# 統計的機械翻訳における適応と応用

秋葉友良 (メディア科学リサーチセンター, セマンティックアーカイブコア)

## 1　はじめに

本稿では、本研究室での統計的機械翻訳に関する研究を紹介する。

## 2　統計的機械翻訳とは

自動機械翻訳の研究は古く、電子計算機の誕生からその可能性について論じられていたという。機械翻訳に対する古典的な手法は、専門家の翻訳知識をプログラムにコーディングする知能工学的アプローチであった。これに対し、1990 年代に IBM ワトソン研究所の Brown が提案した統計的機械翻訳手法は、2 つの言語で表された同じ意味を持つ文ペアの集合 (対訳データ) から翻訳知識を確率モデルとして自動抽出し、翻訳元文からその確率モデルに見合った (確率を最大とする) 翻訳文を探索することで翻訳を行うという、専門家の知識の代わりに統計処理だけを用いる全く新しい手法であった。その後、統計的機械翻訳は、モデル自体の洗練、学習・翻訳プログラムの整備、評価手法の確立など研究が進み、近年の機械可読言語データの整備や計算機処理能力の増大などの背景計算環境の向上と伴って、現在では機械翻訳研究の主流となるに至っている。

## 3　統計的機械翻訳におけるトピック適応

統計的機械翻訳では、より大量のトレーニングデータを使うことが性能向上につながることが知られている。しかしながら、大量のトレーニングデータ中にはいくつかのトピックが存在しており、それらを活用することで翻訳性能が改善できると考えられる。トピック適用のアイデアは多くの研究分野で用いられており、例えば音声認識の研究では、トピック適応させた言語モデルによる性能改善が示されている。

本研究では、特許文を対象にした統計的機械翻訳におけるトピック適応を検討した。特許には様々なトピックが含まれているため、特許文に対しては特にトピック適応の技術が有効であると考えられる。

統計的機械翻訳におけるトピック適応には以下の２つの課題がある。

- トレーニングデータ中のトピックの推定
- 入力文に最も近いトピックの特定

これら課題に対し、前者にはクラスタリング技術、後者には文書検索を利用した。

### 3.1　トピック依存翻訳モデルの学習

トピック依存翻訳モデルは、対訳データのうちの同じトピックを持つサブセット上で学習する。しかし、対訳データ中にどのようなトピックがどれだけ存在するのかは既知ではない。よって、教師無しのトピック推定方法として文書クラスタリングを利用し、トレーニングデータをトピックごとに分割する。トレーニングの流れを図 1 の左端に示す。

### 3.2　トピック依存翻訳モデルを用いた翻訳手法

入力文に対してトピック依存翻訳モデルを適用するためには、まず入力文のトピックを予測する必要がある。本研究では、トレーニング時に形成されたクラスタの中から入力文に最も似ているクラスタを見つけ、それを入力文のトピックとする。この処理には文書検索を利用し、ソース文をクエリ、PPD をターゲットの文書集合とみなして検索を行った。得られた N-best の文書検索結果を使ってクラスタごとに関連文書数を集計し、最も関連文書を多く含むクラスタを入力文のトピックだと予測する。その後、予測されたトピックに対応する翻訳モデルを使って入力文を翻訳する。翻訳の流れを図 1 の右端に示す。

### 3.3　評価実験

特許翻訳の評価型ワークショップ NTCIR-6 Patent Translation Task に参加し、評価実験を行った [2]。以下の手法の比較を行った。

- **Baseline**
  学習データすべてを使って単一の翻訳モデルをトレーニングする手法
- **Cluster-5**
  クラスタ数を 5 に設定し、5 個のトピック依存翻訳モデルをトレーニングする手法
- **Cluster-5-interpolate**
  **Baseline** と **Cluster-5** のフレーズテーブルを混合する方法

評価尺度には、統計的機械翻訳の評価によく用いられる BLEU(参照訳との n-gram 一致度 ($1 \leq n \leq 3$) の幾何平均) を用いた。評価結果を表 1 に示す。

トピック依存翻訳モデルによる性能改善は見られたが、その効果はそれほど大きくない。そこで、Baseline に対するトピック依存モデルの効果について分析を行った。詳細は文献 [6] を参照されたい。

## 4　新言語への適応

統計的機械翻訳を新たな言語対へ適用するためには、その言語対の対訳データが必要となる。しかし、特にマイナーな言語では、電子化された言語資源を用意することが難しく、十分な対訳データが用意できない場合が多い。そこで、言語資源が期待できるター

