

パネル討論「統計的言語処理／音声言語処理における 大規模言語データベースの利用」

鹿野 清宏（奈良先端大）、丸山 宏（日本IBM）、宇津呂 武仁（奈良先端大）、
松岡 達雄（NTT）、竹沢 寿幸（ATR 音声翻訳通信研究所）

あらまし 言語処理、音声処理の研究が進むとともに、多量の言語データの処理も可能となってきた。とくに、新聞記事などの多量のテキストが公開されるにつれて、大規模な言語データベースの作成、利用が大きな研究テーマとなってきている。また、インターネットなどの普及にともない、多種多様な情報の中から有効な情報を検索することが重要になってきている。さらに、マルチメディア情報の普及により、放送などの多量のメディアも利用できるようになりつつある。とくに、音声メディアは、音声認識技術の進展によっては、テキスト情報や言語情報にも変換が可能であり、新しい局面を迎えている。

このような状況をふまえて、このパネルでは、多量のテキストデータベースの構築や利用を行なっている研究者に、現在の状況や将来の方向を述べて頂きます。さらに、会場からの積極的な意見も頂きたいと思います。このパネルで、この分野の今後の方向が少しでも明確になればと思います。

1 大規模言語データベースの効果

鹿野 清宏 *shikano@is.aist-nara.ac.jp*

インターネットなどの急激な普及により、情報化社会が急速に進んでいる。気軽に検索できる情報も飛躍的に増大している。パソコンなどでのマルチメディア情報の蓄積や再生が容易になるとともに、放送、ビデオなどのデジタル化やデジタルライブラリの導入などにより、情報は、さらに急激に増大している。このような状況では、膨大な情報から必要な情報を検索する技術が必要となってくる。

現在でも、ネットワークを通して多くのテキストの検索が可能である。これらのテキストを利用して、言語モデルを構築して、言語処理に利用し、情報の自動インデッキング（キーワード抽出）に利用することが考えられる。さらに、音声メディアも音声認識を通すことにより、テキストに変換することが、ある程度まで可能であると思われる。現在、数万語を含む新聞記事などを対象として、音声言語研究会のWGでも、このようなディクテーションの研究の共通のデータベース（言語データベース、音声データベース）の構築が始められている。ネットワークからの有効な情報の検索のキーの自動抽出を考えても、次のような技術が必要となる。

1. テキストからの抽出
2. 音声からの抽出
3. ビデオからの抽出

そのためには、多量テキストデータ、あるいは、特定の分野のテキストデータからの言語モデルの学習や、学習するためのタギングが必要となる。また、このようなタギングされたテキストデータからの言語モデルを用いた音声認識も可能となり、音声データからのキーワードの抽出なども考えられる。このような音声認識技術は、次のような段階を経て、その適用領域が広がって行くであろう。

1. ディクテーション、キーワードスポッティング
2. 音声認識の結果からのキーワードの抽出による自動インデッキング
3. 動画像情報と音声情報を併用した自動インデッキング

これらの研究を進めるためにも、テキストデータベース、音声データベースなどの整備や、利用方法の研究が必要であり、共通の基盤で音声の研究者と言語の研究者が協力し合っていくことが重要である。

2 統計的言語処理における大規模テキストの利用

宇津呂 武仁 *utsuro@is.aist-nara.ac.jp*

2.1 統計的言語処理

大規模テキストを利用した統計的言語処理においては、様々な研究が行なわれているが、それらを大まかに分類すると以下ようになる。

- 形態素解析・Tagging のパラメータ学習

- 構文解析規則の学習・確率推定
- 共起知識の抽出
- 単語のクラスタリング

ここでは、特に、自然言語処理・音声言語処理の双方に関連の深い前者二つをとりあげる。ただし、対象としては日本語テキストを扱ったものに限定する。

まず、これらの研究で利用された、もしくは、利用されると思われる日本語コーパスおよび品詞体系として、ATR 対話コーパス・EDR コーパス・RWC テキストデータベース・日本語形態素解析システム JUMAN の標準文法の品詞体系をとりあげ、テキストの種類およびサイズ・品詞分類・利用事例を表 1 に示す。以下では特に、品詞体系の違いや語彙化の有無についての問題を重点的に説明する。

