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

伊藤 雄　　秋葉友良
豊橋技術科学大学

1　はじめに

近年、機械翻訳の研究において統計的機械翻訳 (SMT) への期待が高まっており、その中でもフレーズベース統計的機械翻訳が性能の面から注目されている。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を検討する。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。

そこで本研究ではフレーズベース統計的機械翻訳の翻訳モデルにトピック適応を行った。以降、トピック適応させた翻訳モデルのことをトピック依存翻訳モデルと呼ぶ。本研究では、特許文の統計的機械翻訳においてトピック依存翻訳モデルのトレーニング手法、およびそれを利用した翻訳手法を提案し、その効果について分析を行う。

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にどのようなトピックが存在するかの推定
- ソース文に最も近いトピックの特定

これら課題への対策として、トレーニングデータのトピック推定にクラスタリング技術、ソース文に近いトピックの特定に文書検索をそれぞれ利用した。

2　使用するデータセット

本研究では翻訳の対象を特許文とし、NTCIR-7 特許翻訳タスク [1] でのデータを使用した。トレーニングデータは日英対訳特許文書集合 64347 文書 (PPD) と日英特許文のパラレルコーパス 1798571 文 (PSD) から構成される。PPD から自動的に文アライメントを取った結果から、信頼できると判断された対訳文の集合が PSD である。PSD が抽出された際に文書 ID も付加されており、これにより PPD, PSD は互いに参照可能となっている。

3　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、パラレルコーパスのうちの同じトピックを持つサブセット上でトレーニングされる。しかしながら、パラレルコーパス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。処理の詳細は以下のとおり。

1. クラスタリングツールキット CLUTO[8] を使って日本語側の PPD を決められた数のクラスタに分ける。クラスタの個数は人手で指定する。その際、標準型に直した自立語の出現頻度を文書の素性ベクトルとして使う。CLUTO の文書間類似度尺度にはコサイン類似度を使った。クラスタリングアルゴリズムには k-1 repeated bisections を使った。これは、指定されたクラスタ数に達するまで文書集合を 2 分割し続けるクラスタ分割手法である。
2. PSD が抽出された元の PPD 文書を調べ、その PPD 文書が属するクラスタを特定する。これを繰り返して、各クラスタに PSD を割り当てる。最終的にクラスタごとに PSD の集合が形成され、これがトピック依存コーパスとなる。
3. 各クラスタの PSD を使ってトピック依存翻訳モデルをトレーニングする。

   ただし、この手法では 1 つのモデルあたりのトレーニングデータ量が少なくなる。その解決策として、トピック依存のパラレルコーパスから作ったフレーズテーブルと PSD すべてから作ったフレーズテーブルを混合することによってデータ量を補完した手法も提案する。

