文の正しい単語分割結果を加えると非常に効果が大きい。問題は、自動単語分割器の元の学習コーパスと同じ基準で単語に分割し、場合によっては品詞と読みも付与することが容易ではないことである。多くの自動単語分割器の学習データの最小単位は文であるため、助詞や助

[2] 品詞が必要な場合には一般名詞とする。



$$f_r(w_1^n) = P_i \times (1-P_{b_1}) \times P_{e_1} \times (1-P_{b_2}) \times P_{e_2} \times \cdots \times (1-P_{b_n}) \times (1-P_{b_{n+1}}) \times P_{e_n}$$

図 2: 確率的単語分割コーパスにおける単語 n-gram 頻度

動詞の列が必然的に含まれる。このような箇所に対する修正作業は困難である割には言語モデルの性能向上にほとんど寄与しない。

3 確率的単語分割

前節で述べた自動単語分割器による誤分割の問題を軽減する方法として、確率的単語分割 [4] という概念を提案した。確率的単語分割を用いることで、分野適応に際して人手を全くかけない場合でも、自動単語分割器の結果をそのまま用いる決定的単語分割よりもよい言語モデルが構築できる。

3.1 確率的単語分割コーパス

確率的単語分割コーパスは、文字間に単語境界が存在する確率を推定する単語境界モデルを用いて、生コーパスの各文字間に単語境界確率を付与することで得られる。単語境界モデルは、一般的な分野の単語分割済みコーパスから構築しておく。

確率的単語分割コーパスにおける単語 n-gram 確率の値は、それを構成する文字列のコーパスにおける全ての出現位置での期待頻度 (図 2 参照) の合計から計算される (詳しくは文献 [4] 参照)。

これにより、自動単語分割器が分割を誤る単語列であっても、それを構成する文字列の頻度がある程度高い場合には、ゼロより十分大きい頻度が付与され、決定的単語分割よりも予測力が高い言語モデルが構築できる。

3.2 疑似確率的単語分割コーパス

確率的単語分割コーパスに対する単語 n-gram 頻度は、高いコストの計算を要し、従来の言語モデル構築のツールを直接用いることができない。この問題を容易に回避する方法として、単語分割済みコーパスで確率的単語分割コーパスを近似する方法がある。具体的には、確率的単語分割コーパスの各文字間に対して 0 から 1 の間の乱数を発生させ、単語境界確率との大小関係に応じて単語境界か否かを決定する。

表 1: 対象分野のコーパスの単語分割の方法と言語モデルの予測力の関係

適応分野の学習コーパス	倍率	パープレキシティ
決定的単語分割	(×1)	54.28
疑似確率的単語分割	×1	49.45
	×4	46.12
	×16	44.65
	×64	44.24
	×256	43.86
確率的単語分割	–	43.36

これにより、確率的単語分割コーパスに近い単語分割済みコーパスを得ることができる。これを疑似確率的単語分割コーパスと呼ぶ。近似により生じる誤差を軽減するために、M 個の疑似確率的単語分割コーパスを生成し、これらを単語 n-gram 頻度の計数の対象とする。この M を倍率と呼ぶ。

3.3 予測力の比較

単語に分割された辞典の例文 (単語分割済みコーパス; 14,754 文) と単語に分割されていない医療分野のテキスト (生コーパス; 53,915 文) を所与とし、医療分野の言語モデルを構築するという課題で、決定的単語分割と確率的単語分割と疑似確率的単語分割とを比較した。

単語境界モデルは、各文字間に対して前後最大 3 文字を素性とし、その文字間が単語境界である確率を推定する最大エントロピーモデル[3] である。決定的単語分割では、単語境界モデルによって単語境界を決定した。精度は単語 3-gram モデルや CRF [5] を用いた場合と同程度である。

それぞれの方法で生成される医療分野のコーパスから推定した単語 2-gram モデルによる医療分野のテストセットパープレキシティの結果を表 1 に掲げる。この表から、確率的単語分割により決定的単語分割よりもよい言語モデル

[3] パラメータ推定には L-BFGS を使用し、Gaussian Prior を重みとした最大事後確率推定を行った。

が構築可能であり、16 倍程度の疑似確率的単語分割により確率的単語分割と同程度の予測力の言語モデルが従来のツールの枠内で構築可能であることが分かる[4]。

4 未知語の読み推定

単語を単位とする言語モデルを音声認識に用いるためには、語彙の各単語に発音が付与されている必要がある。適応対象の例文の単語分割結果を用いる場合には、高頻度の未知語 (文字列) を語彙に入れる必要があるので、これらに効率的に発音を付与する必要がある。