2.2 形態素解析のパラメータ学習

[永田 94, 永田 95] では、形態素解析済みコーパスから品詞 3 つ組モデル (tri-POS model) を推定することを行なっている。特に、[永田 94] では、ATR 対話コーパスについて、品詞だけでなく活用型・活用形の情報を付加したものを品詞タグとみなして 120 種類の品詞タグを用いて品詞 3 つ組モデルを推定している。一方、[永田 95] では、EDR コーパスについて、15 種類の品詞タグのみを用い、活用形などの情報を用いずに品詞 3 つ組モデルを推定している。両者を比較して、[永田 95] の方が数%適合率・再現率が低下したことが報告されており、原因の一つとして品詞タグの粒度の違いが挙げられている。

[森 96] では、形態素解析済みコーパスからの 2 つ組モデル推定における語彙化に焦点を当て、自立語・付属語ともに語彙化したマルコフモデル・自立語のみ語彙化したマルコフモデル・付属語のみ語彙化したマルコフモデル・品詞マルコフモデルの 4 つのマルコフモデルの重み付き重ね合わせによる 2 つ組モデルを提案している。EDR コーパスについて実験を行ない、[永田 95] よりも数%高い精度を達成している。

[竹内 95] では、数百文の形態素解析済みコーパスから形態素解析プログラム JUMAN の 2 つ組モデルの初期パラメータを推定し、さらに、約 20 万訓練文をこの JUMAN で解析した結果から HMM (Hidden Markov Model) 学習を行ない、JUMAN の 2 つ組モデルのパラメータを推定している。

品詞体系としては、JUMAN の 50 細分類に活用語の活用形を加え、さらに助詞・助動詞を語彙化したものを用い、新聞記事データを対象に実験を行なっている。

2.3 構文解析規則の学習・確率推定

[白井 95] では、EDR の構文解析済みコーパスから、15 品詞および助詞を語彙化した 161 個の非終端記号およびこれらを主辞とする 22 個の非終端記号で記述された確率文脈自由文法を抽出している。

[Hogenhout96] では、EDR の構文解析済みコーパスを訓練文として、[淵 94] の文法体系に基づいて記述された構文解析規則 (の非終端記号に意味主辞を付与したもの) の確率値を Inside-Outside アルゴリズムにより推定する実験を行なっている。

2.4 今後の研究課題

これまでの研究から、形態素解析済みコーパスからの形態素解析のパラメータ学習についてはかなり研究が進んできたが、形態素解析の観点からどのような品詞体系が最適かという問題はまだ未解決であるといえる。現在のところは、先見的知識により得られた品詞体系を用い、付属語をあらかじめ語彙化するなどして対処している¹。今後は、最適な品詞体系についての研究、およびその品詞体系を用いたテキストデータベースの構築、およびそこから言語モデルの構築が望まれる。また、構文解析規則の学習・確率推定に関しては、研究が始まったばかりである。英語を対象とした研究と比較しても、構文解析済みコーパスの整備の遅れのみでなくや出遅れているといえる。こちらについても、どのような文法体系にしたがって構文解析済みコーパスの構築・文法規則の学習を行なうかが重要な問題である。

参考文献

- [春野 96] 春野雅彦, 松本裕治: 文脈木を利用した形態素解析, 情報処理学会研究報告, Vol. 96, No. 27 (96-NL-112), pp. 31-36 (1996).
- [Hogenhout96] Hogenhout, W. R. and Matsumoto, Y.: Experiments with Using Semantical Categories in Parsing Systems, 言語処理学会第 2 回年次大会論文集, pp. 381-384, 言語処理学会 (1996).

¹ 文脈木を用いて可変長の接続規則および接続規則の語彙化を自動学習する手法 [春野 96] も提案されているので、今後の成果が期待される。

表 1: 日本語コーパスおよび品詞体系

	テキストの種類およびサイズ	品詞分類	利用事例
ATR	会議予約対話など 80 万語	26	[永田 94]
EDR	新聞など 75,000 文	15	[永田 95, 白井 95, 森 96, Hogenhout96]
RWC	通商白書など 11,179 文	大分類 12, 細分類 5 段階	—
JUMAN	—	大分類 14, 細分類 50	[竹内 95]

[森 96] 森信介, 長尾真: 形態素 bi-gram と品詞 bi-gram の重ね合わせによる 形態素解析, 情報処理学会研究報告, Vol. 96, No. 27 (96-NL-112), pp. 37-44 (1996).

