2008－SLP－72 (8)
2008／7／18

統計的言語モデル：何が問題なのか？

伊藤 彰則

東北大学大学院工学研究科

統計的言語モデルは，連続音声認識のための言語モデルとして広く用いられている．特に back-off n-gram は連続音声認識の言語モデルの標準といっても良い．一方，n-gram を超える言語モデルは多数提案されてきたが，n-gram(特に trigram) を大きく凌ぐ言語モデルはまだ発見されていない．本稿では，n-gram が持つ問題点，それに対するいくつかの提案について述べ，現在の統計的言語モデルに足りないもの，それを補うかもしれないものについて議論する．

Statistical Language Modeling and Its Problems

Akinori Ito

Graduate School of Engineering, Tohoku University

Statistical language models are widely used as language models for large vocabulary continuous speech recognition. Above all, a back-off n-gram is a *de facto* standard as a language model for speech recognition. Number of models have been proposed so far for overcoming the back-off n-gram, but none of them has achieved large improvement over the back-off trigram. In this paper, various language models are briefly reviewed, and I give some suggestions what is needed for current language models, and discuss possibilities of improving language models.

1 はじめに

統計的言語モデルは，連続音声認識システムの重要な要素のひとつである．特に n-gram は，ほぼ全ての大語彙連続音声認識システムに利用されており，単純かつ強力な優れたモデルである．しかし，同時にその弱点も古くから指摘されており，n-gram を改善するための手法が多数提案されているが [1]，全面的に従来の n-gram に取って代わるには至っていない．本稿では，n-gram を中心とした統計的言語モデルの問題点と，それを克服するための方向について考えてみたい．

2 N-gram モデルとその問題点

2.1 バックオフ n-gram モデル

N-gram は，ある単語の出現確率がその直前の $n-1$ 個の単語のみによって条件付けられるというモデルである．すなわち，

$$P(w_1,\ldots,w_N)=\prod_{i=1}^{N}P(w_i|w_{i-n+1}\ldots w_{i-1}) \quad (1)$$

である．確率の推定に用いられる方法は，基本的には学習サンプルに基づく最尤推定であるが，ゼロ頻度問題を解決するためにバックオフ平滑化 [2] が用いられる．バックオフ平滑化は，単語の n 個組の学習データ内での出現頻度が 0 であったときに，その n-gram 確率を単語の $(n-1)$ 個組の頻度から推定する手法であり，他の平滑化手法 (削除補間法など) と比較して高い性能が得られているといわれる．

2.2 N-gram の問題点

N-gram には明らかな問題点がいくつかある [1]．一つ目はその制約の局所性である．例えば trigram は直前 2 単語にしか依存しないので，それ以前にどんな単語が出現しているのかを考慮することができない．二つ目は，一つ目とも関連するが，言語の構造をまったく考慮していないことである．ある単語の出現には，その直前の単語だけでなく，構造的に関連するもっと前の単語の影響があるはずであるが，n-gram ではそれが考慮できない．三つ目は，確率推定のドメイン依存性である．単語 trigram は学習コーパスに強く依存する．その結果，学習コーパスとは違う話題やスタイルの音声に対しては劇的に性能が低下する．

これらの問題を解決するため，多くの研究が行われた．N-gram の局所性を補償するために，キャッシュモデル [3] や可変長 n-gram[4] が提案された．

また，構造を考慮するために，Structured language model[5] や依存言語モデル [6]，確率文脈自由文法 [7] などが提案された．学習コーパスへの依存性に関しては，多くの言語モデル適応手法 [8] が提案されている．また，PLSA[9] や LDA[10] のように，文書単位での制約を反映するモデルの導入は，大域的制約と言語モデル適応の両方を同時に行う方法だといえる．

これらのうち，n-gram のコンテキストを長くする手法は，限定的な改善しか得られていない．長い n-gram は，ある決まった言い回しが頻出するような特殊な場合にしか有効でない [11]．また，言語モデルに構造を導入する手法は，少なくとも音声認識のための言語モデルとしては成功していない．その理由の一つは，構造を導入したことによって改善できる誤りの数が少ないことによると筆者は考えている．言語モデル適応はある程度の成功を収めているといってよいと思われるが，適応の方法については決定的な方法はない．N-gram カウント混合 (MAP 推定) に基づく方法 [12] は手軽であるが，パラメータを適切に推定することが難しい．最大エントロピー法に基づく適応 [13] は柔軟であり，最適化手法も存在するが，適応時および確率計算時の計算量が多いため，デコーダに組み込んで使うことは難しい．