文字列から読みを推定する研究としては、機械翻訳 [7] [8] や音声合成のフロントエンド [9] や仮名漢字変換 [10] を目的として行われている。読みと発音 (特に実際の発音) はしばしば異なるが、この関係は別のモデル [11] により記述されるべきであると考え、本節では未知語の読み推定について考察する。

4.1 文字と読みの組の n-gram モデル

未知語の読み推定の方法として、文字と読みの組を単位とする n-gram モデルが提案されている [9]。パラメータ推定には、文字毎に読みが付与された単語 (列) の実例が大量に必要となる。これは、以下に例示するように、単語 (列) と読みの組を単漢字辞書を参照しながら文字と読みの対応を取ることで容易に作成できる (可能な対応関係が複数あることはほとんどない)。

銑鋼一貫製鉄/センコウイッカンセイテツ
⇒ 銑/セン 鋼/コウ 一/イッ 貫/カン 製/セイ 鉄/テツ

単語と読みの組は、多くの機械可読の辞書や単語毎に読みが付与された単語分割済みコーパス [12] から容易に得られる。

以上のようにして作成された文字と読みの組を単位とする n-gram モデルにより、未知語の文字列に対して、構成的に可能な全ての読みがその確率と共に得られる。未知語に対する最尤の読みの正解率は 80%以上である。

現在の問題は、単漢字辞書の不備である。具体的には、連濁や促音化、あるいは例外的な読み[5] に網羅的に対応していないことである。アラインメントが取れない単語と読みの組に対して差分を計算することで、追加すべき文字と読みの候補を出力することはできるが、単漢字辞書への追加にあたっては人手によるチェックが必要でる。

[4] 音声認識における評価は文献 [6] で報告されている。
[5] 例はそれぞれ、「会/ガイ 社/シャ」、「日/ニッ テ/テ レ/レ」、「振/フリ 込/コミ」である。

4.2 アルファベット列の読み

文字を単位とすることでは構成的に読めない未知語がある。英単語に代表されるアルファベット列などであり、これらに対しては、文字列と読みの組の n-gram モデルによって読み推定を行う方法が提案されている [8]。パラメータ推定には、適切な文字列に分解され、読みが付与された単語の実例が大量に必要となる。しかしながら、以下の例のように適切な文字列の定義は自明ではなく、慎重に設計する必要があろう。

$$\left\{\begin{array}{l}\text{gou/グ r/ル met/メ}\\ \text{gou/グ r/ル me/メ t/無音}\end{array}\right.$$

$$\left\{\begin{array}{l}\text{die/ダイ}\\ \text{di/ダイ e/無音}\end{array}\right. \quad \left\{\begin{array}{l}\text{fi/ファイ ne/ン}\\ \text{fi/ファイ n/ン e/無音}\end{array}\right.$$

英単語に関しては、読みが付与された単語の実例は比較的多く入手可能であり、文字列と読みの組が定義されれば EM アルゴリズムを用いて対応を取ることでパラメータ推定が可能である。観光案内などの対話システムなどにおいて必要となるであろうイタリア語やフランス語などの単語や専門用語 (例: cepstrum, Xeon) などは、それらの読みが分り、かつその傾向を考慮してモデル化の単位 (文字列) を決定できる作業者が必要である。

4.3 その他の難読文字列

恣意的に命名される商品名やグループ名 (例:「関ジャニ∞/カンジャニエイト」) などや話し言葉での呼称から乖離している正式な表記 (例:「ＵＳＤ／ＪＰＹ/ドルエン」) など、構成的に読むことが困難な文字列がある。これらは、画像や意味などを考慮する深い処理により対処可能であると思われるが、当面は人手で入力されるのを待ち、それを集収するのが現実的であろう [13]。

5 部分的なコーパスの修正

前節まで、確率的単語分割による生コーパスの利用方法と未知語の読み推定方法について述べた。これらを基礎として、適応対象のコーパスや単語リストやその読みが利用可能であるという前提での適応対象の音声認識のための言語モデルの効率的構築方法とそれを支援する言語処理やツールについて考察する。