[永田 94] 永田昌明: 前向き DP 後向き A* アルゴリズムを用いた確率的日本語形態素解析システム, 情報処理学会研究報告, Vol. 94, No. 47 (94-NL-101), pp. 73-80 (1994).

[永田 95] 永田昌明: EDR コーパスを用いた確率的日本語形態素解析, EDR 電子化辞書利用シンポジウム論文集, pp. 49-56 (1995).

[白井 95] 白井清昭, 徳永健伸, 田中穂積: EDR コーパスからの確率文脈自由文法の自動抽出に関する研究, EDR 電子化辞書利用シンポジウム論文集, pp. 57-63 (1995).

[竹内 95] 竹内孔一, 松本裕治: HMM による日本語形態素解析システムのパラメータ学習, 情報処理学会研究報告, Vol. 95, No. 69 (95-NL-108), pp. 13-19 (1995).

[洞 94] 洞武志: 日本語形態素構文解析のための新手法及び含意導出規則の定式化, 博士論文, 東京大学 (1994).

3 新聞記事データベースを用いた大語彙連続音声認識

松岡 達雄 matsuoka@splab.hil.ntt.jp

3.1 まえがき

近年, Wall Street Journal などの新聞記事を用いて, 英語, フランス語, ドイツ語, イタリア語などを対象に大語彙連続音声認識の研究が盛んに行なわれている [1]. しかし, 日本語を対象とした, これらに類する研究報告はない. これは, 主に, 日本語の文章が単語間にスペースなどのデリミタをおくことなく書かれており, 大語彙連続音声認識において重要な役割を果たす単語 N-gram のような言語モデルの導入が容易ではないことによると考えられる. また, 大語彙連続音声認識を対象とした音声データベースもなかった. そこで, 我々は, 新聞記事を用いて, 日本語大語彙連続音声認識のための音声データベースを設計/構築し, それを用いて大語彙連続音声認識の研究を進めている [2, 3, 4, 5].

3.2 データベースの設計

日本経済新聞 CD-ROM90-94 の記事 5 年分のうち, 4 年 9 カ月分を学習セットに, 残り 3 カ月分を

評価セットとした.

文章の読みやすさ, 形態素解析の精度を考慮し, テキストに前処理を施した. 我々の目的は大語彙連続音声認識であり, いわゆるディクテーションではないので, 通常の音声によるコミュニケーションで発声されない記号や括弧は取り除いた. 長すぎる文も読みにくいと削除した. 文の長さが平均単語数 $\pm 2\sigma$ 以内のものを単語頻度リストの作成と N-gram 言語モデルの学習に用いた. テキスト前処理後, 学習セットは 6.8M 文, 180M 単語 (形態素), 評価セットは 342k 文, 9.8M 単語 (形態素) となった.

文章を単語単位にセグメンテーションするため用いた形態素解析は 250k 形態素の辞書を持ち, 日経新聞記事に対する解析精度は 95% である. 本研究ではこの形態素解析に基づき形態素を単語と定義した. 学習セット中の単語を頻度順に並べた単語頻度リストを作成した. その結果, 623k 語からなる単語頻度リストが得られた. 373k 語は未知語として解析されたことになる. 未知語のほとんどは固有名詞や特殊な専門用語であった. 形態素解析の解析誤りを取り除くため, 上位 150k 語 (カバレッジ 99.6%) に含まれない語を含む文は削除した.

表 2 は各言語に対する新聞記事を用いた大語彙連続音声認識タスクの語彙サイズとカバレッジの比較である. WSJ タスクの 5k, 20k と同等のカバレッジを維持するため 7k, 30k の語彙サイズを決定した. 日本語はドイツ語と同様, 複合語が多く, さらに, 日本語では活用形により見掛けの異なり語彙数は多くなる. 日経タスクでの未知語の割合は英語とドイツ語の間であった.

大語彙連続音声認識システムを評価するため, 語彙サイズと一文中の未知語数を設定し, 学習セット, 評価セットごとに 5 種類のサブセットを定義した. 54 名の話者が, 各サブセットから 20 文ずつ, 計 100 文ずつ文を発声した. 50 文は学習セットから, 残り 50 文は評価セットから選んだ.

