統計的機械翻訳のための確率比を用いたリサンプリングによるドメイン適応

野口 敬輔 † 二宮 崇 ††
† 愛媛大学 工学部 情報工学科
†† 愛媛大学 大学院理工学研究科 電子情報工学専攻
{noguchi@ai., ninomiya@}cs.ehime-u.ac.jp

1 はじめに

近年、機械翻訳の分野において統計的機械翻訳 (SMT) が盛んに研究が行われている。SMT は、特定の分野に対する大量の対訳文書を用いて翻訳モデルの学習を行う機械翻訳の手法である。SMT は対訳文書から自動的に学習することが可能である一方、大量の対訳文書を開発することが大きな課題となっており、以前の多くの研究では数万文対程度の対訳文書のみで研究がなされてきた。ここ十年ほどの間、特定の分野では何百万文対もある対訳文書が自動的に生成されており、例えば、日英間での大規模対訳文書は約三百万文対の対訳文書が特許文書から生成されている [1, 2, 3]。このような自動的に生成された対訳文書は、SMT で研究を行うための十分な量を有している。しかし、特許文書のような、自動生成可能な対訳文書の分野は限られており、翻訳対象の分野と自動生成される対訳文書の分野が一致するとは限らない。SMT ではある分野の対訳文書から学習された翻訳モデルは他の分野の翻訳に用いる場合、大きく精度が低下することが知られている [4]。そのため、大量の対訳文書が生成されている分野の対訳文書を他の分野へ利用するためにはドメイン適応が必要である。ドメイン適応では、目標とはしない大量のデータが存する分野を「アウトドメイン (OUT domain)」と呼び、目標の分野を「インドメイン (IN domain)」と呼ぶ。

本稿では、特許文書翻訳から新聞文書翻訳へのドメイン適応を実現するために、共変量シフト下のインスタンス重み付けを言語モデルに基づくリサンプリングにより近似する手法を提案する。本研究の実験では、ランダムリサンプリングとインスタンス重み付けを用いたリサンプリングの二つの手法を比較する。ランダムリサンプリングはアウトドメインの対訳文書から無作為に選択された対訳文書をインドメインの対訳文書に付加してドメイン適応を行う。インスタンス重み付け [5] はアウトドメインのそれぞれの文に対して、インドメインでの出現確率をアウトドメインでの出現確率で割った値を重みとして与えサンプリングを行う。しかし、インスタンス重み付けを正確に行う場合は重みを反映させるために SMT の学習ツールの修正を行う必要があるが、学習ツールは非常に複雑でありブラックボックス的に扱える方が好ましい。そこで、アウトドメインの各文対に対してそれぞれのドメインでの出現確率の比を基にした重みにより、アウトドメインの各文をリサンプリングしてインドメインに付加することでインスタンス重み付けの近似を行う。また、本研究ではインスタンス重み付けに閾値を設け、閾値以上の重みを持つ全ての文書をインドメインへ付加する手法についても性能評価した。

2 ドメイン適応

ドメイン適応とは、転移学習とも呼ばれており、モデルまたはデータを解析対象ではないがデータが大量に存在する分野から解析対象の分野に対して適応させる手法である。ドメイン適応は、解析対象の分野の解析の精度を上昇させることを目的としている。ドメイン適応では、解析対象の分野のモデルまたはデータを「インドメイン」または「ターゲットドメイン」と呼び、解析対象の分野と異なる分野のモデルまたはデータは「アウトドメイン」または「ソースドメイン」と呼ばれている。本稿では、解析対象の分野をインドメイン、解析対象ではない分野をアウトドメインと呼ぶ。

