統計的機械翻訳システムにおける LDA モデルの統合手法に関する研究

システム情報工学研究科 2 年 福富 崇博
指導教員 山本 幹雄
2008 年 6 月 26 日

1 はじめに

ある自然言語を別の言語に自動的に変換する技術は機械翻訳とよばれ、これまで盛んに研究されてきた分野のひとつである。現行の機械翻訳システムは主に人手で翻訳規則を記述する、ルールベースと呼ばれる手法で開発されている。しかしながら自然言語の規則には例外や曖昧性があるため、翻訳規則を全て列挙することは困難であり、また、規則の優先順位を人手で制御するのに膨大なコストを必要とする。一方、計算機の発達によって近年注目を浴びているのが、統計的な手法を用いた機械翻訳 (以下、統計的機械翻訳)[1] である。統計的機械翻訳とは原文と訳文が対になった対訳例データから、機械学習によって翻訳規則を自動抽出する手法である。この方式では翻訳規則を確率で表現するため、翻訳規則の一貫性、網羅性の面で優れている。

統計的機械翻訳は翻訳精度に貢献する複数の特徴関数から構成される。一般的に、これらの特徴関数はその性質から、翻訳らしさをモデル化する翻訳モデルと、言語らしさをモデル化する言語モデルに区別される。従来の統計的機械翻訳システムでは言語モデルとして、$n$gram モデルと呼ばれる手法を用いている。$n$gram モデルは近距離の単語の依存関係をモデル化するが、例えば文書全体の属するトピック等の、より広範囲の依存関係を考慮することはできない。統計的機械翻訳システムにこのような広範囲に渡る文脈情報を組み込むことで、翻訳精度の向上が期待できると考えられる。

大域的な文脈情報をモデル化する手法としては、LDA モデル [2] がよい性能を示すことが知られている。当研究室ではこれまで、LDA モデルを統計的機械翻訳システムに組み込む手法について研究を進めてきた [5][6]。本稿では新たに、LDA モデルを統計的機械翻訳の新たな特徴関数として追加する統合手法を提案する。

2 統計的機械翻訳

統計的機械翻訳では、原言語 E から目的言語 J への翻訳を Noisy channel model(雑音のある通信路モデル) でモデル化する。このモデルでは目的言語 J は雑音のある通信路を通過したことによって原言語 E に変換されてしまったものと仮定し、翻訳は原言語 E から目的言語 J への復号化 (decode) であるとみなす。復号化の誤りを最小化するような翻訳候補 $\hat{J}$ を求める式は以下のように表現される。

$$\hat{J} = \arg\max_J P(J|E) \tag{1}$$
$$= \arg\max_J P(J)P(E|J) \tag{2}$$

上述の式は、復号化誤り確率を最小化するものと解釈できる。我々は真の確率分布 $P(J|E)$ を知る事ができないため、実際には $P(J|E)$ を近似するモデル $p(J|E)$ を推定する。真の確率分布 $P(J|E)$ を近似する手法としては、log-linear model を用いる手法 [3] が提案されている。以下に、log-linear model を用いた統計的機械翻訳の理論的枠組みを示す。

$$Pr(J|E) = p_{\lambda_1^M}(J|E) \tag{3}$$
$$= \frac{\exp\left[\sum_{m=1}^{M} \lambda_m h_m(J,E)\right]}{\sum_{J_1'^I} \exp\left[\sum_{m=1}^{M} \lambda_m h_m(J_1'^I, E)\right]} \tag{4}$$

式中 $h_m(J,E)$ は、翻訳精度に貢献する性質を持つ M 個の特徴関数である。それぞれの特徴関数はモデルパラメータ λ_m を持つ。式 (4) の分母は任意の翻訳候補集合に対して一定の値となることから、これを式 (1) に代入すると以下のように書くことができる。

$$\hat{J} = \arg\max_J \sum_{m=1}^{M} \lambda_m h_m(J,E) \tag{5}$$

すなわち、それぞれの特徴関数をモデルパラメータで重み付けし、足し合わせた値が最大となるような翻訳候補の探索問題となる。式 (2) から、従来の統計的機械翻訳の特徴関数は $P(J)$ を近似する言語モデル $p(J)$ と、$P(E|J)$ を近似する翻訳モデル $p(E|J)$ とに大別される。