表 2: 各言語の大語彙連続音声認識タスクの語彙サイズとカバレッジ

	日経	WSJ	Le Monde	Frankfurter Randschau	Sole 24
学習テキストサイズ	180 M	37.2 M	37.7 M	36 M	25.7 M
異なり語彙数	623 k	165 k	280 k	650 k	200 k
5 k coverage (%)	88.0	90.6	85.2	82.9	88.3
7 k coverage	90.3	-	-	-	-
20 k coverage	96.2	97.5	94.7	90.0	96.3
30 k coverage	97.5	-	-	-	-
40 k coverage	98.2	99.2	97.6	-	98.8
65 k coverage	99.0	99.6	98.3	95.1	99.0
20 k OOV rate (%)	3.8	2.5	5.3	10.0	3.7

表 3: N-gram の種類 (上段) と平均出現頻度 (下段)

	Unigram	Bigram	Trigram
7k	7000	2.1 M	17.1 M
	24388.0	65.1	7.2
30k	30000	4.9 M	30.5 M
	6121.9	33.8	5.1

表 4: テストセットパープレキシティ

語彙サイズ	言語モデル	パープレキシティ		
		日経	WSJ	
			VP	NVP
7k/5k	Unigram	597	-	-
	Bigram	82	80	118
	Trigram	38	44	68
30k/20k	Unigram	693	-	-
	Bigram	124	158	236
	Trigram	64	101	155

3.3 言語モデル

表3に, 学習セット中に観測された unigram, bigram, trigram の種類数と平均出現頻度を示す. bigram, trigram は, ほとんどが一回だけしか観測されない singleton であった. 我々は, 言語モデルに対して, Katz による back-off スムージングを適用した. 表4に日経タスクと WSJ タスクのテストセットパープレキシティを示す. 日経タスクでは, 句読点だけは, 文の切れ目として残しているため, パープレキシティの比較では, WSJ の VP (Verbalized Punctuation) の場合と比較するのが妥当と考えられる. 引用符の有無や, 単語の定義自体の違いなどがあるため厳密な比較はできないが, オーダーとしては近い値を示していることは興味深い.

3.4 連続音声認識実験

7k 語彙について, 今回収録した音声データベース中の 10 名の話者の音声を用いて, 連続音声認

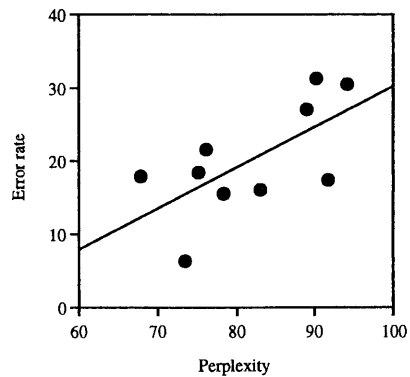


図 1: テストセットパープレキシティと単語誤り率

識実験を行った. 音素文脈独立の音響モデルと no-grammar 言語モデルを用いたベースラインの正解精度は, 17.2%であった. bigram 言語モデルの導入により正解精度は 63.7% まで改善された. さらに, 音素文脈依存音響モデル, そして, 対数パワーを用いることで, 80.0%まで改善された. 言語モデルの導入により誤り率が $1/2$ となり, 音響モデルを改善することで, 誤り率をさらに $1/2$ とすることができた. 各話者は異なる文章を読み上げているため, 文章自体の難しさ, つまり, パープレキシティが認識精度の違いに影響していると考えられる. 図1は, 各話者ごとに, テストセットパープレキシティを求め, 単語誤り率との関係をプロットしたものである. この結果より, パープレキシティが高いほど単語誤り率が高くなっており, 互いに相関があると考えられる. 実線は, 一次の回帰曲線である. この線からのばらつきは, 各話者の音響的な特徴の違いに起因するものと考えられることができるであろう.

謝辞

日本経済新聞 CD-ROM90-94 の研究利用を許諾して下さった日本経済新聞社に感謝します.

