日本語における口語体言語モデル

伊東 伸泰, 西村 雅史, 森 信介
日本アイ・ビー・エム 東京基礎研究所

1 はじめに

音声認識の分野では N-gram に基づく言語モデルが広く用いられているが、日本語音声認識システムの多くは新聞やニュース原稿などの「書き言葉」を学習データとして用いている [1]。著者らは講演や会議、対談といったものの書き起こしをアプリケーションとして読み上げではない自発的な発声 (Spontaneous speech) の認識システム構築を目指しているが、そのためにはいわゆる「話し言葉」に対応した言語モデルが必要である。文献 [2] で報告した言語モデルにおいては、口語体に近いデータとして、パソコン通信の電子会議室に投稿された文を学習データの一部として用いているが、これらは新聞、ビジネス文書などとは明らかに異なるスタイルで書かれているものの、あくまで「口語体風の」文であって、自発的な発話の「言い回し」に対しての検証は十分ではない。

旅行の申し込み、観光案内といったタスクでは日本語の対話音声とその書き起こしテキストが ATR、日本音響学会、JEIDA 等により提供されている [3] が、対象とする分野が限定されているため、講演・会議の書き起こしシステムにおける言語モデルのベースとすることは困難である。そこで著者らは自由発話の基礎データとして、「放送大学」に着目し、その書き起こしコーパスを作成した。本研究ではこの「話し言葉」コーパスを用いて、特に分野を限定しない自由発話用言語モデルを試みたので、その内容について報告する。

2 放送大学コーパス

'98 年度にテレビ放送された放送大学の授業から、のべ 148 回分 (45 分/回) を選択した後、書き起こし、単語単位への分割を行い、放送大学コーパスを作成した [4]。図 1 にその例[1]、表 1 にその緒元を示す。ただし、異なり単語数には不要語を含まない。間をつなぐための間投詞的表現 (つなぎ語) と、言い淀んだ単語の断片は、'< >' でタグ付けし、音を記述しており、後者については別途タグ (図中'd') を設けて区別している。さらに、長音化した語尾にもタグ付けした (i.e. を < オー >)。なお、用いているテキストの分割単位については文献 [2] を参照されたい。

表 1: 放送大学コーパスの緒元

科目数	78
話者数	97
総単語数	1,098,888
異なり単語数	23,929
不要語	99,419 (9.1%± 4.0)
Filler	8.75%
Fragment	0.35%

<エト> 、 私 も 、 この 式 を<オー> 、
<エー> に 参加 して 、 <アノ> d<カン> 参加
致し ました けれども 、 人々 の 、 <ソノー> 、
熱意 、 それから 日本 では 見られ なく
なった ような 、 <ソノ> 、 学校 に おいて ...

図 1: 書き起こしの例

ここで、不要語という用語の定義について、述べておきたい。Shriberg ら [10] によれば、自発的発話においては、本来意図した、正しい (fluent) 発声が、言語的に変形しているものを *Disfluency* と呼び、そのタイプを分類したところ、Filled pause、Repetition、Deletion の 3 つが約 85% を占めると指摘している。Filled pause はいわゆる Pause filler となるものであり、日本語では間投詞的表現中「エー」、「アー」といったものが相当するであろう。Repetition は隣接した繰り返し、Deletion は (正しい表現ならば存在すべき単語が) 発声されていないことを意味する。さらに Switchboard [11] における調査では Repetition や Deletion の約 25%で、単語を構成しない部分列が発声されており、それらを *Fragment* と呼んでいる。われわれの放送大学で不要語としてタグ付けしているものは、図 1 に示すように Filled pause を含むつなぎ語 (表中では Filler と記す) と Fragment であり、その他の Repetition や Deletion

[1] 知的所有権に抵触するのを避けるため一部内容語を変更している。

については、今後の課題としたい[2]。

3 自由発話のための言語モデル

理想的には、自由発話を書き起こしたコーパスが十分あり、それから N-gram モデルを学習すればよいわけであるが、タスクを限定しない限りそれは現実的ではない。そこでより大きなコーパスから作成された「書き言葉」言語モデルと放送大学を組み合わせることにより、自由発話に対応することを考える。その際、主として学習すべき対象は以下の 2 つである。