3 言語モデル

3.1 言語モデルとは

言語モデルは文の生起確率を与えるモデルであり、評価対象がより「言語らしい」ほど高い確率を付与することが求められる。例えば、次の 2 つの文を考える。

- 今日は全体ゼミの発表だ。
- 発表全体だ。は今日の

ここに M_1、M_2 の 2 つの言語モデルがあったとして、それぞれが上の 2 つの文を与えられたとする。このとき、言語モデル M_1 が

- P_1(今日は全体ゼミの発表だ。) = 0.12
- P_1(発表全体だ。は今日の) = 0.02

言語モデル M_2 が

- P_2(今日は全体ゼミの発表だ。) = 0.02
- P_2(発表全体だ。は今日の) = 0.12

のように確率を付与したとすると、我々の感覚としてより「言語らしいもの」に高い確率を付与しているのは M_1 である。つまりこのとき、M_1 は M_2 と比較して言語モデルとしての性能が高いといえる。我々は真の言語モデル $P(J)$ をみることはできないため、近似的にモデルを推定する必要がある。

3.2 $n$gram モデル

言語モデルを近似するモデルとしては $n$gram モデルが良い性能を示すことが知られており、従来の統計的機械翻訳システムにおいても、$n$gram モデルを用いるのが一般的である。$n$gram モデルはある単語列 $w_1^n = w_1 \cdots w_n$ の i 番目の単語 w_i の生起確率が直前の $N-1$ 単語のみに依存すると仮定したモデルであり、任意の

単語列 w_1^n の生起確率 $P(w_1^n)$ は以下で定義される。

$$P(w_1^n) = P(w_1)P(w_2|w_1)\cdots P(w_N|w_1^n) \tag{6}$$

$$\fallingdotseq \prod_{i=1}^{N} P(w_i|w_{i-N+1}^{i-1}) \tag{7}$$

理論的には N を大きくするほど長距離の依存関係を記述できるようになるが、実際に学習に用いるデータは有限なため、N を大きくするに従って、学習データ中に一度も出現しない単語連続が爆発的に増加する。このような単語連続群には明確な確率を割り振ることができない。ゆえに $n$gram モデルの性能の向上は、N がある程度大きくなると飽和する。統計的機械翻訳システムでは特に、$N = 5$ で良い性能を示すことが知られている。

4 LDA モデル

$n$gram モデルによって近距離の単語の依存関係をモデル化できることはこれまでに述べた。しかし、実際の単語出現確率は近距離の単語の依存関係だけではなく、大域的な文脈情報[*1]によって変動する。例えば「携帯電話」に関する文書中では他のトピックの文書と比較して、「Docomo」「Au」「Softbank」等の単語が頻繁に観察されるであろう。$n$gram のような近距離の単語の依存関係を扱うモデル (局所的言語モデル) に対し、このような大域的な文脈情報を扱うモデルを大域的言語モデルと呼ぶ。次節では大域的言語モデルのひとつである、LDA(Latent Dirichlet Allocation) モデル [2] について概説する。

LDA では、文書の生成過程について以下のような仮定をおく。

step 1. ディリクレ分布 $P_D(\boldsymbol{\theta}|\boldsymbol{\alpha})$ に従って K 個のトピックの比率 $\boldsymbol{\theta} = \theta_1 \cdots \theta_K$ を決定

step 2. 以下を N 回反復する

  step 2.1. $\boldsymbol{\theta}$ をパラメータとする多項分布 $p(z|\boldsymbol{\theta})$ からトピック z_n を選択

