機械翻訳によって生成された追加テキストを使った言語モデルの適応*

中嶋秀治, 山本博史, 渡辺太郎
ATR 音声言語コミュニケーション研究所
〒619-0288 京都府相楽郡精華町光台 2-2-2

1 はじめに

N-gram などの統計的言語モデルの利用範囲（タスク）を新たなものに変更する場合には、その新たなタスクでの大規模コーパスを用いて再訓練を行なうことが望ましい。しかしながら、新たな大規模コーパスの入手はしばしば困難であり、その場合には新たなタスク（適応先タスク）の小規模コーパスを作成し、それを用いて既存の言語モデルが効果的に働くように調整する（適応化を行なう）。言語モデルが話し言葉の多言語音声翻訳器のためのものである場合には各言語での言語モデルが必要となるため、その利用範囲の拡大において、各言語での適応先タスクの小規模コーパスが必要となる。ところが、適応先タスクの単一言語での小規模コーパスの収集すらコストがかかり困難であり、多言語のコーパスの収集は更に困難である。そこで本論文では、ある1つの言語で書かれた適応先タスクのコーパスを、タスク適応が必要な言語モデルの言語に機械翻訳し、その翻訳結果を適応先タスクのコーパスとして利用して、言語モデルの適応化を行う方法を提案し、性能評価を行なう。

辞書、N-gram、用例などの機械翻訳器の翻訳用の知識には、単語の連接制約に関する情報が保持されていることが期待できる。翻訳結果が仮に訳文全体として不自然であったり間違いを含んでいたりしても、局所的な文脈ではN-gram にとって最も重要と思われる適切な単語の連接制約が反映されていることが多いと考えられる。また、適応先タスクのコーパスを翻訳するので、翻訳により生成されたコーパスもトピックや文のスタイルが適応先タスクのそれを反映していると考えられ、言語モデルの適応用のコーパスとして利用できることが期待できる。

本論文では上記の方法の提案と評価を行なう。まず2節で本研究の言語モデル適応の状況と方法を述べ、用いる機械翻訳の概要を3節で説明する。そして、4節で評価実験の条件と結果を述べる。5節で従来研究との関連を述べ、まとめを6節で行なう。

*Language Model Adaptation with Additional Texts Generated by Machine Translation

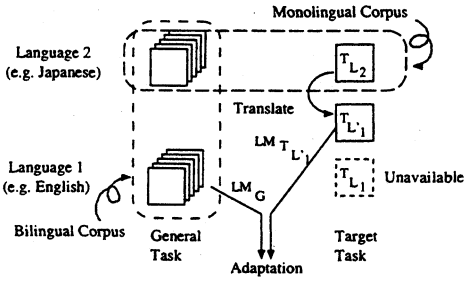

図 *1.* 翻訳結果を使った適応

2 言語モデルの適応方法

本研究での適応の状況は図1のようになる。すなわち、図1で *Language* 1 の言語モデルの適応において、適応先タスクの少量コーパス T_{L_1} が存在しない場合に、その代わりとなる言語データ $T_{L'_1}$ を、そのタスクでの他の言語の少量コーパス T_{L_2} から機械翻訳によって生成する。そして、これを大規模の一般コーパス（図1左側の *Language* 1 側の *General Task* コーパス）とともに使って言語モデルを適応させる。

適応後のモデルの推定手法として、MAP 適応[6]や、モデルを線形結合する方法[1]などの様々な手法があるが、本研究ではモデルの線形結合を利用した。

適応の手順をまとめると次のフローの通りとなる。

Step 1 翻訳器を用意する。

Step 2 一般タスクの言語モデル（図1の LM_G）を作成する。

Step 3 翻訳器を使い、一般タスクの言語モデルと同じ言語の適応先タスクのコーパス（図1の $T_{L'_1}$）を生成する。

Step 4 生成されたコーパスを使い、適応先タスク特定の言語モデル（図1の $LM_{T_{L'_1}}$）を作成する。

Step 5 Step 2の言語モデルとStep 4の言語モデルとを使い、適応先タスクに適応化した言語モデルを作成する。

3 機械翻訳システム

本提案手法には、様々な翻訳器の利用が可能であるが、本論文の評価実験では、統計的機械翻訳器を利用した。

本研究の評価実験で用いる翻訳器で用いられる統計的なモデルはBrownら[5]のIBM Model 4に基づいている。このモデルでは、翻訳単位として句への考慮が含まれていることからIBM Model 3よりも高い翻訳精度が期待でき、さらに、モデルのパラメータの総量が少ないためモデルの推定がIBM Model 5よりも速いことが期待できるので、実験に向いていると考え、IBM Model 4を選択した。モデルのパラメータの推定には、OchらのGIZA++[7, 8]を用い、そして、翻訳結果の探索には、Tillmannらの行なったビームサーチ[9]と同様の手法を用いた。翻訳器のモデルのパラメータは全て学習用のバイリンガルコーパス（例えば図1のGeneral TaskのBilingual Corpus）から推定される。