- 不要語
- 分野に依存せず、講演を中心とする「話し言葉」に出現する単語

3.1 予備調査

これらのモデル、および学習を考えるための予備調査として、放送大学コーパスに出現する単語の種類を不要語とそれ以外の単語（以下「通常語」と呼ぶ）に分けて調べた。表 2 に示したのが、不要語を取り除いた場合の 75K 語彙（主に新聞・パソコン通信の書き込みから作成したもの）によるカバレージで、比較のために 2 つの新聞についても列挙している。これによれば本語彙は、放送大学コーパス中の単語を比較的よくカバーしており、新聞の場合と大きな差はない。一方不要語については、出現頻度順にとったときのカバレージを図 2 に示す。不要語の種類について調べた研究は過去にもいくつか存在するが、それらの報告と同様、少数の単語が多くを占めており、たとえば 24 個で約 90%となる。さらに書き起こしに揺れのあるものを同一化すれば、これらは 12 個に集約される。

表 2: 既存語彙 (75K) のカバレージ

放送大学	98.6%
日経新聞	99.6%
毎日新聞	98.7%

3.2 不要語モデル

Filled pause をはじめとする不要語について、日本語においてはそれを統計的なモデルとして扱い、コーパスから積極的に学習しようとする試みはほとんどない。従来は不要語が存在しないコーパスから、通常語

[2]これらが、言語理解のために「不要」かどうかも議論があるかもしれないが、ここでは発話の命題内容に影響を与えないという意味でこう呼ぶこととする。

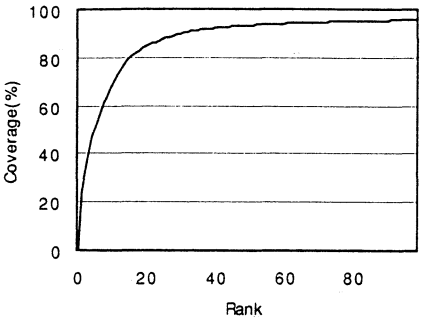

図 2: 不要語のカバレージ

の N-gram モデルを学習し、不要語は任意の位置に一様確率で出現するか [6] 、または文頭・読点といった位置に出現しやすいという観察を元に、当該状態を不要語と等価とみなしたり、その確率の一部を不要語に配分するといった手法が用いられている [7] [8]。

本研究では、本節の最初で述べたように、放送大学コーパスから学習したモデルをベースの言語モデルと組み合わせるわけであるが、ベースモデルは不要語がないデータから学習したものであり、不要語については、組み合わせることが予測精度を向上させるとは考え難い。そこでここでは、不要語と通常語で異なるモデルを参照する以下の手法を比較・評価する。

[モデル 1] (Dis1)

不要語予測時には放送大学コーパスから学習した Unigram 確率を用い、通常語予測時には、不要語をスキップし透過単語とした上で、放送大学およびベースの Trigram モデルを補間する。数式で表現すれば、以下のようになる。ただし V_Dは不要語集合、P_{au}、P_{base}はそれぞれ放送大学コーパスから学習した確率とベースの言語モデルから得られる確率、h は履歴、h_sは特に不要語をスキップした場合の履歴、そしてλは補間係数を表す。

$$P(w_n \mid h) = \begin{cases} P_{au}(w_n) & (\text{if } w_n \in V_D) \\ \lambda P_{au}(w_n \mid h_s) + & \\ \quad (1-\lambda)\delta_{\overline{D}} P_{base}(w_n \mid h_s) & (\text{if } w_n \notin V_D) \end{cases}$$

$$\delta_{\overline{D}} = 1 - \sum_{w_n \in V_D} P_{au}(w_n \mid h_s)$$

[モデル 2] (Dis2)

不要語予測時には放送大学コーパスから学習したTrigramを用い、その他はモデル1と同じとする。すなわち上式の不要語の場合を以下のように計算する。

$$P(w_n \mid h) = P_{au}(w_n \mid h_s) \quad (\text{if } w_n \in V_D)$$

[モデル3] (Dis3)
通常語予測時における$P_{au}()$項の履歴について、透過単語の場合と、透過単語としないBi-gramを補間して用いる(補間係数γ)。不要語予測時にはモデル2と同じとする。すなわち、通常語の場合の確率を以下の式で求める。

$$P(w_n \mid h) = \lambda P_{au}(w_n \mid h) + (1-\lambda)\delta_{\overline{D}} P_{base}(w_n \mid h_s) \ (\text{if } w_n \notin V_D)$$
ただし
$$P_{au}(w_n \mid h) = \gamma P_{au}(w_n \mid h_s) + (1-\gamma) P_{au}(w_n \mid w_{n-1})$$