5.1 単語単位の言語モデル

言語モデルの構築に先立って単語単位の確定が必要になるが、これには形態素解析器を流用することが多い。形態

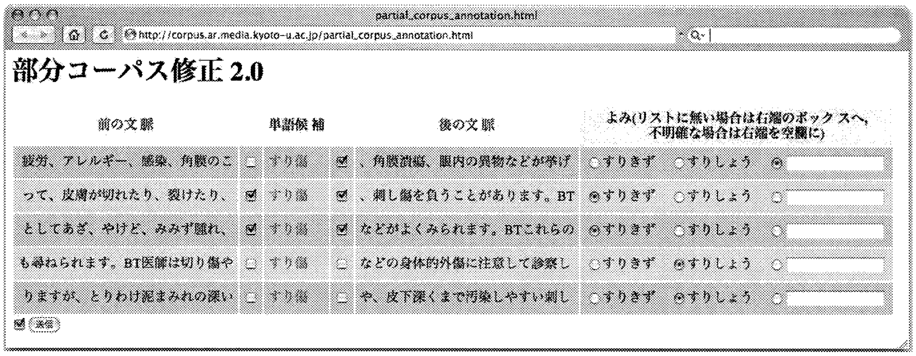


図 3: KWIC 形式のコーパスの部分修正ツール

素解析では、単語境界情報に加えて品詞や読みが付与されるので、単語 (表記) と品詞と読みの 3 つ組を言語モデルの単位とすることがしばしば行われている。正しく情報付与された同量のコーパスからパラメータを推定する場合は、単語を単位とするモデルよりも単語と品詞と読みの 3 つ組を単位とするモデルのほうが高い予測性能となる。しかしながら、コーパスの各単語に品詞と読みを適切に付与する作業は、高いコストを必要とする。さらに、品詞付与などの基準を作業者に徹底することは非常に困難である。一方で、第 2 節で述べたように、言語モデルの適応には、適応対象のコーパスを少量であっても適切に単位に分割し、自動単語分割器を再学習することが非常に効果的である。このことから、言語モデルの単位を単語 (表記のみ) とすることが適切である。

言語モデルの単位を単語とすることは、世界の主要言語の音声認識において一般的に行われていることであり、日本語の先行事例もある [14]。ごく稀に、単位を単語とすることに起因する誤り (例: … 聞/キ く/ク 夜/ヤ …) が生じることがあるが、学習コーパスを作成するコストや適応できる言語処理の範囲を考慮すると単語を単位とすることは合理的である。

単語の定義には、短い単位 [15] を用いるのがよいと考える。短い単位に基づく言語モデルの利点は、小さい語彙で高いカバー率を実現することができることである。平均単語長が短くなると、単語 n-gram モデルが参照する履歴が短くなり、予測精度が低下する。この問題には、予測に有用な単語列を選択的に語彙に追加することで、既存の単語 3-gram モデルの実装が利用可能な範囲で、予測精度の向上を図ることができる [16]。

5.2 コーパスの部分的修正

単語を単位とすることでコーパス修正のコストは大きく低減されるが、助詞や助動詞などが連続する箇所を基準に従って適切に単語に分割することは容易ではない。そのような箇所は、あまり分野に依存しないので、分野適応の観点からは、対象分野特有の単語の周辺に修正コストを集中するのがよい。分野特有の単語リストが与えられる場合、単語数の割合にしてコーパス全体の 5%程度の修正で、単語数にして 45%を文単位で修正する場合と同程度の予測力が実現できる例がある [17]。音声認識には読みが必要なので、この文献で提案されている修正インターフェイスを拡張し、図 3 のように読みの推定結果や辞書検索結果を付与し、それを同時にチェックする仕組みを構築するのがよいと考えられる。

対象分野特有の単語リストが与えられない場合も十分考えられるので、コーパスから未知語候補を高精度で抽出することも重要な課題である [18, 19, 20]。ある単語の修正作業の結果により、周辺文字列が単語候補であるか否かの判断が変ることがあり得るので、作業の対象とすべき単語候補を逐次的に計算し提示する能動学習のような枠組みが望ましい。

5.3 不完全な情報からの自動単語分割器の学習

前述の文献 [17] の研究では、自動単語分割器の再構築を行っていない。これを行い、修正していない箇所を再分割することでさらなる精度向上が実現可能と考えられる。

図 3 のコーパス修正インターフェイスの出力として得られるコーパスには、部分的に単語境界か否かの情報が付与されている。また、多くの専門用語辞書に記載されている

単語の多くは、短単位の単語定義から見れば複合語 (単語列) である。自動単語分割器は、これらの不完全な情報からの学習を許容する必要がある。第3節で述べた点推定の単語境界モデルは、これが容易に実現できる。CRF のような系列として推定するモデルでも、不完全な情報からの学習の研究がなされている[6] [22]。

これらの自動単語分割器を拡張し、専門用語辞書に記載されいてる単語列や、さらには未知語に対応した仮名漢字変換システムや音声認識システムを運用することで獲得される未知語候補 [23, 24] を参照し単語分割の精度を向上することで、言語モデルの性能が向上すると考えられる。

6 おわりに

本稿では、対象分野の生コーパスが利用可能であるとの前提で、音声認識のための言語モデル構築のための言語処理について述べた。コーパスを単語に分割したり、単語に読みを付与する作業者を効率的に活用するために、コーパスに情報を断片的に付与することを容認する枠組みが現実的である。また、音声認識の結果や仮名漢字変換などの運用によって得られる不確実な情報を用いることも単語分割や読み推定の精度向上に寄与する可能性がある。これらの断片的あるいは不確実な情報を最大限に活用する言語処理システムを構築することでさらなる音声認識の精度向上が図れる。

参考文献

[1] 松本裕治. 形態素解析システム「茶筌」. 情報処理, Vol. 41, No. 11, pp. 1208–1214, 1996.

