バイリンガルコーパスクラスタリングを用いた統計翻訳モデルのオンラインタスク適応*

山本 博史 †, 隅田 英一郎 †‡
† 情報通信研究機構 音声言語グループ
‡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

近年、N-gram に代表される統計ベースのモデルが音声認識や統計翻訳といった自然言語処理の分野でも、広く用いられるようになってきている。この統計ベースのモデルを用いるにあたっては処理対象のドメインに依存したモデルを用いた方が性能が向上することが知られている。しかしながら、ドメイン依存モデルの利用には、次の二つの問題がある。一つ目はデータスパースネスの問題であり、これを解決する手法がタスク適応 [1] である。二つ目はドメイン推定である。タスク適応のためには、ドメインが既知でなければならないが、ドメインが未知、あるいは動的に変化する場合の方が一般的である。このような場合にタスク推定を行うためには、適応対象のドメインを推定する必要がある。この場合、ドメイン推定とタスク適応をオンラインで行うことになる。本稿では、オンラインでのドメイン推定とタスク適応を統計翻訳に適用することを試みる。

2 統計翻訳におけるドメイン依存モデル

統計翻訳の目的は翻訳元言語の文 f が与えられた時に、確率が最大となる翻訳先言語文 e を見つけることであり、次の式であらわされる。

$$\underset{e}{argmax}\ P(e|f) \tag{1}$$

この式では、翻訳先言語文 e は翻訳元言語文 f にのみ依存するが、実際はそれだけではなく、ドメイン D に強く依存する。そこで、このドメイン D を、新たな確率変数として式 (1) に導入するごとによって、次式が得られる。

$$\underset{e}{argmax}\ P(e|f,D) \tag{2}$$

さらに、この式はベイズ則により、次のように書き換えることができる。

$$\underset{e}{argmax}\ P(e|D)P(f|e,D) \tag{3}$$

ここで、$P(f|e,D)$ はドメイン D 依存の翻訳モデル、$P(e|D)$ はドメイン D 依存の言語モデルをあらわしている。

*On-line Adaptation for Statistical Machine Translation Using Bilingual Corpus Clustering

ドメイン D が既知の場合は、あらかじめこれらのドメイン依存モデルを用意しておき、そのモデルを用いて翻訳を行うことができるが、ドメインが未知、あるいは動的に変化する場合の方が一般的である。この場合は、ドメイン D そのものを翻訳先言語文 e と同時に推定する必要があり、次式であらわされることになる。

$$\begin{aligned} & \underset{e,D}{argmax}\ P(e,D|f) \\ = & \underset{e,D}{argmax}\ P(D|f)P(e|f,D) \\ = & \underset{e,D}{argmax}\ P(D|f)P(e|D)P(f|e,D) \end{aligned} \tag{4}$$

この式において $P(D|f)$ はドメイン推定をあらわしており、$P(e|f,D)$ はドメイン依存の翻訳をあらわしている。

3 提案法の概要

ここでは、提案する統計翻訳に対する、オンラインでのドメイン推定とタスク適応の概要に関して記述する。提案法は二つの手順から構成されており、事前にバッチ処理的に行っておくオフラインプロセスと、実際の翻訳文が与えられた時に動的に行うオンラインプロセスに分けられる。

3.1 オフラインプロセス

このプロセスでは、まず訓練コーパスのクラスタリングを行う。クラスタリングの結果として、サブコーパスが得られるが、そのサブコーパスをドメインと見なし、ドメイン依存モデルを構築する。統計翻訳のための訓練コーパスはバイリンガルコーパスであるが、これらは、以下に示す手順でクラスタリングされ、クラスタすなわちドメイン依存モデルが作成される。

1. バイリンガルコーパスの対訳文対をクラスタリングする (詳細は 4.2)。
2. 各クラスタごとに、クラスタ (ドメイン) 依存翻訳モデルを作る。
3. 各クラスタごとに、クラスタ (ドメイン) 依存言語モデルを作る。