  step 2.2. $P(w_n|z_n, \boldsymbol{\beta})$ から w_n を選択

ここで、$\boldsymbol{\alpha}$ はディリクレ分布のモデルパラメータ、K は考慮するトピック数、$\boldsymbol{\theta} = \theta_1 \cdots \theta_K$ は各トピックにおける多項分布 $P(z|\boldsymbol{\theta})$ のパラメータ、N は文書に含まれる単語数、z_k は k 番目のトピック、$\boldsymbol{\beta} = \beta_1 \cdots \beta_K$ は各トピックにおけるユニグラムモデル[*2]、w_n は文書中 n 番目の単語、$P(w_n|z_k, \boldsymbol{\beta})$ はトピック z_k におけるユニグラムモデルを表している。LDA モデルは同一文書中の各単語が異なるトピックから生成されることを許すため、1 つの文書に複数のトピックを考慮することができる。上記過程を M 回繰り返すと、M 個の文書が生成される。このとき生成される文書群 $\boldsymbol{D} = \boldsymbol{w_1} \cdots \boldsymbol{w_M}$ の生起確率 $P(\boldsymbol{D}|\boldsymbol{\alpha}, \boldsymbol{\beta})$ は、以下のように表現される。

$$p(D|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \tag{8}$$
$$\prod_{d=1}^{M} \int p(\boldsymbol{\theta_d}|\boldsymbol{\alpha})(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{d_n}|\boldsymbol{\theta_d}) p(w_{d_n}|z_{d_n}, \boldsymbol{\beta}))d\theta_d$$

実際に文書確率を求めるにはモデルパラメータ $\boldsymbol{\alpha}$、$\boldsymbol{\beta}$ を学習する必要がある。$\boldsymbol{\alpha}, \boldsymbol{\beta}$ の推定はモデルパラメータ $\boldsymbol{\alpha}$、$\boldsymbol{\beta}$ が与えられたときの文書 $\boldsymbol{w}$ の出現確率 $p(\boldsymbol{w}|\boldsymbol{\alpha}, \boldsymbol{\beta})$ の対数尤度関数

$$\log p(\boldsymbol{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \tag{9}$$
$$\log \int p(\boldsymbol{\theta}|\boldsymbol{\alpha})(\prod_{n=1}^{N} \sum_{z_n} p(z_n|\boldsymbol{\theta}) p(w_n|z_n, \boldsymbol{\beta}))d\theta$$

の最大化問題として求めるが、この問題は解析的に解くことが困難であるため一般には変分ベイズ法、マルコフ連鎖モンテカルロ法等の近似的手法を用いる。

LDA モデルでは、これまで出現した単語の履歴 h を与えれられたときの事後分布 $p(\boldsymbol{\theta}|h)$ を用いて、次に出現する単語 $w*$ の期待値を以下のように表すことができる。

$$p(w*|h) = \int p(w*|\boldsymbol{\theta}) p(\boldsymbol{\theta}|h) d\boldsymbol{\theta} \tag{10}$$

このように単語の履歴に従って θ の分布を変化させることにより、トピックを考慮した単語出現確率の推定が可能となる。

5 LDA モデルを用いた統計的機械翻訳システム

5.1 先行研究

当研究室で行われた先行研究において、ユニグラムリスケーリング法、及びリスコアリング法によって LDA モデルを用いた統計的機械翻訳システムを実現する手法が提案されている [5][6]。これらの統合手法について、以下に概説する。

ユニグラムリスケーリング法ではトピックの影響による単語の出現確率の変動を $n$gram モデルにおける unigram 確率[*3]と LDA モデルにおける単語出現確率の比として表し、これを $n$gram モデルと掛け合わせることで統合する。具体的には、リスケーリング確率 $P(w_i|h, w_{i-n+1}^{i-1})$ は以下のように書ける。

$$P(w_i|h, w_{i-n+1}^{i-1}) \propto \frac{P(w_i|h) \times P(w_i|w_{i-n+1}^{i-1})}{P(w_i)} \tag{11}$$

ここで、$P(w_i|h)$ は LDA モデルにおける単語出現確率、$P(w_i|w_{i-n+1}^{i-1})$ は $n$gram モデル、$P(w_i)$ は $n$gram モデルにおける unigram 確率を示す。ただし、実際の計算では確率の正規化計算に膨大な計算を要するため、この統合手法は翻訳に非常に長い時間がかかるという欠点を抱えている。

