LG_003

統計翻訳のための言語モデルのオンラインタスク適応

Online Language Model Task Adaptation for Statistical Machine Translation

山本 博史[1]　隅田 英一郎[1,2]
[1]情報通信研究機構 第二研究部門
知識創成コミュニケーション研究センター音声言語グループ
[2]ATR 音声言語コミュニケーション研究所
{hirofumi.yamamoto, eiichiro.sumita}@nict.go.jp

1　はじめに

近年、N-gramに代表される統計言語モデルは音声認識、統計翻訳をはじめとする言語処理において広く用いられている。統計言語モデルはその性格上、学習データと異なるタスクに対しては性能が劣化してしまう。この問題を防ぐために、通常「タスク適応」の手法が用いられる。本稿では、この「タスク適応」を統計翻訳に適用することを試みる。

「タスク適応」は、特定のタスクに特化したモデルであるタスク依存モデルを作成することが目的である。目的のタスクと一致した学習コーパスが大量に得られる場合は、特別な処理は不用であり、そのコーパスのみを用いて学習することによって、タスク依存モデルを作成することができる。しかしながら、一般に目的のタスクと一致した学習コーパスを大量に収集することは難しい。そのために、少量の目的のタスクと一致した学習コーパスと特定のタスクに偏らない大量の一般的な学習コーパスの二つを組み合わせてタスク依存モデルを作成する手法が「タスク適応」である。

このように、対象のタスクが既知である場合には、あらかじめ「タスク適応」を用いてタスク依存モデルを作成しておき、それを音声認識や統計翻訳に利用することができる。しかしながら、タスクをあらかじめ想定しておくことが困難な場合も多く、このような場合は通常の「タスク適応」の手法を用いることはできない。処理対象のタスクが未知の場合には、タスクの推定と、推定されたタスクに対する適応を同時に行う必要がある。このための手法がオンラインタスク適応である。本稿ではこのオンラインタスク適応を統計翻訳のための言語モデルに対して適用することを試みる。本手法では、まず入力された翻訳原文から、そのタスクを推定し、そのタスクに依存した翻訳目的言語のタスク依存言語モデルを用いることで、オンラインタスク適応の問題の解決を図る。

2　タスク適応

統計翻訳は次式に示されるように、与えられた翻訳原言語単語列 e に対し、確率が最大となる翻訳目的言語単語列 f を見つける問題である。

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

この式においては、翻訳先目的単語列 f は翻訳原単語列 e のみで決るが、実際はタスク T の環境の影響を大きく受ける。この、タスク T が既知の場合は T を新たな変量として式(1)に導入することにより、次式が得られる。

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

この式はベイズ則を用いて次式のように書き換えることができる。

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

ここで、$P(e|f,T)$ がタスク依存翻訳モデル、$p(f|T)$ がタスク依存言語モデルである。

2.1　オンライン適応

式(2)を用いる場合は、あらかじめ適応モデルを構築しておく、すなわちオフライン適応が可能であるが、タスク T は必ずしも既知ではない。この場合はタスク T と翻訳目的単語列 f を同時に推定する問題、すなわちオンライン適応として、次式のように表わされることになる。

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

この式と式(2)との大きな違いは、式(4)においてはタスクを表わす変量 T が隠れ変数となっていることである。またこの式の右辺の $P(T|e)$ はタスク推定、$P(f|e,T)$ はタスク適応を表わしていることになる。

3 タスク推定

式(4)を満たす翻訳目的単語列fを求めるためには$P(T|e)$と$P(f|e,T)$を同時に最大化する必要がある。しかしながら、これは困難であるため、近似としてまず$P(T|e)$を最大化し、それによって求まったTを用いて$P(f|e,T)$を最大化するという手順をとることにする。

3.1 タスクの規定

オフライン適応の場合、タスクはトピック等の人間の感覚に合ったものとして、あらかじめ規定されていることが多い。しかしながら、オンライン適応の場合はタスクは隠れ変数として用いられ、外部に出力する必要もない。このため、まずタスクそのものを自由に規定しておくことが可能である。この場合のタスクは、必ずしも人間の感覚に合ったものである必要はないため、タスクTは統計的な観点から、式(4)をなるべく大きく、すなわち$P(T|e)$と$P(f|e,T)$をなるべく大きくできるように規定することが望ましい。