[2] 永田昌明. 統計的言語モデルと n-best 探索を用いた日本語形態素解析法. 情処論, Vol. 40, No. 9, pp. 3420–3431, 1999.

[3] 森信介, 長尾眞. 形態素クラスタリングによる形態素解析精度の向上. 自然言語処理, Vol. 5, No. 2, pp. 75–103, 1998.

[4] 森信介, 宅間大介, 倉田岳人. 確率的単語分割コーパスからの単語 n-gram 確率の計算. 情処論, Vol. 48, pp. 892–899, 2007.

[5] 工藤拓, 山本薫, 松本裕治. Conditional random fields を用いた日本語形態素解析. 情報処理学会研究報告, 第 NL161 巻, 2004.

[6] 倉田岳人, 森信介, 西村雅史. 講義関連コーパスを利用した音声認識システムの自動適応. 電子情報通信学会論文誌, Vol. J90-D, No. 9, pp. 1780–1789, 2005.

[6] CRF による自動単語分割器は単語境界確率を計算することも可能である [21]。

[7] Kevin Knight and Jonathan Graehl. Machine transliteration. *Computational Linguistics*, Vol. 24, pp. 599–612, 1998.

[8] 齋藤邦子, 篠原章夫, 永田昌明, 小原永. 音声制御ブラウザ VCWeb の英日シームレス化. 人知誌, Vol. 17, No. 3, pp. 343–347, 2002.

[9] 長野徹, 森信介, 西村雅史. N-gram モデルを用いた音声合成のための読み及びアクセントの同時推定. 情処論, Vol. 46, , 2006.

[10] 森信介. 無限語彙の仮名漢字変換. 情処論, Vol. 48, pp. 3532–3540, 2007.

[11] 秋田祐哉, 河原達也. 話し言葉音声認識のための汎用的な統計的発音変動モデル. 電子情報通信学会論文誌, Vol. J88-DII, No. 9, pp. 1780–1789, 2005.

[12] 日本電子化辞書研究所. EDR 電子化辞書仕様説明書, 1993.

[13] 中野鐵兵, 佐々木浩, 藤江真也, 小林哲則. 集合知を利用した語彙情報の収集・共有・管理システム. 情報処理学会研究報告, 第 SLP71 巻, 2008.

[14] 伊東伸泰, 西村雅史, 荻野紫穂, 山崎一孝. 単語単位による日本語言語モデルの検討. 自然言語処理, Vol. 6, No. 2, pp. 9–28, 1999.

[15] 小椋秀樹, 小磯花絵, 冨士池優美, 原裕. 『現代日本語書き言葉均衡コーパス』形態論情報規程集. 独立行政法人国立国語研究所, 2008.

[16] 森信介, 山地治, 長尾真. 予測単位の変更による n-gram モデルの改善. 情報処理学会研究報告, 第 SLP19 巻, pp. 87–94, 1997.

[17] 森信介. 単語リストと生コーパスによる確率的言語モデルの分野適応. 自然言語処理, Vol. 13, No. 4, pp. 33–48, 2006.

[18] 森信介, 長尾眞. n グラム統計によるコーパスからの未知語抽出. 情処論, Vol. 39, No. 7, 1998.

[19] Haodi Feng, Kang Chen, Xiaotie Deng, and Weimin Zheng. Accessor variety criteria for chinese word extraction. *Computational Linguistics*, Vol. 30, No. 1, pp. 75–93, 2004.

[20] 永田昌明. 単語頻度の期待値に基づく未知語の自動収集. 情報処理学会研究報告, 第 NL-116 巻, 1996.

[21] 岡野原大輔, 工藤拓, 森信介. 形態素周辺確率を用いた確率的単語分割コーパスの構築とその応用. NLP 若手の会 第1回シンポジウム, 2006.

[22] Yuta Tsuboi, Hisashi Kashima, Shinsuke Mori, Hiroki Oda, and Yuji Matsumoto. Training conditional random fields using incomplete annotations. *To appear in Proc. of the COLING 2008*, 2008.

[23] 森信介, 小田裕樹. 自動未知語獲得による仮名漢字変換システムの精度向上. 言語処理学会年次大会, 2007.

[24] 笹田鉄郎, 森信介, 河原達也. 音声とテキストからの語彙獲得による読み推定精度の向上. 言語処理学会年次大会, 2008.