3.3 スタイルの学習

3節の最初で、学習すべき単語として、分野に依存せず(講演調の)話し言葉に共通して現れる単語をあげた。これは言い替えればスタイルを学習することに他ならない。Seymore[9]は、語彙をGeneral、On-topic、Off-topicの3種類に分け、そのそれぞれで、異なるモデルを用いることを提案したが実験結果によれば、パープレキィティで線形補間より悪い。そこでわれわれは線形補間の利点を生かしながら、語彙種別に応じてより良い言語モデルを作成することを考えた。その基本は科目に依存せず出現する単語(不要語以外)をCommon単語(V_C)、科目により偏っている単語をTopic単語(V_T)とすると、Common単語は放送大学から学習したモデルの確率(P_{au})を、より信じるべきであるし、Topic単語は、より大きなコーパスから学習したベース言語モデルの確率(P_{base})の方がより信頼できるという仮定に基づいている。したがって補間係数(λ)を、単語種別により変更することになる。

$$P(w_n \mid h) = \begin{cases} \lambda_C P_{au}(w_n \mid h) + (1-\lambda_C)\delta_{\overline{D}} P_{base}(w_n \mid h) \\ \quad w_n \in V_C \\ \delta_T(\lambda_T P_{au}(w_n \mid h) + (1-\lambda_T)\delta_{\overline{D}} P_{base}(w_n \mid h)) \\ \quad w_n \in V_T \end{cases}$$

ただし、δ_Tは、確率を正規化するための係数であり、具体的には条件式

$$\sum_{w_n \in V_C, V_T} P(w_n \mid h) = 1$$

より、以下の式により計算される。

$$\delta_T(h) = \frac{1 - \sum_{w_n \in V_C} \lambda_C P_{au}(w_n \mid h) + (1-\lambda_C)\delta_{\overline{D}} P_{base}(w_n \mid h)}{\sum_{w_n \in V_T} (\lambda_T P_{au}(w_n \mid h) + (1-\lambda_T)\delta_{\overline{D}} P_{base}(w_n \mid h))}$$

ある単語wが分野(科目)tに依存する程度を判定する基準は、特に情報検索の分野で幅広く研究されており、相対エントロピー、χ^2値、相互情報量などがあるが、ここでは、加藤ら[7]が用いた相互情報量を採用した。

$$(1) \qquad I(T;w) = -\sum_t P(t) log P(t) + \sum_t P(t \mid w) log P(t \mid w)$$

4 実験

ベースとした語彙・言語モデルは[4]で報告を行ったものであるが、単語数は75K、学習用コーパスは98-99年度の新聞を中心に増やし、合計193M単語となっている。一方、放送大学コーパスは、全体を科目単位で無作為に10個のサブセットに分け、うち9セットをN-gramの学習・スムージングに、残り1セットをさらに2つに分け、各モデルの補間係数の推定用とテストデータとした。テストデータは、5つに講義(世界の教育、金融論、計測と制御、カオスの数理と技術、それに現代生物学の冒頭)各60文、計300文である。また、言語モデルの作成・評価にあたっては頻度の高い句読点の扱いが大きな影響を与えることが知られているが、放送大学コーパスは、学習・テストデータともすべての句読点を削除している[3]。

この9個のサブセットを各分野(t)として式(1)に基づき分野独立な単語を抽出した。不要語は、同学習セット中で頻度の高いもの30個を独立した単語として、その他出現した不要語はまとめて1クラスとして語彙に登録して学習した。通常語はベースの75Kのみで特に追加していない。

実験結果(各モデルに対するテストセットパープレキシティ)を表3に示す。ただし、AirUnivは放送大学のみで作成した言語モデル、Styleはスタイル学習、括弧

[3] ベースの言語モデルは学習データ全体の約20%で、句読点を削除している。

表 3: 実験結果

モデル	パープレキシティ
AirUniv	230.7
Baseline	189.3
Dis1	185.4
Dis2	182.4
Dis3	176.5
Dis3 + Style(293)	161.5
Dis3 + Style(1296)	159.1
Dis3 + Style(6018)	165.6