2.1 関連研究

ドメイン適応はモデル適応 [6] とインスタンス重み付け [5] のふたつに大きく分類することが出来る。イ

ンスタンス重み付けは更に、メトリクスアプローチ、重み付け最適化アプローチ、共変量シフトの三種類に分類することが出来る。メトリクスアプローチはインドメインとアウトドメインのそれぞれの文書の距離を重みとして与える手法である。SMT ではいくつかメトリクスアプローチについての研究があり、それぞれの文対に対してクロスエントロピーを用いる手法 [7] やクロスエントロピー距離を用いる手法 [8] がある。メトリクスアプローチでは、閾値を設定し、その値に基づいてアウトドメインから文対を選択している。重み付け最適化アプローチでは、学習用データの文/句に対して重み付けを行い、その重みが目的関数に対し最良の結果になるようにインドメインの学習用データの重み付けを調整する手法である [9, 4]。共変量シフト [10, 5] はそれぞれの文に対してインドメインの出現確率とアウトドメインの出現確率の確率比を重みとして与える手法である。つまり、それぞれの文のインドメインでの出現頻度を重みとして考えている。最近の研究では重みを目的関数に直接組み込む手法と、重みについての閾値で文を選択する手法が提案されている。本研究の手法はインスタンス重み付けであるが、SMT に使用している点と、確率比に基づいたリサンプリングによって近似を行っている点が相違点である。

本研究に最も近い研究として Gascó らの研究 [11] が挙げられる。彼等の手法も SMT のためのリサンプリングによるドメイン適応である。本研究との相違点は、本研究ではインドメインでの出現確率とアウトドメインの出現確率を用いた共変量シフトによって重み付けをしているのに対し、彼等の研究ではインドメインの確率のみを利用していることである。

2.2 SMT のためのインスタンス重み付け

与えられた学習用データの対訳文書 $(s_i, t_i)_{i=1}^N$ に対してのパラメータ推定式を示す。ただし、s_i は原言語であり、t_i は目的言語である。

$$\begin{aligned}\hat{\theta} &= \arg\max_\theta \sum_{s\in\mathcal{S}}\sum_{t\in\mathcal{T}} p(s,t)\log p(t|s;\theta) \\ &\approx \arg\max_\theta \sum_{s\in\mathcal{S}}\sum_{t\in\mathcal{T}} \tilde{p}(s,t)\log p(t|s;\theta) \\ &= \arg\max_\theta \sum_{i=1}^{N} \log p(t_i|s_i;\theta)\end{aligned} \tag{1}$$

S は原言語、T は目的言語を表し、$\tilde{p}$ は経験分布を表す。インスタンス重み付けは式 2 のように導出することができる。

$$\begin{aligned}\hat{\theta} &= \arg\max_\theta \sum_{s\in\mathcal{S}}\sum_{t\in\mathcal{T}} p_{in}(s,t)\log p(t|s;\theta) \\ &= \arg\max_\theta \sum_{s\in\mathcal{S}}\sum_{t\in\mathcal{T}} \frac{p_{in}(s,t)}{p_{out}(s,t)} p_{out}(s,t)\log p(t|s;\theta) \\ &\approx \arg\max_\theta \sum_{i=1}^{N} \frac{p_{in}(s_i,t_i)}{p_{out}(s_i,t_i)}\log p(t_i|s_i;\theta)\end{aligned} \tag{2}$$

p_{in} はインドメインの出現確率であり、p_{out} はアウトドメインの出現確率である。インスタンス重み付けはそれぞれの文対に対してインドメインとアウトドメインの確率比を計算して重みとして与える。つまり、それぞれの文対にはインドメインでの出現頻度が重みとして与えられることになる。

3 SMT のためのランダムリサンプリング

ドメイン適応には二種類の単純かつ優れたベースライン手法がある。ひとつはインドメインの学習用データのみで翻訳モデルを学習する手法であり、もうひとつがインドメインとアウトドメインの学習用データを結合させて翻訳モデルを学習する手法である。しかし、後述する本実験ではインドメインの学習用データのみで学習を行った場合と比べ、インドメインとアウトドメインの学習用データを結合させて学習を行う場合の方が精度が低下した。原因として、インドメインよりもアウトドメインの方が対訳文書が大量にあるため、インドメインに対してアウトドメインの影響を与えすぎたことが考えられる。