3.2 オンラインプロセス

オンラインプロセスでは、翻訳元文が入力されるつど、ドメイン推定とドメイン依存の翻訳が繰り返される。

1. 翻訳元文が所属するクラスタを求める。
2. 求まったクラスタ (ドメイン) 依存のモデルを用いて、翻訳を行う。

4 ドメイン推定

式 (4) を満足させるためには、$P(D|f)$ と $P(e|f,D)$ を最大化する D と e を同時に求める必要がある。しかしながら、これは困難であるため、近似を導入することにする。この近似では、まず $P(D|f)$ と最大化する D を求め、求まった D のもとで $P(e|f,D)$ を最大化する e を求めることとする。この近似を導入により、式 (4) は次のように書き直すことができる。

$$\underset{e}{argmax}\ P(e|f, \underset{D}{argmax}\ P(D|f)) \tag{5}$$

4.1 ドメインの定義

ドメインがあらかじめ与えられる場合は、トピックなどがドメインとして用いられる。従って、このような場合はドメインの定義は人間の感覚にあったものであることが望ましい場合が多い。しかしながら、提案法が対象とするドメインの推定を動的に行うような場合は、必ずしも人間の感覚にあったものである必要はない。統計的な立場から考えた場合、良いモデルとは、正しい対訳ペア $\{e,f\}$ を与えた時に、なるべく大きな $p(D|f)p(e|f,D)$ を与えるものである。これと同様に、$p(D|f)p(e|f,D)$ を最大化するようなドメイン D の定義が好ましいことになる。提案法では $p(D|f)p(e|f,D)$ の近似として、$p(D|f)p(e|D)$ を最大化するようにドメイン D を定義することとする。この最大化の対象である $p(D|f)p(e|D)$ はベイズ則を用いて次のように書き換えることができる。

$$\begin{aligned} & P(D|f)P(e|D) \\ = \ & P(e|D)P(f|D)P(D)/P(f) \end{aligned} \tag{6}$$

ここで、$P(f)$ は D に依存しないため考慮する必要がなく、さらに、$P(D)$ を定数と見なす近似を導入することにより、望むべきドメインの定義は、次式であらわされることになる。

$$\underset{D}{argmax}\ P(e|D)P(f|D) \tag{7}$$

そして、この式は対訳文ペア $\{e,f\}$ に対し、エントロピー基準でクラスタリングを行った結果のクラスタをドメインとして定義することを意味している。

4.2 バイリンガルコーパスのクラスタリング

上で述べたように、ドメイン定義として、バイリンガルコーパスをエントロピー基準クラスタリングしたものを用いる。クラスタリングの手順は以下の通りであり、これはモノリンガルコーパスのクラスタリング [2] を拡張したものになっている。

- あらかじめ、ドメイン数すなわちクラスタ数を定めておく。
- バイリンガルコーパス中の全ての対訳ペアをランダムにクラスタに割り当てる。
- それぞれのクラスタごとに e と f のクラスタ依存言語モデルを作成する。
- 各クラスタごとに e と f のエントロピーをクラスタ依存言語モデルを用いて計算し、全てのクラスタのエントロピーの和を総エントロピーとする。
- 個々の対訳ペアを総エントロピーが最も小さくなるクラスタに移動させる。
- 以上の操作を総エントロピーの減少が閾値より小さくなるまで繰り返す。

4.3 ドメインの推定

上に示したように、提案法ではドメインは訓練データのクラスタとして定義される。従って、ドメイン推定はすなわちクラスタ選択ということになり、与えられた翻訳元文 f に対し、$P(D|f)$ が最大となるような D を選択することになる。最大化の対象の $P(D|f)$ は、ベイズ則で $P(f|D)P(D)/P(f)$ と書き換えられる。さらに $P(f)$ は定数であり、ドメイン定義の際に用いた $P(D)$ が定数という近似を導入すれば、クラスタ選択の際の最大化の対象は $P(f|D)$ となる。これは、すなわち f に対し、パープレキシティを最大にするクラスタ D をドメインとして推定すれば良いことをあらわしている。

5 ドメイン依存の翻訳

ドメイン推定の後、推定されたドメイン依存の翻訳を行う。ドメイン依存の翻訳は $P(e|f,D)$ を最大化する e を求めることであるが、この最大化の対象はベイズ則で次のように書き換えることができる。

$$\begin{aligned} & P(e|f,D) \\ = \ & P(f|e,D)P(e,D)/P(f,D) \\ = \ & P(f|e,D)P(e|D)P(D)/P(f,D) \end{aligned} \tag{8}$$

ここで、D、f は既知であるため、最大化の対象は $P(f|e,D)P(e|D)$ となる。この内、$P(f|e,D)$ はドメイン依存翻訳モデル、$P(e|D)$ はドメイン依存言語モデルである。