参考文献

- [1] L. Lamel and R. De Mori, "Speech recognition of European languages," *IEEE ASR Workshop*, pp. 51-54, Snowbird, 1995-12
- [2] 大附, 森, 松岡, 古井, 白井, "新聞記事を用いた大語彙連続音声認識の検討," 信学技報 SP95-90, 1995-12
- [3] 森, 大附, 松岡, 古井, 白井, "新聞読み上げタスクを用いた大語彙連続音声認識における言語モデルの検討," 音講論 pp. 159-160, 1996-03
- [4] 大附, 森, 松岡, 古井, 白井, "新聞読み上げタスクを用いた大語彙連続音声認識における音響モデルの検討," 音講論 pp. 161-162, 1996-03
- [5] T. Matsuka, K. Ohtsuki, T. Mori, S. Furui, and K. Shirai, "Large-vocabulary continuous-speech recognition using a Japanese business newspaper (Nikkei)," *Proc. ARPA Speech Recognition Workshop*, Harriman, 1996-02

4 ATR の旅行会話データベース

竹沢 寿幸 takezawa@itl.atr.co.jp

4.1 目的と概要

ATR 音声翻訳通信研究所では音声翻訳の研究を進めている。音声翻訳システム開発のためには、音声認識、言語翻訳、音声合成という3つの要素技術の研究が必要となる。ところが、音声翻訳システムを意識した研究と、各要素技術の研究では、必要なデータが異なる。音声翻訳、音声認識、言語翻訳の3つの観点から、必要なデータベースの要件を述べる。

(1) 音声翻訳研究のためのデータベースの要件

収集した音声データと言語データを統合して管理し、研究に活用できるようにすることが必要である。「近未来の音声翻訳システム」がどんな話題でも扱えるとは考えにくいので、音声翻訳システムが役立ちそうな場面や状況設定のもとで、ある種の目的を達成するような対話を収録したい。

(2) 音声認識研究のためのデータベースの要件

テキストを読み上げる朗読音声 (read speech) だけではなく、決められたテキストのない、自然で自発的な発話 (spontaneous speech) を扱おうとしている。限られた少数の場面設定のもとで収録した対話音声データが求められる。技術的には、韻律 (プロソディ) 情報の処理の重要性が増す。一方、大語彙の連続音声認識も相変わらず重要な研究テーマである。また、不特定話者の音声認識研究のためには、多数の話者が多様な発話様式で発話した音声データが必要である。

(3) 言語翻訳研究のためのデータベースの要件

話し言葉翻訳のためのコーパスが必要である。言語は文化を反映するので、異なる言語を話す人々の間では、対話の進め方が異なる。そのため、バイリンガル対話を集めることが必須である。しかも、広範な被覆率 (broad coverage) が望まれるので、多様な発話状況あるいは話題の設定のもとで、対話データを収集したい。

技術的には、いわゆる「不適格文」ないし断片文の構文解析技術が重要となる。さらに、状況に依存する発話の理解や適切な訳し分けのための文脈処理の研究も期待される。

これらの要件に合わせて、それぞれ、音声言語データベース、音声データベース、言語データベースを収集・整備している。それらの主な違いを表5に示す。

多くの人に親しみの持てる領域で、かつ、音声翻訳システムが役立ちそうな応用分野として「旅行に関する対話」を選び、対話を集めている。旅行者と旅行代理店あるいはホテル従業員との対話が典型的なものである。窓口業務のような各種案内を原則的に想定している。話題は複数のものを集めている。

以下では、音声認識から言語翻訳まで共通に研究利用する予定の音声言語データベースを中心に述べる。

4.2 データ収集方法

音声言語データベースの特徴の1つは、日本語話者と英語話者の、通訳者を介した対話を収集している点にある。通訳の質を高めるために、日英方向と英日方向の2名の通訳者を介している。「近未来の音声翻訳システム」のための基礎資料を目指しているため、通訳者は1回の発話 (発声単位) 毎に逐次的に通訳を行なう。通訳者が正確に伝えられるために、1回の発話は10秒以内としている。また、相手の話している間に割り込むことは禁止した。

どこかのホテルの窓口や旅行代理店で実際の対話を収録することは難しいので、模擬的な対話で音声言語データを収録している。ホテル担当者やホテル滞在者であるという設定資料を用意し、それをもとに模擬対話を行なっている。役割や設定をいろいろ変化した上で、多くの話者に演じてもらい、多様な音声言語現象を収録している。