バックオフ平滑化を改良する試みは一時期盛んに行われたが，結局 Good-Turing 法 [2], Witten-Bell 法 [14, 15], Kneser-Nay 法 [16] あたりで落ち着いており，どれを使ってもそう大きな差はないとされる．しかし，そもそもバックオフが起きている部分で致命的な誤りが発生するという指摘もあり [17]，バックオフ平滑化が本当に良いのかどうかについて再検討が必要に思われる．

2.3 さらに問題点

上記の問題点は n-gram に特有の問題であったが，もっと広く「音声認識のための言語モデル」に関する問題点を二つ挙げたいと思う．一つは、学習データの問題である．言語モデル適応技術により，少量の学習データであっても有効に利用できるようになってきてはいるが，その「少量の学習データをどこから手に入れるのかという問題がある．これは言語モデルの問題というよりも，言語モデルを利用するシステムとその開発に関する問題であるといえる．二つ目は，未知語 (未登録語) の問題である．未知語を検出すること自体が難しいが，その言語確率を推定することも一般には困難である．

3 言語モデル改良の方向性

これまで述べたさまざまな問題に対して，言語モデルをさらに改良するための方向性について考えてみたいが，そもそも「それがわかれば苦労はない」わけで，まるで見当はずれの可能性も十分ある．しかし，ここではそれを恐れずに，思うことを書いてみたい．

3.1 欲しいデータはどこにあるのか

まず学習データの問題である．学習データがなければ話が始まらないが，少なくとも現在の言語モデルでは「広く大量に言語データを集めれば，どんな内容でも認識できる」とは考えられていない．さまざまな話題を含む巨大なコーパスを用いるよりも，これから認識しようとする特定の話題に関連する小/中規模のコーパスを使って言語モデルを学習したほうが，性能も上がるしモデルも小さくてすむ．

コーパスを集めるとすれば，まず情報源として考えられるのは World Wide Web である．初期段階では，広く Web からダウンロードするよりもニュースサイトなど特定のソースから集めたほうがよいとされていたが [18]，現在では一般の Web サイトからデータを集めることも一般に行われている [19]．しかし，ただ闇雲にデータを集めてモデルを学習しても得られるものは少ないので，「どんなモデルを作りたいのか」についてはっきりした目標を持つことが重要ではないかと思う．データ収集法の研究は，どこまで言語モデルの研究といっていいのか悩ましいところではあるが，今後ますます重要になるだろう．特に，集めたいデータをどう記述するかが問題である．現在は数個のキーワードを使うことが一般的であるが，利用可能なデータをデータ収集後にフィルタリングするためにはキーワードだけでは不足である．一方，従来の学習データ選択のように，集めたいデータを記述するために統計モデルを利用する [20] のは本末転倒である．人間が例文を作って与えたり，簡単な文法を書いたりすることが有効かもしれない．あるいは，ウェブ検索におけるコンテキストサーチ [21] のようなものがヒントになるのかもしれない．

3.2 私の知らない単語

未知語問題は音声言語処理にとって永遠のテーマである．初期の未知語処理では，認識語彙でカバーされていない未知語区間の検出を行っていた [22]．その後，未知語自体をモデル化して言語モデルに組み入れる手法が使われるようになった．この場合，未知語自体を独自のクラスとして n-gram に組み込み，そのクラスからの未知語音韻列の生成確率を利用するのが一般的である [23]．一方，未知語をモデル化するのではなく，他の情報ソースから語彙を取り入れることで未知語を減少させるというアプローチもある [24]．近年は Wikipedia やはてなダイアリーキーワードなどの情報源が Web 上に整備されてお

り，これらの内容を利用する研究も出てきている [25]．3.1 において Web を利用するアプローチも，言語モデル適応と同時に未知語を減少させる効果がある．

おそらく，新語・流行語・専門用語などについてはモデル化が難しいので，Web 上の情報源からそのつど取り入れて語彙を作り直すことが有効であろう．一方，人名や地名などの固有名詞は，まともに登録すると数が膨大すぎるという特徴があるので，モデル化によって縮退させる方法が有効であろう．