ここで、統計的翻訳について、本論文に必要な事項についての概要説明を行なう。

統計的翻訳では、翻訳の問題を、雑音のある通信路での復号問題と見て、モデリングを行う。例えば、日本語（J）から英語（E）への翻訳を考えると、日本語の文から英語の文への翻訳で最も尤もらしい翻訳結果（ここではこれをE^*とする）を得るという問題は

$$E^* = argmax\ P(E|J)$$

と表わされる。この式をベイズの公式で変形し、

$$E^* = argmax\ \{P(J|E)P(E)/P(J)\}$$

とし、ある1つの日本語の文を入力に定めた場合には分母が定数項となるので、分子のみで翻訳結果E^*の決定が行われる。ここで、分子の$P(J|E)$は「翻訳モデル」、$P(E)$は「言語モデル」と呼ばれる。言語モデルには単語N-gramがしばしば用いられる。

次節の実験では、この言語モデル$P(E)$を2節の図1のLM_Gとして、これを新たなタスク（図1のTarget Task）に合った言語モデルに適応させ、評価を行なう。

翻訳対象が学習用コーパスの外の新たなタスクの場合でも、モデルのパラメータが正しく推定されていれば適応前のモデルの制約（例えば$P(E)$）が働いて、出力される翻訳結果の中に局所的な文脈では適応先タスクで用いられる単語の比較的正しい並びが得られ、この翻訳結果を言語モデル適応用コーパスとして利用できることが期待できる。

4 評価実験

以下、本手法の有効性を、言語モデルの単語予測性能を示すテストセット文での単語パープレキシティー（PP）によって確認する。

4.1 一般タスクおよび適応先タスクのデータ

本実験では日英の対訳コーパスを利用する。文の内容は旅行時の会話表現である。このコーパスはおよそ16万文からなる。

このコーパスの各表現は、あらかじめ人手によって、「空港」、「飛行機内」、「レストラン」などの場面を主とした複数のカテゴリに分類されている。分類カテゴリの例を表1に示す。この

表 *1.* 文の分類カテゴリ

基本	空港	飛行機	連絡
両替	宿泊	帰国	レストラン
軽食	飲物	移動	トラブル
買物	観光	美容	コミュニケーション

中の「空港」での会話表現を評価実験の適応先タスクに設定し、残りを一般タスクに設定する。これらの内訳は表2の通りである。表2の「一般」は一般タスクを意味する。適応先タスクのデータの規模と適応の効果との関係を調べるために、サイズの異なる適応先タスクコーパスを3通り用意した。それらは、表2の「適応 1000」などである。これらとは別に、適応先タスクでの評価用コーパスとして同じ表2の「評価」を用意した。なお、文と単語の総数は英語で数えた値である。

表 *2.* 一般タスク、適応先タスク、及び評価コーパスのサイズ（英語で数えた値）

コーパス名	文数	単語数
一般	152,857	1,197,691
適応 1000	1,000	7,269
適応 2000	2,000	15,415
適応 4739	4,739	36,737
評価	4,739	36,191

4.2 実験の手順

本実験では、日本語から英語への翻訳（日英翻訳）を行なって得られた英語文を利用して英語の言語モデルの適応を行なう場合、及び、その逆の英日翻訳を行なって日本語の言語モデルの適応を行なう場合について、適応前後のPP値で評価する。本実験では、言語モデルには単語トライグラムを用いる。

また、上記の対照実験として、人手によって作成された適応先タスクの各言語の小規模コーパスを使った場合の言語モデルの適応によって達成されるPP値を調べる。この実験には、対訳コーパス内の訳文をそのまま適応に用いる。

表 3. データ量とパープレキシティー（日英方向での下限値（PP_1）と本手法での性能（PP_2））

データ量	PP_1	R [%]	PP_2	R [%]
一般	32.0	-	32.0	-
一般 + 1000	23.5	26.7	27.8	13.1
一般 + 2000	21.8	31.9	27.9	12.8
一般 + 4739	19.8	38.1	27.9	12.8