この問題を解決する手法として、SMT でドメイン適応を行う場合のランダムリサンプリングによる手法 [12] について説明する。この手法は元々データの二値分類を行う場合のデータ不均衡問題を解決するための手法であるが、ドメイン適応に応用することができる。ランダムリサンプリングにはオーバーリサンプリングとアンダーリサンプリングの二種類があり [12]、どちらの手法でも元々のインドメインとアウトドメインの学習用データセットを利用する。オーバーサンプリングはインドメイン内から無作為に選択された文を更に加えることでインドメインの学習用データを増加させる手法である。アンダーサンプリングはアウトドメインから無作為に選択された文を削除することでアウトドメインの学習用データを減少させる手法である。本研究では、アンダーサンプリングをランダムリサンプリングの手法として実験を行った。

4 SMT のためのリサンプリングによるインスタンス重み付け

SMT のためのリサンプリングによるインスタンス重み付けの提案手法について説明する。式 2 を近似することで式 3 を得ることが出来る。

$$\begin{aligned}\hat{\theta} &= \arg\max_{\theta} \sum_{i=1}^{N} \frac{p_{in}(s_i,t_i)}{p_{out}(s_i,t_i)} \log p(t_i|s_i;\theta) \\ &\approx \arg\max_{\theta} \sum_{i=1}^{N} \frac{p_{in}(t_i)}{p_{out}(t_i)} \log p(t_i|s_i;\theta)\end{aligned} \tag{3}$$

ただし、$p_{in}(t)$ はインドメインの言語モデル、$p_{out}(t)$ はアウトドメインの言語モデルを表している。それぞれの言語モデルは n-gram の言語モデルで定義される。

本実験では Kneser-Ney スムージングによる 5-gram の言語モデルを n-gram の言語モデルとする。5-gram の Kneser-Ney スムージングによる言語モデルは SRILM ツールキットを用いて学習した。式 4 でそれぞれの文の出現確率を示す。

$$p(w_1 w_2 \ldots w_n) \approx \prod_{j=1}^{n} p(w_j|w_{j-4}w_{j-3}w_{j-2}w_{j-1}) \tag{4}$$

ただし、w は各文のそれぞれの単語であり、$p(w_j|w_{j-4}w_{j-3}w_{j-2}w_{j-1})$ は単語 w_j に対する 5-gram の確率を表す。ある文 t が与えられた時、$p_{in}(t)$ と $p_{out}(t)$ はそれぞれドメインで学習された 5-gram の言語モデルから計算される。

ある文 t が与えられたとき、式 3 の $p_{in}(t)/p_{out}(t)$ より重み $w(t)$ が求められる。重み $w(t)$ は文 t のインドメインでの出現頻度に相当する。提案手法では、アウトドメインの対訳文 (s,t) を $w(t)$ が 1 未満の場合は $w(t)$ を確率とみなしてサンプリングを行い、$w(t)$ が 1 以上の場合は一度だけ (s,t) をインドメインへ付加する。$w(t)$ が 1 以上の時にサンプリングしないのは、$w(t)$ の値が非常に大きな値をとる可能性があるからである。リサンプリング回数を $w'(t)$ として式 5 に示す。

$$w(t) = \frac{p_{in}(t)}{p_{out}(t)}, \qquad w'(t) = \begin{cases} w(t) & \text{if } w(t) < 1 \\ 1 & \text{otherwise} \end{cases} \tag{5}$$

また、先行研究では閾値を用いて文を選択する手法があるため、本研究でも式 5 の重み $w(t)$ に対して閾値を設定して文対 (s,t) の選択を行った場合についても実験を行う。