5.1 先行研究との相違点

5.1.1 クラスタ言語モデル

ドメインの定義、推定の方法として、すでにクラスタ言語モデル [3] が提案されている。この手法は三つの手順からなっている。まず、ドメイン定義として、人手で定義された Reguler Expression にしたがって翻訳先コーパスがクラスタリングされる。続いて、翻訳元文が入力された時に、翻訳元文から Reguler Expression が抽出され、その Reguler

Expression に対応するクラスタがドメインとみなされる。最後に、ドメイン依存言語モデルを用いた翻訳が行われる。クラスタ言語モデルと提案法の主な相違点は以下の通りである。

- クラスタ言語モデルのクラスタリング対象は、翻訳先言語のみであるが、提案法では、翻訳元、先双方が対象である。
- クラスタ言語モデルでは、人手でクラスタが定義されるが、提案法では自動的に決定される。
- クラスタ言語モデルでは、クラスタ依存言語モデルのみが用いられるが、提案法ではクラスタ依存翻訳モデルも用いられる。

5.1.2 文混合モデル

式 (4) において、D を隠れ変数とした場合、次のようになる。

$$\underset{e}{argmax} \sum_{D} P(D|f)P(e|D)P(f|e,D) \tag{9}$$

ここで、$P(D|f)$ の近似として $P(D) = D_\lambda$ を用い、ドメイン依存翻訳モデル $P(f|e,D)$ の代わりに非依存モデル $P(f|e)$ を用いるならば、この式は、文混合モデル [4] を表すことになる。

$$\underset{e}{argmax} \sum_{D} D_\lambda P(e|D)P(f|e) \tag{10}$$

文混合モデルと、提案法の主な相違点は以下の通りである。

- 文混合モデルでは、文混合比 D_λ は定数であるが、提案法では、入力文ごとに変わる。
- 文混合モデルでは全てのクラスタが考慮されるが、提案法では最尤のクラスタのみが考慮される。
- 提案法ではクラスタ依存翻訳モデルも用いられる。

6 評価実験

6.1 日英翻訳実験

6.1.1 実験コーパス

提案法の評価のために旅行対話コーパス [5] を用いた実験を行った。言語対は日英翻訳であり、訓練、デベロップメント (Dev)、評価コーパスは表 1 に示す通りである。この訓練セットは IWSLT2006 [12] 日英オープントラックに用いられたもので、評価セットは同じく IWSLT05 の評価セットとして用いられたものである。

6.1.2 実験条件

ドメイン定義のためのクラスタ数は予備実験で 10 とし、パープレキシティ計算は unigram を用いた。クラスタ依存言語モデルは SRI 言語モデルツールキット [8] を持ちて作成し、Good turing ディスカウント [6] を用いた 3-gram である。翻訳モデルの作成には GIZA++[7]、デコーディングには PHARAOH[9] を用い、デコーディングパラメータのチューニングには、minimum error training[10] を用いた。この際の参照訳の数は 16 である。デコーディング時には、ドメイン依存言語モデルは新しい特徴量として加え (すなわちドメイン非依存モデルとの log linear 補間)、ドメイン依存翻訳モデルは非依存モデルと linear 補間することで用いた。依存モデルと非依存モデルの補間比率は言語モデルで 6:4、翻訳モデルで 3:7 である。いずれの場合もデコーディングパラメータはベースラインのものと同じものを用いている。また、言語モデル関しては、ドメイン依存モデルと非依存モデルの重みの和が、ベースラインの現尾重みと同じになるように設定した。

表 *1.* 日英翻訳実験コーパス

	文数	総単語数	異なり単語数
訓練 (日)	40K	355K	12.5K
訓練 (英)	40K	315K	9.2K
Dev(日)	510	3,525	918
Dev(英)	510×16	57,388	2,118
評価 (日)	506	3,647	951

6.1.3 実験結果

以上の条件での実験結果を表 2 に示す。評価時の参照訳の数は 16 である。ドメイン依存の言語、翻訳モデルは単独でも Bleu、Nist 双方のメジャーで性能が向上しており、両者の組合わせでは Bleu で約 2.7 ポイントの向上となっている。

表 *2.* 日英翻訳実験結果

	Bleu	Nist
ベースライン	52.38	9.316
言語モデルのみ	53.66	9.349
翻訳モデルのみ	54.30	9.333
言語+翻訳モデル	55.09	9.451