表 4. データ量とパープレキシティー（英日方向での下限値（PP_1）と本手法での性能（PP_2））

データ量	PP_1	R [%]	PP_2	R [%]
一般	23.0	-	23.0	-
一般 + 1000	17.7	22.8	20.0	12.8
一般 + 2000	16.6	27.6	20.1	12.6
一般 + 4739	15.4	32.9	20.2	11.9

統計的機械翻訳器の作成には、表2の「一般」のバイリンガルコーパスだけを用いる。一般タスクの言語モデル LM_G は片側の言語のコーパスのみを用いて作成する。そして、例えば日英翻訳の場合には、機械翻訳器へ表2の「適応 1000」などの形態素解析済みの日本語形態素列をそれぞれ入力する。出力された翻訳結果の第1位候補の英語文（単語への分割済）を適応用の擬似コーパスとして利用し、適応先タスク特定の英語の言語モデル LM_T を作る。そして、2つの言語モデル LM_G と LM_T とを線形結合することにより、適応化された言語モデルを作成する。日英翻訳を行なう場合には、適応用追加コーパス中の対応する英文は用いない。

言語モデルの適応の効果を調べるので、語彙サイズの変化による未知語への確率の配分の違いを排除するため、語彙サイズを適応前の辞書のサイズに固定した。

翻訳性能はしばしば音声認識の評価と同じ Word Error Rate (WER)、つまり、正解文の単語総数を T、正解文に対する翻訳結果の置換誤り数、挿入誤り数、削除誤り数を、それぞれ、Sub, Ins, Del とすると、

$$\mathrm{WER}\ [\%] = 100.0 \times (Sub + Ins + Del) \,/\, T$$

で定義される尺度で評価される。本設定での翻訳器の WER はおよそ 80% であった†。

4.3 結果と考察

日英翻訳を行ない英語の言語モデルの適応を行なった場合の結果を表3に示す。表3の1行目の「一般」の行は一般タスクの英語コーパスだけで作られた言語モデルでの PP 値を意味する。"+1000" などで追加される英文の数を示す。PP_1 の列の下3行は理想的な状況として対訳コーパスの訳（英文）を使って適応した場合の PP 値を、PP_2 の列の下3行は本手法で統計的機械翻訳によって生成された擬似的コーパスを追加した場合の PP 値である。各 R は各 PP 値の削減率である。

†対訳コーパスでの機械的照合の結果、翻訳への同じ入力文に対して異なる出力文が存在する場合は、それらをどちらも正解と設定した状況で計った。また、翻訳結果が大量であるため主観評価は行なわなかったが、概ね正しい結果が比較的多く見られた。

同様に、英日翻訳を行ない日本語の言語モデルの適応を行なった場合の結果を表4に示す。

表3や表4の PP_1 の変化のように、一般タスクのコーパスだけから作られた言語モデルでの PP に比べて、一般タスクの言語モデルとターゲットタスクの言語モデルとの適応化された言語モデルでの PP 値のほうが大幅に小さくなっている。また、これらの結果は、本実験の設定で、人手で作成されるような翻訳文を生成できる翻訳器が仮に存在した場合の PP 値の削減限界を示しており、適応前に比べて相対値で 30.0% 以上 PP 値を下げられることを示している。以上の点から、何らかの方法によって適応先タスクのデータを準備すべきであることが分かる。

本研究では統計的翻訳によって適応先タスクの擬似データを準備した。それを使った適応の結果が表3と表4の PP_2 から右の列である。

翻訳が WER の点では 80% であったにも関わらず、適応先タスクのコーパスとしての翻訳結果を使った適応によって、日英両言語において、適応前に比べて相対値で最大で 13% 前後の PP 削減が得られることが確認された。これは人手で作成されたコーパスを用いるという理想ケースの PP 値の削減率のおよそ 30 から 50% であり、効果は大きい。

しかし、機械翻訳によって生成された擬似コーパスを使った適応では、1000 文追加以後は性能が横ばい状態になっている。また、その適応の効果は、適応先のタスクでのコーパスが存在するという理想的なケースに比べると、およそ 50% 前後に留まっいる。

この原因は、翻訳性能が低いことのほかに、翻訳結果として第1位候補のみを用いたため、訳文内に確率の高い表現だけが現われ、多様な表現が十分に得られなかったためと考えられる。今後、翻訳結果の N-Best やラティス、さらには、複数の翻訳器の出力結果を使うなどの方法を取れば、それらの中から多様な表現が取り出されることにより改善されると考えられる。しかし、人手でコーパスを作成した場合と同じ効果を得るには、尤度に基づく N-Best だけではなく、同じ適切な意味を持つ複数の結果を出力したり、未知語に対して十分な対策を備えるなどのさらなる翻訳器の改善が必須であるが、これは今後の統計翻訳の課