表 1　データセット

分野	学習用セット (文量)	開発用セット (文量)	テストセット (文量)
特許	3,166,284	-	-
新聞	130,000	500	10,000

5 実験

本実験では、英日間のランダムリサンプリングと提案手法の翻訳精度の比較を行った。

5.1 実験設定

本実験では、インドメインに新聞文書、アウトドメインに特許文書を使用した。英日の特許の対訳文書には共用タスクセット NTCIR10(PatentMT) から $3,166,284$ 文対を使用した。英日の新聞の対訳文書には JENAAD から $130,000$ 文対を使用した。表 1 は本実験で使用したデータセットの詳細である。言語モデルは Kneser-Ney スムージングによる 5-gram の言語モデルを用いた。また、インドメインの言語モデルの学習には新聞文書の学習用データセット、アウトドメインの言語モデルの学習には特許文書の学習用データセットをそれぞれ用いた。

アライメントには GIZA++ 1.0.7 を使用し、言語モデルの学習には SRILM 1.5.12 を使用した。日本語文のトークン化には Mecab 0.98 と ipadic 2.7.0 を用いた。SMT ツールには Moses と MERT を使用した。精度評価は BLEU スコアを用いた。翻訳方向は英日方向のみの翻訳だけを行った。

実験は、提案手法の他に、インドメインの対訳文書のみを用いて学習した場合、インドメインの対訳文書とアウトドメインの対訳文書を結合して学習した場合、アウトドメインをアンダーサンプリングした場合、インスタンス重みに閾値を用いた場合の五種類について行った。

5.2 実験結果

表 2 は、実験の各手法の結果を表している。「ベースライン (インドメイン)」は、インドメインの学習用データセットのみで学習した結果を表している。「ベースライン (インドメイン + アウトドメイン)」はインドメインとアウトドメインの学習用データを結合した学習用データセットで学習した結果を表している。「ランダム

リサンプリング」はアウトドメインの学習用データにアンダーサンプリングを行い、インドメインの学習用データと組み合わせて学習した場合の結果を表している。「インスタンス重み付け (リサンプリング)」は、提案手法のリサンプリングを用いたインスタンス重み付けの手法であり、「インスタンス重み付け (閾値設定)」は、インスタンス重み付けの重みに対して閾値を設定してアウトドメインの学習用データからデータを選択した場合の結果を表している。

インスタンス重み付けによるリサンプリングは、アウトドメインの学習用データセットから 441 文対しか選択されなかった。しかし、表 2 からわかるように、BLEU スコアは他の手法と比較して最大になり、インドメインのみで学習した場合と比べ、0.65 ポイントの上昇が確認できた。ランダムリサンプリングの結果は、インドメインの学習用データのみで学習した場合と比較すると、0.20 ポイント精度が低下した。インスタンス重み付けに閾値設定をした場合の結果は、インドメインの学習用データのみで学習した場合と比べ、0.41 ポイントの上昇が確認できた。

表 3 は、開発用セットに対するランダムリサンプリングとインスタンス重み付けの閾値毎の学習結果を表している。ランダムリサンプリングでは、10,000 文を追加した場合に開発用セットで最も高い精度を得ることができた。また、インスタンス重み付けの閾値設定では、10,000 文を追加した場合に最も高い結果を示した。

6 おわりに

本稿では、特許文書翻訳から新聞文書翻訳へのドメイン適応を実現するために、共変量シフト下のインスタンス重み付けを言語モデルに基づくリサンプリングにより近似する手法を提案した。インスタンス重みを用いたリサンプリングは、アウトドメインの各文に対してインドメインとアウトドメインの確率比を重みとしたサンプリングを行ってインドメインに適応させる。インスタンス重みによって選択される文対の集合は、アウトドメインのデータの中でもインドメインらしいデータが選択されている。また、インスタンス重みに閾値を設定し、閾値よりも重みが大きい文を選択してインドメインに適応させる手法についても実験を行った。本研究では、n-gram の言語モデルを使ってインドメインとアウトドメインでの各文対の生成確率の計算を行った。実験では、提案手法であるリサンプリング