3.3 バックオフの限界

バックオフ n-gram は非常に強力な言語モデルであるが，実際にバックオフが起きたときには，計算される確率はコンテキストの制約を受けにくくなる．コンテキストの制約がゆるくなった部分で認識誤りが起きると，そこからの回復が難しい [17]．近年注目されている誤り訂正モデル [26, 27] が有効である理由はこの辺にあるのかもしれない．

ある単語の並びが学習データに出現しない場合でも，ある程度コンテキストの制約を保つ言語モデルが必要なのかもしれない．例えば，単語 n-gram の出現頻度が 0 の場合には品詞 n-gram にバックオフするという手法が提案されたことがある [28]．単語パープレキシティから見ると，通常のバックオフ n-gram と変わらないのでその後発展していないが，再検討してみる価値があるかもしれない．

確率計算の際にコンテキスト制約を失わないようにしても，結果的に単語正解制度という点ではほとんど変わらないかもしれない．しかし，連続音声認識でよく見られる「部分的に無意味な音節列のような認識結果になる」という誤りが改善できるかもしれない．そういう意味では，単に単語誤り率を減少させるのではなくて，「どのような誤りを減らしたいのか」という目標が必要なのかもしれない．

3.4 PLSA など

言語モデル研究の周辺でここ数年最も流行しているものといえば，PLSA[9] や LDA[10] などの適応手法であろう．最大エントロピー法によって特徴を取り入れた大局的制約のモデル化 [13] から PLSA のような「大まかな」モデルへの流行の移り変わりは，話者適応における VFS から MLLR への移行を連想させて興味深い．

その流れから言うと，PLSA のように「ある話題を話題空間の中の点として表現する」というモデル化のままで，さらにモデルを詳細化するというアプローチが自然に思いつく．PLSA における「話題」を，文字通りの意味における「話題」と「話者スタイル」や「発話スタイル」に分解する考え方は，すでに提案されている [29, 30]．そこからの発展として考えれば，PLSA を直交するいくつかの要因に自動的に分解して，分解された PLSA を個別に学習するという方法がありうるかもしれない．

4 まとめ

バックオフ n-gram，その他のこれまで提案された言語モデルについて概観し，現在の n-gram に足りないもの，それを補うヒントになるかもしれない方向性について議論した．放送ニュースや会議音声などの実環境音声を認識するには，現在の言語モデルでは不十分なことは確かであるが，どこがどう足りないのかを指摘することはなかなか難しく，現在も試行錯誤が続いている．この議論が次の言語モデル改良に少しでも役立てば幸いである．

参考文献

[1] Rosenfeld, R.: Two Decades of Statistical Language Modeling: Where Do We Go from Here?, *Proceedings of IEEE*, Vol. 88, pp. 1270–1278 (2000).

[2] Katz, S. M.: Estimation of probabilities from sparse data for the language model component of a speech recognizer, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 35, No. 3, pp. 400–401 (1987).

[3] Kuhn, R. and Mori, R. D.: A cache-based natural language model for speech recognition, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 12, No. 6, pp. 570–583 (1990).

[4] Kneser, R.: Statistical language modeling using a variable context length, *Proc. ICSLP*, Vol. 1, pp. 494–497 (1996).

[5] Chelba, C. and Jelinek, F.: Recognition performance of a structured language model, *Proc. Eurospeech*, Vol. 4, pp. 1567?–1570 (1999).

[6] Chelba, C., Engle, D., Jelinek, F., Jimenaz, V., Khudanpur, S., Mangu, L., Printz, H., Ristad, E., Rosenfeld, R., Stolke, A. and Wu, D.: Structure and performance of a dependency language model, *Proc. Eurospeech*, Vol. 5, pp. 2775?–2778 (1997).

[7] 堀　智織，加藤正治，伊藤彰則，好田正紀：音声認識のための確率文脈自由文法に基づく言語モデルの構築と評価，信学論 D-II, Vol. J83-D-II, No. 11, pp. 2407–2417 (2000).

[8] Bellegarda, J. R.: Statistical language model adaptation: review and perspectives, *Speech Communication*, Vol. 42, No. 1, pp. 93–109 (2004).

[9] Hofmann, T.: Unsupervised Learning by Probabilistic Latent Semantic Analysis, *Machine Learning*, Vol. 42, No. 1–2, pp. 177–196 (1999).

[10] Blei, D. M., Ng, A. Y. and Jordan, M. I.: Latent Dirichlet Allocation, *Journal of Machine Learning Research*, Vol. 3, pp. 993 – 1022 (2003).