題と考える。

5 議論と関連研究

本研究では、統計的翻訳器と、そのパラメータ学習に適応先タスク以外の大量コーパスを用いた。適応先タスクの表現との類似表現が学習コーパスに含まれている可能性があるが、その中のどの表現が適応先タスクの表現と類似するかは未知の状況で実験を行なった。本研究では適応に有効な表現を機械翻訳によって取出し利用することで、言語モデルの適応化を達成した。

言語モデルの作成と利用においては基本的には、適用先タスクでの大規模コーパスを作成することが最善の方法であるが、それはしばしば難しい。そこで、従来研究では、集められた小規模コーパスを用いて目的のタスクに適応化を行なう[1]か、World Wide Web (WWW) から得られた文書を用いて適応化が行なわれていた[2]。これらの研究では対象がディクテーションであり、話し言葉に比べて大量に存在する書き言葉を集めて利用する場合がほとんどであった。また、データの集めにくい医療所見のディクテーションや、マンマシンインタフェースの分野では、タスクでの会話を記述する文脈自由文法 (Context Free Grammar; CFG) を人手で作成し、それを使って人工的に生成したデータを使った適応が行われてきた[3, 4]。その一方で、話し言葉のコーパスはそもそも少なく、新たな利用先のタスクの小規模なコーパスすら存在しない状況がある。また、話し言葉のCFGでの記述も困難である。このような状況で、ある言語の言語モデルのタスク適応において、別の言語で書かれた適応先タスクのコーパスを機械翻訳し、その翻訳結果を利用するという本提案のような従来研究は見られないようである。

6 おわりに

本論文では、ある言語での言語モデルの新しいタスクへの適応をその言語での適応先タスクのデータが存在しない状況下でも可能とするために、その適応先タスクの別の言語で書かれたコーパスを機械翻訳し、生成された擬似的な適応先タスクのデータを使って、適応を達成する手法を提案した。旅行用会話文を対象としたテストセット単語パープレキシティーを評価尺度とする実験において、本手法によって、適応前にくらべて、およそ13%前後のパープレキシティーが削減されることが確認された。また、この削減量はコーパスが存在する理想的な状況のおよそ50%前後の削減量にあたり、これらの結果から本手法による言語モデル適応の有効性が確認された。また、このときの翻訳器のWERが80%という条件でも、言語モデルのタスク適応のためのデータ生成には十分に利用可能であることが確認された。

REFERENCES

[1] A. I. Rudnicky: "Language Modeling with Limited Domain Data," *Proc. of the ARPA Spoken Language Systems Technology Workshop*, pp. 66-69 (1995).

[2] A. Berger, et al.: "Just-In-Time Language Modeling," *Proc. of the ICASSP*, pp. 705-708 (1998).

[3] 伊東伸泰ら "文法を利用したN-gramモデルのタスク適応," 言語処理学会第4回年次大会発表論文集, pp. 610–613 (1998).

[4] Y. Wang, et al ed., "A Unified Context-Free Grammar and N-gram Model for Spoken Language Processing," *Proc. of ICASSP*, pp. 1639–1642 (2000).

[5] P. Brown, et. al, "The mathematics of statistical machine translation: Parameter estimation" *Computational Linguistics*, 19(2), pp. 263–311 (1993).

[6] H. Masataki, et. al: "Task Adaptation using MAP Estimation in N-gram Language Modeling," *Proc. of the ICASSP*, pp. 783-786 (1997).

[7] F. J. Och, and H. Ney, "A Comparison of Alignment Models for Statistical Machine Translation", *Proc. of COLING-2000*, pp. 1086–1090 (2000).

[8] F. J. Och, and H. Ney, "Improved Statistical Alignment Models", *Proc. of ACL-2000*, pp. 440–447 (2000).

[9] C. Tillmann and H. Ney, "Word Reordering and DP-based Search in Statistical Machine Translation", *Proc. of COLING-2000*, pp. 850–856 (2000).

[10] F. Jelinek and R. L. Mercer, "Interpolated Estimation of Markov Source Parameters from Sparse Data," *Proc. of the Workshop on Pattern Recognition in Practice*, pp. 381-397, North Holland, Amsterdam (1980).