リスコアリング法は従来手法の統計的翻訳システムにで良い翻訳候補と評価されたものを順番に N 候補を出力したのち、出力された N 候補に対し、LDA モデルを用いて再評価 (リスコアリング) する手法である。この手法は評価効率がよいという利点がある一方で、改善する可能性が従来手法の出力する N 個の翻訳候補に限られるという欠点を持つ。

5.2 提案手法

従来の統計的機械翻訳システムでは、翻訳候補の翻訳らしさを以下の式で計算する。

$$score(s) = \lambda_{LM} L(s) + \lambda_{TM} T(s) + \lambda_D D(s) \tag{12}$$
$$- \lambda_{WP}|s| - 100 \cdot unk(s)$$

ここで、s は翻訳候補文を表す。$L(s)$ は $n$gram モデルによる尤度、$T(s)$ は翻訳モデルによる尤度、$D(s)$ は歪モデルによる尤度、$|s|$ は s に含まれる単語数、$unk(s)$ は s に含まれる未知語の数で

[*1] 例えば文書の種類、分野、作成された時期、著者等

[*2] $N = 1$ としたときの $n$gram モデル

[*3] n=1 のときの $n$gram 確率

あり、翻訳精度に貢献する性質を持つ特徴関数のひとつである。それぞれの特徴関数が持つモデルパラメータは、Minimum Error Rate Training(以下、MERT)[3] によって決定する。本研究では、LDA モデルを統計的機械翻訳システムの新たな特徴関数として追加する手法を提案する。すなわち提案手法における翻訳候補の評価には、以下の式を用いる。

$$score(s) = \lambda_{LM} L(s) + \lambda_{TM} T(s) + \lambda_D D(s) - \lambda_{WP}|s| - 100 \cdot unk(s) + \lambda_{LDA} LDA(s) \quad (13)$$

この手法では LDA モデルを直接翻訳システムに組み込むため、リスコアリング法に比べて探索空間が広がり、より大幅な性能の向上が期待できる。また、ユニグラムリスケーリング法のような正規化計算を必要としないため、翻訳にかかる時間の増大も抑えることが可能となる。

6 LDA モデル性能評価実験

6.1 テストセット・パープレキシティ

言語モデルの性能評価には一般に、テストセット・パープレキシティと呼ばれる、情報理論に基づく客観的評価手法を用いる。テストセット・パープレキシティは次式で計算される。

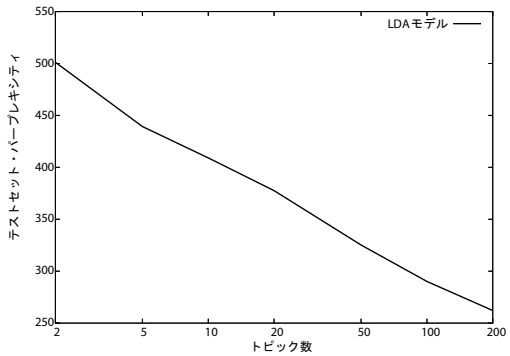
$$PP = P_M(w_1 \cdots w_N)^{-\frac{1}{N}} \quad (14)$$

ここで $w_1 \cdots w_N$ は評価用のテキスト集合、P_M は言語モデル M による $w_1 \cdots w_N$ の生成確率をあらわす。テストセット・パープレキシティは直感的な意味合いとして、「言語モデルを用いることで次の単語の予測候補をどの程度絞りこむことができたかを評価する指標」と解釈できる。テストセット・パープレキシティは言語のパープレキシティと比べて常に過大であることが理論的に保証されているため、テストセット・パープレキシティの値が低いほど言語モデルの性能が高い[*4]と評価できる。

6.2 実験の概要

特許文対訳コーパスを用いて LDA モデルを作成し、性能評価実験を行った。対訳となっている文書は同一トピックを持つ 1 つの文書としてモデル学習を行った。また、評価時は評価データとして日本語と英語の対訳文集合を用い、それぞれ英語文を単語の出現履歴として与えたときの、日本語文の言語らしさを評価した。言語モデルとしての性能が高いほど単語の候補を絞り込む能力に優れるため、翻訳システムへの貢献も高くなることが期待できる。実験条件の詳細を表 1 に示す。