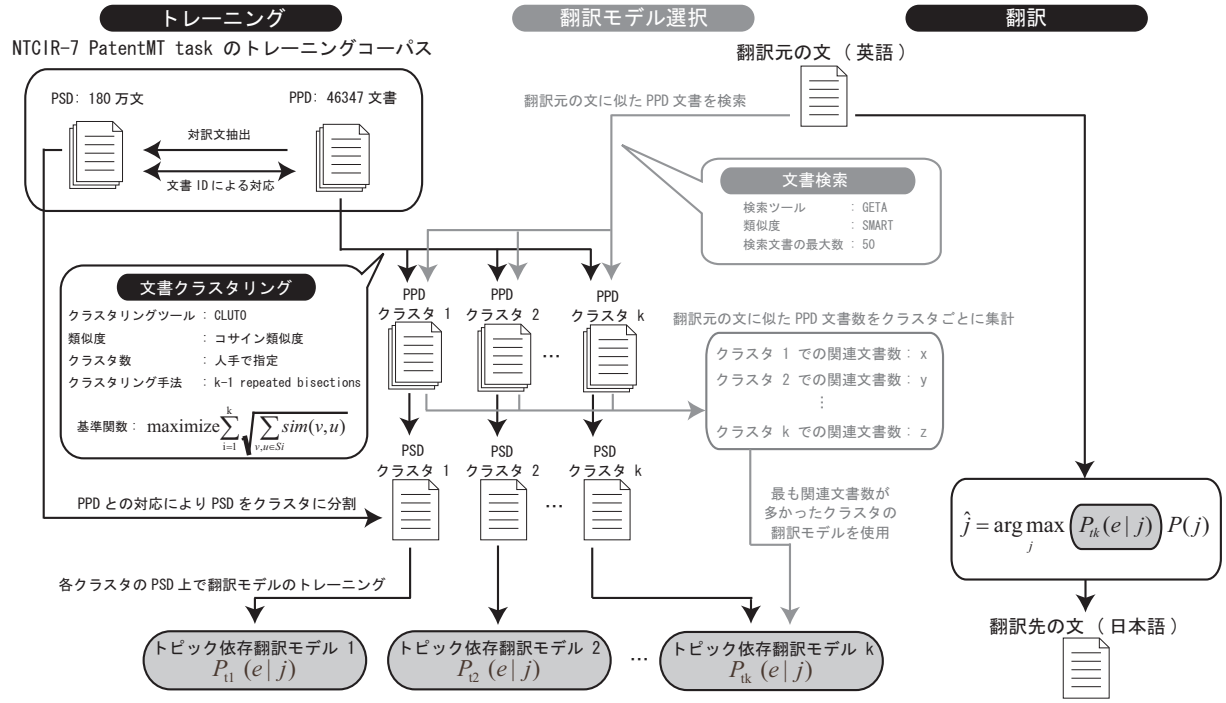


Fig. 1　トレーニングと翻訳の流れ

Table 1　NTCIR-7 formal-run テストデータでの評価実験結果

| 英日翻訳 | クラスタあたりの平均トレーニング文数 | BLEU |
|---|---|---|
| Baseline(1 cluster) | 1798571 | 29.80 |
| Cluster-5 | 339843 | 29.71 |
| C5-interpolate | (1798571) | 29.92 |

ゲット言語と三番目の言語 (中間言語) の対訳データを用いて、言語資源の少ないソース言語からターゲット言語への機械翻訳を実現する手法を研究している。ここで、ソース言語と中間言語の間では、辞書データのみを利用するものとする。

研究では、ベトナム語から日本語への翻訳を対象に、英日統計翻訳で用いる対訳データから越日対訳データを自動生成して、統計的機械翻訳を構成した。詳細は、文献 [3] を参照されたい。

## 5　音声ドキュメント検索への応用

音声ドキュメントを対象とした内容検索は、検索対象の音声文書を音声認識したテキストを対象とした文書検索問題として扱うことができる。しかし、音声認識誤りによるテキストの劣化による検索性能の低下が問題である。そこで、統計的機械翻訳を応用し、音声認識による書き起しテキストに現れる単語 $e$ が正解書き起しテキストにおいて単語 $f$ として現れる翻訳確率 $t(f|e)$ を用いて、正解テキストを予測して検索を行う手法を開発した [1, 5]。$t(f|e)$ の推定には、統計翻訳と同様に、音声認識結果の自動書き起しテキストと、人手による書き起しテキストのペアによる対訳データを用いる。また、翻訳モデルのような確率的手法と親和性の高いと考えられる言語モデルに基づく検索モデルへの適用も試みている [4]。

## 6　おわりに

統計的機械翻訳の適用範囲は広く、本稿で紹介した研究の他に、言語横断質問応答への適用も行っている。詳細は、平成 18 年度の研究報告を参照されたい。

## 発表論文


[1] T. Akiba and Y. Yokota. Spoken document retrieval by translating recognition candidates into correct transcriptions. In *Proceedings of International Conference on Speech Communication and Technology (Eurospeech)*, pp. 2166–2169, 2008.

[2] T. Ito, T. Akiba, and K. Itou. Effect of the topic dependent translation models for patent translation – experiment at ntcir-7. In *Proceedings of the 7th NTCIR Workshop*, pp. 425–429, 2008.

[3] Le Tuan Anh, 秋葉. パラレルテキストの自動的生成に基づく越日統計的機械翻訳. 言語処理学会第 14 回年次大会, pp. 997–1000, 2008.

[4] 秋葉, 本田耕一郎. 翻訳モデルを用いた講演音声ドキュメントの内容検索 – 文脈情報の利用と言語モデリング検索手法の適用. 第 3 回音声ドキュメント処理ワークショップ講演論文集, pp. 1–8, 2009.

[5] 秋葉, 横田. 認識候補から正解テキストへの翻訳に基づく講演音声ドキュメントのアドホック検索. 情報処理学会論文誌, 50(2):514–523, 2009.

[6] 伊藤, 秋葉. トピック依存翻訳モデルを利用した特許文の統計的機械翻訳. 言語処理学会第 15 回年次大会, pp. 244–247, 2009.