表 2　結果

	BLEU (%)
ベースライン (インドメイン)	13.93
ベースライン (インドメイン + アウトドメイン)	12.67
ランダムリサンプリング	13.73
インスタンス重み付け (リサンプリング)(提案手法)	14.58
インスタンス重み付け (閾値設定)	14.34

表 3　閾値毎のリサンプリングの結果

	追加文量	ランダムリサンプリング (BLEU (%))	インスタンス重み付け (閾値設定) (BLEU (%))
ベースライン (インドメイン)	0	13.44	-
リサンプリング	10,000	13.97	13.69
	20,000	13.36	13.48
	40,000	13.10	13.48
	80,000	13.78	13.54
	500,000	12.79	12.30
	1,000,000	12.35	12.10
	2,000,000	12.35	11.59
ベースライン (インドメイン + アウトドメイン)	3,166,284	12.27	-

を用いたインスタンス重み付けが最も高い精度を示した。提案手法は、閾値を設定する必要がなく、一度のサンプリングでドメイン適応のためのデータセットを決定することができるため、閾値を設定するためのコストが必要ない点も利点に挙げられる。

謝辞

本研究は JSPS 科研費 25280084 の助成を受けたものである。ここに謝意を表する。

参考文献

[1] Utiyama, M. and Isahara, H.: Reliable Measures for Aligning Japanese-English News Articles and Sentences, *Proc. of ACL '03*, pp. 72–79 (2003).
[2] Koehn, P.: Europarl: A Parallel Corpus for Statistical Machine Translation, *Proc. of MT Summit 2005*, pp. 79–86 (2005).
[3] Callison-Burch, C., Koehn, P., Monz, C. and Schroeder, J.: Findings of the 2009 Workshop on Statistical Machine Translation, *Proc. of WMT 2009*, pp. 1–28 (2009).
[4] Foster, G., Goutte, C. and Kuhn, R.: Discriminative Instance Weighting for Domain Adaptation in Statistical Machine Translation, *Proc. of EMNLP 2010*, pp. 451–459 (2010).
[5] Jiang, J. and Zhai, C.: Instance Weighting for Domain Adaptation in NLP, *Proc. of ACL 2007*, pp. 264–271 (2007).
[6] Foster, G. and Kuhn, R.: Mixture-Model Adaptation for SMT, *Proc. of WMT 2007*, pp. 128–135 (2007).
[7] Yasuda, K., Zhang, R., Yamamoto, H. and Sumita, E.: Method of Selecting Training Data to Build a Compact and Efficient Translation Model, *Proc. of IJCNLP 2008*, pp. 655–660 (2008).
[8] Axelrod, A., Li, Q. and Lewis, W.: Applications of Data Selection via Cross-Entropy Difference for Real-World Statistical Machine Translation, *Proc. of IWSLT 2012*, pp. 201–208 (2012).
[9] Matsoukas, S., Rosti, A.-V. I. and Zhang, B.: Discriminative Corpus Weight Estimation for Machine Translation, *Proc. of EMNLP 2009*, pp. 708–717 (2009).
[10] Shimodaira, H.: Improving predictive inference under covariate shift by weighting the log-likelihood function, *Journal of Statistical Planning and Inference*, Vol. 90, No. 2, pp. 227–244 (2000).
[11] Gascó, G., Rocha, M.-A., Sanchis-Trilles, G., Andrés-Ferrer, J. and Casacuberta, F.: Does more data always yield better translations?, *Proc. of EACL 2012*, pp. 152–161 (2012).
[12] He, H. and Garcia, E. A.: Learning from Imbalanced Data, *IEEE Trans. on Knowl. and Data Eng.*, Vol. 21, No. 9, pp. 1263–1284 (2009).