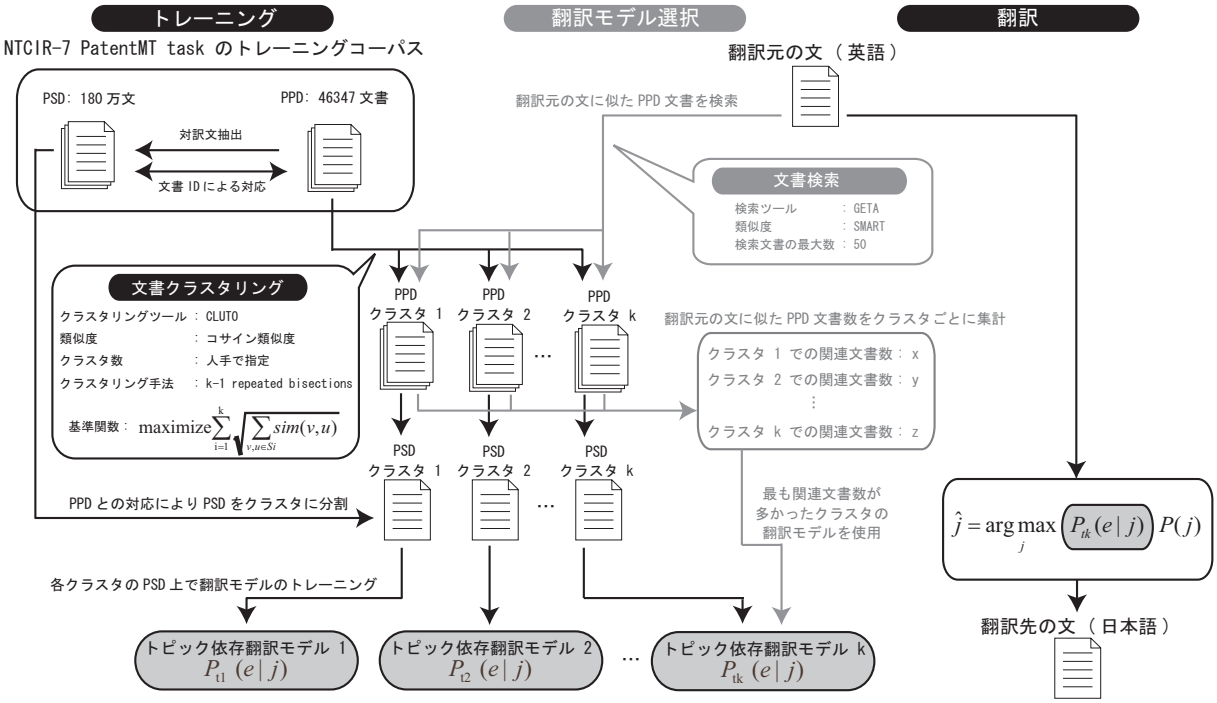

図 1: トレーニングと翻訳の流れ

4 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している。

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。

本研究では文書検索に GETA[9] を使用した。文書の索引付けの際には内容語のみを考慮し、単語重み付けにはピボット正規化をした TF-IDF を用いた。

5 評価実験

評価実験では、翻訳モデルトレーニングとデコーディングに Moses[4]、言語モデルトレーニングに SRILM[6] を使っている。言語モデルには、1 から 5gram までの補完モデルに Kneser-Ney smoothing を使用している。文書検索時の検索結果は 50-best までを使った。統計的機械翻訳ではより正確な翻訳結果を得るために、一般に depelopment データを用いて、言語モデルの確率、フレーズ翻訳確率、フレーズの並び替え確率などの重みパラメータをチューニングする。しかし本研究ではチューニングは行わず、Moses の初期設定を人手で修正して利用している。

トレーニングには NTCIR-7 特許翻訳タスクのトレーニングデータである PSD 1798571 文と、PPD 46347 文書を使用した。テストデータには NTCIR-7 特許翻訳タスクの formal-run テストデータ 1381 文を使用した。

なお、英語側の PPD には International Patent Classification(IPC) によるトピック情報が付加されている。これを利用してのトピック分類も可能だが、本研究では教師無しのトピック分類を目的としているため、IPC などの事前知識は使用していない。

実験では以下の手法を比較した。なお、言語モデルに関しては、すべての手法において PSD すべてを使った言語モデルを使っている。

- ベースライン
  - **Baseline**
    PSD 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
- 提案手法
  - **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

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

日英翻訳	クラスタあたりの平均トレーニング文数	BLEU
Baseline(1 cluster)	1798571	23.96
Cluster-10	179823	23.29
Cluster-5	339843	23.52

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-10	179823	29.29
Cluster-5	339843	29.71
C5-interpolate	(1798571)	29.84

- **Cluster-10**
  クラスタ数を 10 に設定し、10 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した方法

評価尺度には BLEU[7] を使った。評価結果を表 1 に示す。

結果から単純にトピック依存翻訳モデルを使った **Cluster-5** と **Cluster-10** がベースラインよりも低い性能であることが分かった。これは、**Baseline** に比べて、提案手法のクラスタあたりのトレーニングデータ量が少なくなっていることが原因であると考えられる。

トレーニングデータ量を補完した **C5-interpolate** ではベースラインを超える性能を示すことができた。これは **Baseline** と **Cluster-5** のフレーズテーブルを一対一で混合した場合の手法であり、混合の際の重みを変えた場合に性能がどう変化するかも調査した。結果を図 2 に示す。フレーズテーブル混合の比率を 1, 2, 3, 5 に変えて実験を行った。1:1 は **C5-interpolate** の場合、1:2, 1:3, 1:5 は **Cluster-5** のフレーズテーブルを 2, 3, 5 倍の量にして **Baseline** とフレーズテーブルを混合させた場合である。結果では 1:2 のときに性能のピークをむかえ、以降減少する傾向が示された。

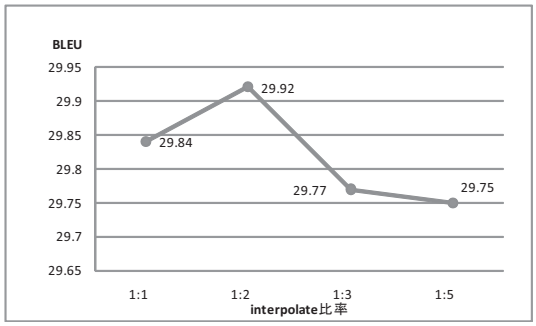

図 2: interpolate 比率による性能の変化

6 分析

トピック依存翻訳モデルによる性能改善はされたが、その貢献は少なかった。トピック依存フレーズ翻訳モデル $P_t(t|s)$ が有効に働くのは、トピック適応によりソースフレーズ s に対応するターゲットフレーズ t を絞り込むことができるときである。しかし、もし s に対して t がほぼ一意に決まってしまうならば、トピック適応の効果は薄いと考えられる。この仮定を調査するために、以下に示すフレーズ翻訳モデル $P(t|s)$ の perplexity を計算した。