中の数字は Common (V_C) 単語として選択した数を意味する。未知語についてはカウントからはずしている。比較のための Baseline としてはベースの Tri-gram モデルの確率 (P_{base}) と放送大学から学習した Tri-gram モデルの確率 (P_{au}) を単純に線形補間した場合の結果を示した。まず Baseline (Tri-gram、単純に線形補間) と Dis2 (Tri-gram、透過単語) の結果から不要語を透過単語とした方がよいことがわかる[4]。Dis1 は、Dis2 と比較することによって不要語を Uni-gram とするか、あるいは履歴から予測する (Tri-gram) の方がよいかを検証するために行ったモデルである。結果から見ると、不要語といえどもコンテキストに依存して出現しているが、その差は比較的小さいとも言えよう。モデル Dis3 はモデル Dis2 の拡張である。やや、意外なことであるが、この拡張はパープレキシティを改善している。すなわち、不要語は、それに続く通常語の予測に役立つことがわかる。過去、英語では Shriberg、Stolcke ら [10] により、Switchboard コーパスを用いて Filled pause を中心とする Disfluency の統計的性質が、明らかにされ、「単純な clean-up モデル (本論文でいう透過単語、Dis2 に相当する) の拡張が必要」だと述べられているが、この Dis3 モデルは、その解の 1 つと考える。

スタイル学習はパープレキシティをさらに 15 程度改善した。

5 おわりに

不要語も含めて正確に書き起こしたコーパスの作成は著しくワークロードがかかるものであり、多量に用意することは不可能に近い。最もデータが豊富な英語でも、明らかな Spontaneous speech の書き起こしコーパスは Switchboard で 2M 単語、Callhome で約 0.15M 単語であり、ニュースや新聞に比べ明らかに少ない [11]。したがって、タスクを限定しない限り、十分な学習データを得ることは難しく、スタイルの適応化が必要であると思われる。

本研究では、実際の講演を書き起こした放送大学コーパスを用いて、口語体特有の不要語やスタイル語彙を学習することにより、「書き言葉から学習した」言語モデルを講演スタイルに適応化する手法について述べ、実験結果を示した。結果としてパープレキシティで単純な線形補間よりも、よい結果を得た。

毎日新聞社 (CD 毎日新聞 91-95)、日本経済新聞社、および放送大学に深謝する。

[4] 正確には、不要語予測において補間をおこなっていないことの差も存在する。

参考文献

[1] 伊藤他: 日本語ディクテーションのための言語資源・ツールの整備, 情処 音声言語情報処理研究会, SLP 26-1, pp. 31-38 (1999).

[2] 伊東他: 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. 9-27, (1999).

[3] 山本: 音声対話データベースの現状, 日本音響学会誌, pp. 797-802, (1998).

[4] 伊東, 西村: 口語体言語モデルのためのコーパス, 情処 自然言語処理研究会, NL-134-2, pp. 9-14, (1999).

[5] Stolcke, A., Shriberg, E.: Statistical Language Modeling for Speech Disfluencies, *Proc. of ICASSP 96*, pp. 405-408, (1996).

[6] 西村他: 放送音声の書き起こしに関する検討, 音響学会秋季全国大会, 1-R-14, pp. 145-146, (1998).

[7] 加藤他: 講演ディクテーションのための話題独立言語モデルと話題適応, 情処 音声言語情報処理研究会, SLP 26-2, pp. 9-16, (1999).

[8] Ohtsuki et al.: Improvements in Japanese Broadcast News Transcription, *Proc. of DARPA Broadcast News Workshop*, pp. 231-236, (1999).

[9] Seymore, K. et al.: Nonlinear Interpolation of Topic for Language Model Adaptation, *Proc. of ICSLP-98*, pp. 2503-2506, (1998).

[10] Shriberg, E., Stolcke, A.: Word Predictability After Hesitations: A Corpus-based Study, *Proc. of ICSLP-96*, (1996).

[11] Godfrey, J.J. et al.: Switchboard: Telephone Speech Corpus for Research and Development, *Proc. of ICASSP 92*, pp. 517-520, (1992).