6.2 先行研究との比較

先行研究であるクラスタ言語モデル (CLM) と文混合モデル (SMix) との比較実験を行った。実験コーパスは表 3 に示す通りで、言語対は中英である。このコーパスもまた IWSLT06 で用いられたものであり、英語のコーパスは、前の実験と同じものである。また、参照訳の数は 7 である。クラスタ言語モデルと文混合モデルの実験に関しては、IWSLT06 におけるアーヘン大学の実験結果 [11] との比較とした。この実験結果の訓練、評価データは表 3 と全く同じであるが、翻訳

システムが異なるため、若干ベースラインの性能が異なる (Bleu 値がアーヘン大学で 21.9、我々のシステムで 21.7)。このため、性能比較はそれぞれのベースラインからの向上値で評価を行った。実験結果を表 4 に示す。提案法は Bleu において、クラスタ言語モデルと文混合モデルを合わせたものより性能向上が大きく、Nist においてはクラスタ言語モデルと文混合モデルとも劣化しているにもかかわらず、提案手法では向上している。

表 *3.* 先行研究との比較実験コーパス

	文数	総単語数	異なり単語数
訓練 (英)	40K	315K	9.2K
訓練 (中)	40K	304K	18.7K
評価 (中)	489	5,110	1.3K

表 *4.* 先行研究との比較実験結果

	Bleu	Nist
RWTH	21.9	6.31
本システム	21.7	6.79
CLM	+0.6	-0.22
SMix	+0.2	-0.06
提案法	+1.1	+0.17

7 おわりに

本稿ではオンラインタスク適応による、ドメイン依存モデルの統計翻訳への利用を試みた。統計モデルにおいて、ドメイン依存モデルを利用することにより、性能が向上することが知られている。多くの場合、ドメインは未知、あるいは動的に変化する。このような場合、ドメインを動的に推定し、かつ推定されたドメイン依存のモデルを用いて翻訳を行う必要がある。提案法ではドメインを統計翻訳モデルの訓練データであるバイリンガルコーパスのサブコーパスとして定義する。このサブコーパスはエントロピーを基準としたクラスタリング手法により、自動的に構築される。翻訳元の入力文が与えられた時、その文に対し最も高い確率を与えるクラスタを選択することで、ドメイン推定が行われる。ドメイン推定が行われた後、そのクラスタから作られたクラスタ依存言語モデルと翻訳モデルを用いた翻訳が行われる。

IWSLT06 のデータを用いた評価実験の結果ベースラインのクラスタ非依存モデルを用いた時に比べ、Bleu 値で 2.7 ポイント (52.4 から 55.1) 性能が向上した。また、従来法である、クラスタ言語モデル、文混合モデルとの比較実験においてもより高い性能を示し、提案手法の有効性が確認できた。

REFERENCES

[1] K. Seymore, R. Rosenfeld, "Using Story Topics for Language Model Adaptation," Proc. EUROSPEECH, pp. 1987-1990, 1997.

[2] David Carter, "Improving Language Models by Clustering Training Sentences," Proc. ACL, pp. 59-64, 1994.

[3] S. Hasan, H. Ney, "Clustered Language Models Based on Regular Expressions for SMT," Proc. EAMT, Budapest, Hungary, May 2005.

[4] R. M. Iyer and M. Ostendorf, "Modeling Long Distance Dependence in Language: Topic mixture versus dynamic cache models," IEEE Transactions on Speech and Audio Processing, 1994.

[5] Genichiro Kikui, Eiichiro Sumita, Toshiyuki Takezawa, Seiichi Yamamoto, "Creating Corpora for Speech-to-Speech Translation," Proc. EUROSPEECH, pp. 381-384, 2003.

[6] S. M. Katz, "Estimation of Probabilities from Sparse Data for Language Model Component of a Speech Recognizer," IEEE Trans. on Acoustics, Speech, and Signal Processing, pp. 400-401, 1987.

[7] F. J. Och, H. Ney, "A Systematic Comparison of Various Statistical Alignment Models," Computational Linguistics, No. 1, Vol. 29, pp. 19-51, 2003.

[8] A. Stolcke, "SRILM - An Extensible Language Model Toolkit," http://www.speech.sri.com/projects/srilm/

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

[10] F. J. Och, "Minimum error rate training for statistical machine trainslation," Proc. ACL, 2003.

[11] A. Mauser, R. Zens, E. Matusov, S. Hasan, H. Ney, "The RWTH Statistical Machine Translation System for IWSLT 2006 Evaluation," IWSLT 2006, Nov. 2006.

[12] M. Paul, "Overview of the IWSLT 2006 Evaluation Campaign," IWSLT 2006, Nov. 2006.