[11] 加藤直人，浦谷則好，江原暉将，安藤彰男：ニュース音声認識のための (n ≧ 4)-gram を併用する言語モデル，信学論 D-II，Vol. J85-D-II, No. 6, pp. 967–975 (2002).

[12] 伊藤彰則，好田正紀：N-gram 出現回数の混合によるタスク適応の性能解析，信学論 D-II, Vol. J83-D-II, No. 11, pp. 2418–2427 (2000).

[13] Rosenfeld, R.: A maximum entropy approach to adaptive statistical language modeling, *Computer Speech and Language*, Vol. 10, pp. 187–228 (1996).

[14] Witten, I. H. and Bell, T. C.: The Zero-Frequency Problem : Estimating the Probabilities of Novel Events in Adaptive Text Compression, *IEEE Trans. Information Theory*, Vol. 37, p. 4 (1991).

[15] Placeway, P., Schwartz, R., Fung, P. and Nguyen, L.: The Estimation of Powerful Language Models from Small and Large Corpora, *Proc. ICASSP*, Vol. 2, pp. 33–36 (1993).

[16] Kneser, R. and Ney, H.: Improved backing-off for m-gram language modeling, *Proc. ICASSP*, Vol. 1, pp. 181–184 (1995).

[17] 浅見太一，野田喜昭，高橋　敏：ピットフォールエラーに着目した音声認識誤りの分析，音講論，pp. 53–54 (2008).

[18] Berger, A. and Miller, R.: Just-In-Time language modeling, *Proc. ICASSP*, Vol. II, pp. 705–708 (1998).

[19] Nisimura, R., Komatsu, K., Kuroda, Y., Nagatomo, K., Lee, A., Saruwatari, H. and Shikano, K.: Automatic N-gram Language Model Creation from Web Resources, *Proc. Eurospeech*, pp. 2127–2130 (2001).

[20] Bigi, B., Huang, Y. and Mori, R.: Vocabulary and Language Model Adaptation using Information Retrieval, *Proc. ICSLP*, pp. 602–605 (2004).

[21] Lawrence, S.: Context in Web Search, *IEEE Data Engineering Bulletin*, Vol. 23, No. 3, pp. 25–32 (2000).

[22] Hayamizu, S., Itou, K. and Tanaka, K.: Detection of Unknown Words in Large Vocabulary Speech Recognition, *Journal of Acoustical Society of Japan (E)*, Vol. 16, No. 3, pp. 165–171 (1995).

[23] Tanigaki, K., Yamamoto, H. and Sagisaka, Y.: A Hierarchical Language Model Incorporating Class-Dependent Word Models for OOV Words Recognition, *Proc. ICSLP*, Vol. 3, pp. 123–126 (2000).

[24] 小林彰夫，今井　亨，安藤彰男，中林克己：ニュース音声認識のための時期依存言語モデル，情報処理学会論文誌，Vol. 40, No. 4, pp. 1421–1429 (1999).

[25] 鈴田健太郎，西村竜一，河原英紀，入野俊夫：Web 知識を二段階利用した単語辞書更新手法，音講論，pp. 123–124 (2008).

[26] Roark, B., Saraclar, M. and Collins, M.: Corrective Language Modelling for Large Vocabulary ASR with the Perceptron Algorighm, *Proc. ICASSP*, Vol. 1, pp. 749–752 (2004).

[27] Oba, T., Hori, T. and Nakamura, A.: An Approach to Efficient Generation of High-Accuracy and Compact Error-Corrective Models for Speech Recognition, *Proc. Interspeech*, pp. 1753–1756 (2007).

[28] Niesler, T. R. and Woodland, P. C.: Combination of Word-based and Category-based Language Models, *Proc. ICSLP*, Vol. 1, pp. 220–223 (1996).

[29] Akita, Y. and Kawahara, T.: Language Model Adaptation Based on PLSA of Topics and Speakers for Automatic Transcription of Panel Discussions, *IEICE Trans. Inf. & Syst.*, Vol. E88-D, No. 3, pp. 439–445 (2005).

[30] 栗山直人，鈴木基之，伊藤彰則，牧野正三：情報量基準で語彙分割した PLSA 言語モデルによる話題・文型適応，情処研報 2006-SLP-64, Vol. 2006, No. 136, pp. 233–238 (2006).