6.3 実験結果



[*4] テストセット・パープレキシティの値が低くなるほど、言語のパープレキシティに接近するため。

表 1　実験条件

学習用データ	特許文対訳コーパス 文書数:32.522 総文数:1,818,885
学習用データ語彙数	英語:139,491 日本語:121,815
評価用データ	特許文対訳コーパス 総文数:899 文
評価用データ語彙数	英語:3,967 日本語:3,736
トピック数	1,2,5,10,20,50,100,200
変分ベイズ繰り返し演算回数	学習:20 回 適応:100 回

縦軸はテストセットパープレキシティ、横軸は考慮したトピックの数を示す。トピック数を増やすに従い、テストセットパープレキシティが順調に減少しているのがわかる。これは、考慮するトピック数が増えることによってより精密なモデル化ができるようになるためと考えられる。

7 翻訳システム性能評価実験

7.1 BLEU

統計的機械翻訳における性能評価では一般に、BLEU(BiLingual Evaluation Understudy) と呼ばれる客観的評価指標が用いられる。BLEU の計算式は以下で定義され、人手による評価と高い相関があることが知られている。

$$BP = \begin{cases} 1 & c > r \text{ のとき} \\ e^{1-\frac{r}{c}} & c \leq r \text{ のとき} \end{cases} \quad (15)$$

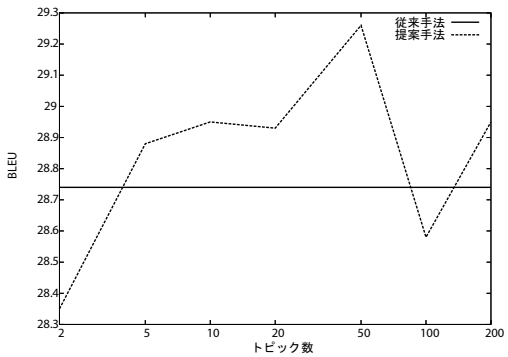
$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \quad (16)$$

ここで、c は翻訳結果の単語数、r は正解文の単語数、p_n は $n$gram[*5]の適合率を表す。w_n は各 N における重みであり、本実験では一様分布として, 各重みに $\frac{1}{N}$ を用いた。

7.2 実験の概要

従来の統計的機械翻訳システムの特徴関数として新たに LDA モデルを加え、日英方向の翻訳精度評価実験を行った。LDA モデルとしては、前節の実験と同じものを用いた。実験条件の詳細を表 2 に示す。各々特徴関数のモデルパラメータは、考慮するトピックの数毎に MERT によって決定した。

7.3 実験結果



[*5] N 連続の単語列。$N = 4$ がよく用いられており、本実験においてもこれにならって、$N = 4$ として性能評価を行った。

表 2　実験条件

言語モデル	SRILMToolkit 5gram モデル
ディスカウンティング法	Modified Kneser-Ney
線形補間	あり
単語対応	GIZA++ IBM Model 4
調整用データ	特許文対訳コーパス 総文数:915
評価用データ	特許文対訳コーパス 総文数:899 文
トピック数	1,5,10,20,50,100,200

横軸は考慮するトピックの数、縦軸は BLEU 値を示す。有意水準 5% のもとで二項検定を行ったところ、トピック数 5,50 のときに有意であった。全体として、考慮するトピック数が増えるに伴い、翻訳精度が向上する傾向にあるといえる。トピック数 100,200 において、BLEU は低い値を示した。この理由としては、MERT を行った際に局所解に陥った可能性が考えられる。

従来手法、提案手法それぞれの翻訳結果を観察すると、両システムの間には大きな差異がみられた。以下より、いくつかの翻訳例を紹介する。本提案手法はより正確な単語候補の選出に貢献するものであるため、ここでは特に、従来手法と比べて正しい単語が含まれているかについて検討を行う。