すべての模擬対話は録音スタジオで同じマイクを使って DAT (デジタルオーディオテープ) に収録

表 5: ATR の旅行会話データベース

	音声データベース Speech-Specific Databases	音声言語データベース Common Database	言語データベース Language-Specific Databases
主要な利用目的	音声認識	音声翻訳 (音声認識と言語翻訳)	言語翻訳
音声波形データ	あり	あり	なし (DAT のみ)
書き起こしテキスト	あり	あり	あり
形態素情報の付与	あり	あり	あり
構文解析情報の付与	なし	あり	あり
話者・音声現象の多様性	大	中	—
話題・言語現象の多様性	小	中	大
典型的な対話の形態	日本語話者同士の対話	日本語話者と英語話者の通訳者を介する対話	日本語話者と英語話者の通訳者を介する対話

表 6: 音声言語データベースの品詞一覧

固有名詞	サ変名詞	形容名詞
普通名詞	代名詞	人名
住所名	日時	数詞
本動詞	形容詞	副詞
連体詞	接続詞	感動詞
助動詞	補助動詞	格助詞
係助詞	副助詞	連体助詞
並立助詞	準体助詞	接続助詞
終助詞	引用助詞	接頭辞
接尾辞	語尾	記号
間投詞	その他	

している。ただし、言語データベースはオフィスレベルの環境で収録している。

4.3 データベース作成ツールと利用方法

音声波形データと書き起こしテキストだけではなく、区切りおよび品詞情報からなる形態素情報や、構文木からなる構文解析情報も付与されている。日本語形態素情報は、音声認識用言語モデルの研究を含む、すべての日本語処理研究の基礎となるものである。日本語形態素情報の品詞一覧を表 6 に示す。語の単位認定の原則は次の通り。

- 固有名詞類は長い単位で 1 語と認定する。
- 普通名詞類は短い単位で 1 語と認定する。
- 基準となる辞書を選び、それを参照する。

また構文解析情報のデータベースでは、形態素にはない次のような品詞も付与している (表 7)。

1 つの構文木にまともならないものは、複数の部分木に分けている。確率文法の研究や、係り受け・格関係などの統計情報の抽出・評価実験に活用する予定である。

表 7: 構文解析情報データベースで追加した品詞

句点	読点	疑問符
感嘆符	中黒	文副詞
副詞的名詞	助動詞語幹	使役助動詞語幹
受身助動詞語幹	補助動詞語幹	名詞句

4.4 今後の展望

音声翻訳という研究目的のために利用可能な品質のデータを大量に収集するためには、通訳者を介した模擬会話が適していると考えられる。しかしながら、多様な対話の現象や音声・言語的現象を集めるという観点からは足りないものがある可能性がある。それらをすべて ATR で研究プロジェクトで網羅的に集めることは難しいので、あちこちの研究機関で収集された多様なコーパス・データベースが広く研究利用に公開されることが望まれる。ATR で現在収集しているデータベースは整備が済み次第、広く研究利用のために公開する予定である。

なお、データベースに関しては文献 [1, 2, 3] を参照して欲しい。また、データベースに現れる言語現象については文献 [4] に報告がある。

参考文献

- [1] 浦谷則好, 竹沢寿幸, 田代敏久, 森元逞, 勾坂芳典: “ATR の新音声言語データベース”, 情報処理学会第 48 回 (平成 6 年前期) 全国大会, 3Q-4, Vol. 3, pp. 79-80 (1994-03).
- [2] Morimoto, T., et al.: “A Speech and Language Database for Speech Translation Research”, *Proc. of ICSLP '94*, pp. 1791-1794 (1994-09).
- [3] 竹沢寿幸, 古瀬蔵, 中村篤: “音声言語データベース—話し言葉を収集し, 音声的・言語的特徴を探る—”, *ATR ジャーナル*, No. 17, pp. 4-5 (1994 秋).
- [4] 竹沢寿幸, 田代敏久, 森元逞: “音声言語データベースを用いた自然発話の言語現象の調査”, 人工知能学会 言語・音声理解と対話処理研究会 (第 10 回), SIG-SLUD-9403-3, pp. 13-20 (1995-02).