$$Perplexity(s) = 2^{-\sum_t P(t|s) log_2 P(t|s)}$$

perplexity は、ソースフレーズ s が与えられたときに候補となるターゲットフレーズ t の平均の個数を表している。

Baseline のフレーズテーブルを使い、ソースフレーズ長（含まれる単語数）ごとに perplexity の平均を計算した。その結果を表 2 に示す。この調査結果から、ソースフレーズ長が長くなるほど平均 perplexity が減少し、長さ 4 以上のフレーズは平均して 2 つ以下のターゲットフレーズしか持たないことがわかる。すなわち、フレーズ翻訳モデルの場合には、長いソースフレーズに対してはターゲットフレーズがほぼ一意に決定されてしまう。このことがトピック依存翻訳モデルの貢献が期待よりも微少だったことの理由だと思われる。

7 関連研究

Utiyama and Isahara[2] は PPD に付加されている IPC 分類を使って PSD からトピック依存翻訳モデルをトレーニングしている。彼らはトピック依存翻訳モデルよりも、PSD すべてを使ったモデルの性能が良いと述べており、この結論は本研究と矛盾しない。ただし、本研究におけるフレーズテーブル混合のような手法は適用されていない。トピック依存翻訳モデルではモデルあたりのトレーニングデー

表 2: ソースフレーズ中の単語数ごとの平均 perplexity

日英翻訳		
フレーズ中の単語数	フレーズ数	perplexity
1	59382	7.29
2	996557	4.35
3	3795944	2.7
4	6292354	2.01
5	7155370	1.69
6	6666304	1.52
7	5567892	1.41

英日翻訳		
フレーズ中の単語数	フレーズ数	perplexity
1	122127	6.02
2	1594698	4.11
3	5146112	2.61
4	7302523	1.90
5	6845443	1.58
6	5096549	1.42
7	3352540	1.33

タデータの不足が問題になり、本研究ではフレーズテーブル混合によってデータ不足を解決できたことが最終的な性能の改善につながったと考えられる。

Yamamoto and Sumita[3] は旅行対話タスクのコーパスを使って、翻訳モデルと言語モデルの両方にドメイン適応させた手法を提案している。彼らの手法は本提案手法と似ているが、クラスタリング手法とモデル選択手法に大きな違いがある。またデコーダには Pharaoh[5] が使われている。なお、彼らはドメイン適応が効果的であるという結論に到っている。

8 おわりに

本論文では、統計的機械翻訳システムの翻訳モデルに対するトピック適応手法を提案し、その効果について調査を行った。今後の課題としては、クラスタリング手法の改善がある。また、クラスタリングの性能が明らかでないので調査をする必要がある。それに加えて、クラスタリング性能と翻訳性能の関係についても調査を予定している。

参考文献

[1] Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, Takehito Utsuro, "Overview of the Patent Translation Task at the NTCIR-7 Workshop", Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access, 2008.

[2] Masao Utiyama, Hitoshi Isahara, "A Japanese-English Patent Parallel Corpus", MT summit XI, pp. 475-482, 2007.

[3] Hirofumi Yamamoto, Eiichiro Sumita, "Bilingual Cluster Based Models for Statistical Machine Translation", PRoc. of Conference on Empirical Methods in Natural Language Processing(EMNLP), pp.514-523, June 2007.

[4] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin and E. Herbst, "Moses:Open source toolkit for statistical machine translation", ACL 2007 demonstration session, pp.177-180, 2007.

[5] P. Koehn, "PHARAOH: A beam search decoder for phrase-based statistical machine translation models", http://www.isi.edu/publications/licensed-sw/pharaoh/

[6] A. Stolcke, "SRILM - an extensible language modeling toolkit", ICSLP, pp.901-904, 2002.

[7] K. Papineni, S. Roukos, T. Ward, W. -J. Zhu, "Bleu: a method for automatic evaluation of machine trainslation", Proc.ACL, 2002.

[8] George Karypis, "CLUTO - A Clustering Toolkit", Technical Report 02-017, Dept. of Computer Science, University of Minnesota, 2002. Available at http://www.cs.umn.edu/cluto.

[9] "汎用連想検索エンジン GETA", http://geta.ex.nii.ac.jp/