原言語文　on the other hand , when the spindle motor 3 is started at one of current values i2 , i3 , and i4 , the control skips step s4 .
正解文　なお、電流値 i2,i3,i4 のいずれかでスピンドルモータ 3 が起動 した場合には、ステップ s4 はスキップされる。
従来手法　一方、スピンドルモータ 3 の一方の電流値 i2,i3,i4, スキップ制御が 開始 されると (ステップ s4)、
提案手法　一方、スピンドルモータ 3 が 起動 されると、スキップ s4 で、制御電流値の i2,i3,i4 である。

原言語文の「the spindle motor 3」からトピックを捉えることにより、提案手法では「start」が「開始」よりもふさわしい「起動」に変化したと考えられる。これはまさしく、LDA モデルの機能が働いた好例といえる。

原言語文　the output terminal of the amplifier 41 is connected to the inverted input terminal of a comparator 46 .
正解文　前記 増幅器 41 の 出力 端子 は 、 コンパレータ 46 の 反転 入力 端子 に 接続 さ れ て いる 。
従来手法　増幅器 41 の 反転 入力 端子 に は 、 比較 部 46 の 出力 端子 が 接続 さ れ て いる 。
提案手法　増幅 回路 の 出力 端子 と コンパレータ 46 の 反転 入力 端子 に 接続 さ れ て いる 。

これも、LDA モデルが機能した例といえる。原言語文のトピックを捉えることによって、従来手法では「比較部」と訳されていた英単語「comparator」が、提案手法では「コンパレータ」と正しく訳された。

一方で、LDA モデルが機能したがために、従来手法より悪化する例もみられた。

原言語文　the reader unit 1 is further provided with an operation unit 115 for effecting various settings on the composite image input / output apparatus .
正解文　また 、 リーダ 部 1 に は 、 本 複合 画像 入出力 装置 に対して 各種 設定 を 行う ため の 操作 部 1 1 5 が 設け られ て いる 。
従来手法　また 、 リーダ 部 1 が 設け られ た 操作 部 1 1 5 の 複合 画像 入出力 装置 の 各種 設定 を 行う 。
提案手法　また 、 リーダ 部 1 の 複合 画像 入出力 装置 の 各種 設定 を 行う ため の 演算 部 1 1 5 が 設け られ て いる 。

従来手法では英単語「operation」が正しく「操作」に訳されていたが、提案手法ではトピックを捉えたことにより、「演算」に変化してしまったものと考えられる。

8　まとめと今後の課題

統計的機械翻訳システムの新たな特徴関数として、LDA モデルを導入した。LDA モデルの性能は考慮するトピック数を増やすに従って向上し、翻訳システムに統合すると、トピック数 5,50 において有意な性能の向上がみられた。実際の翻訳結果を観察すると、LDA モデルが機能することによって改善される例がみられたが、一方で悪化する例もあった。

今後の課題としては、先行研究との性能比較が挙げられる。統合手法以外の条件を同じにした上で、ユニグラムリスケーリング法、リスコアリング法、そして本稿で提案した統合手法のうちどれが良い性能を示すか、検証を急ぎたい。

参考文献

[1] Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: parameter estimation. Computational Linguistics, 19, 2, pages.264-311,1993.

[2] David M.Blei,Andrew Y.Ng,and Michael I.Jordan. Latent dirichlet allocation. Journal of Machine Learning Research,3,pages.933-1022,2003.

[3] Franz Josef Och. Minimum Error Rate Training in Statistical Machine Translation Proceedings of the 41th Annual Meeting of the Association for Computational Linguistics, July 2003, pp.160-167.

[4] 貞光九月. ”階層ベイズモデルを用いた混合ディリクレモデルのスムージング法”. 筑波大学大学院博士課程システム情報工学研究科修士論文,2006.

[5] 大塚悠平. ”言語横断大域的言語モデルを用いた統計的機械翻訳システム”. 筑波大学大学院博士課程システム情報工学研究科修士論文,2007.

[6] 西尾拓. ”トピック言語モデルを用いた統計的機械翻訳システム”. ” 筑波大学第三学群情報学類卒業研究論文,2008.