そこで、この近似として、$P(f|e,T)$を$P(f|T)$で置き換えた$P(T|e)P(f|T)$を最大化するようなTを規定することを試みる。最大化の対象である$P(T|e)P(f|T)$は、ベイズ則を用いて次式のように書き換えることができる。

$$P(f|T)P(e|T)P(T)/P(e) \tag{5}$$

ここで、$P(e)$はTに無関係であり、さらに$P(T)$を定数とする近似を導入することによって、規定すべきタスクTは次式で表わされることになる。

$$\underset{T}{argmax}\ P(f|T)P(e|T) \tag{6}$$

この式はfとeに対して同時に確率を最大化する、すなわちfとeの尤度の和を最大化するようにTを規定することを意味している。

3.2 対訳文対クラスタリング

前節で述べたように、タスクTとしてはfとeの尤度の和を最大化するように設定すればよい。これはすなわち、fとeの対訳文対をエントロピー最小化の基準の元にクラスタリングを行えばよいことを意味している。本稿では、具体的には以下の手順でクラスタリング[1]を行うことにする。

1. あらかじめタスク数、すなわちクラスタ数を定めておく。
2. 全ての対訳文対をランダムにクラスタに配置する。
3. 各クラスタに含まれる対訳文対を用いてfとeに対する言語モデルをそれぞれ、クラスタごとに作成する。
4. クラスタごとに作成された言語モデルを用いてfとeに対するエントロピーを求め、それの全クラスタに対する和を総エントロピーとする。
5. 全ての対訳文対一つ一つに対して、総エントロピーが最小となるようなクラスタへの移動を行う。
6. 総エントロピーの変化量があらかじめ定めておいた閾値よりも小さくなるまで、上の操作を繰り返す。

3.3 クラスタ選択とタスク適応

前節で、タスクが規定されたため、続いてこのタスクTを入力されたeから推定することになる。これはすなわち$P(T|e)$を最大化するTを見つけることである。$P(T|e)$はベイズ則を用いて、$P(e|T)P(T)/P(e)$と書き換えることができ、クラスタリング時に用いた、$P(T)$を定数とする近似を導入すれば、$P(e|T)$を最大化すればよいことになる。これはすなわち、最尤推定を用いた場合にはeに対して、最大尤度を与えるタスク、すなわちクラスタTを選べばよいことになる。

次に推定されたTを用いてタスク適応、すなわち$P(f|e,T)$の最大化を図る。$P(f|e,T)$はベイズ則を用いて次式のように書き換えることができる。

$$\begin{aligned} &P(f|e,T) \\ &= P(e|f,T)P(f,T)/P(e,T) \\ &= P(e|f,T)P(f|T)P(T)/P(e,T) \end{aligned} \tag{7}$$

ここで、eとTは既知であるため、上式を最大化するためには$P(e|f,T)P(f|T)$を最大化すればよいことになる。この式で、$P(e|f,T)$がタスク依存(タスク適応後)翻訳モデル、$P(f|T)$がタスク依存(タスク適応後)言語モデルであり、これらのタスク依存モデルをもちいてeからfを推定することを意味している。本稿では、このタスク依存モデルのうち、言語モデルのみについて言及することとする。

4 タスク適応の手順

前章で述べた方法を用いて、実際にタスク適応を行う手順をのべる。この手順には、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、実際に翻訳原文が入力された時に行うオンラインプロセスが含まれる。

4.1 オフラインプロセス

タスク規定のための対訳文対のクラスタリングと、そのクラスタ依存言語モデルを作成する手順である。

1. 翻訳モデル、言語モデルの学習データである対訳文対に対し、クラスタリングを行う。
2. 各クラスタごとに翻訳原言語、翻訳目的言語のクラスタ依存言語モデルを作成する。

4.2 オンラインプロセス

入力された翻訳原文それぞれに対し、以下の手順を行う。

1. 翻訳原文に対し、最も高い尤度を与えるクラスタを選択する。
2. 選択されたクラスタの翻訳目的言語のクラスタ依存言語モデルを用いて翻訳を行う。

5 評価実験

5.1 実験コーパス

本手法の有効性を確認するため実験を行った。対象としたドメインは旅行対話で、用いたコーパスは「ATR旅行対話基本表現集 (BTEC)」[2] である。翻訳言語対は、日本語から英語であり、学習、および評価コーパスのサイズ等は表1、表2に示すとおりである。

表 1: 評価実験学習コーパス

言語	文数	総単語数	異なり単語数
日本語	162K	1,449K	18.7K
英語	162K	1,303K	21.0K

5.2 言語モデル

作成した言語モデルは、日本語、英語とも Good-Turing[3] で平滑化された単語 3-gram である。入力された日本語文に対し、最大尤度を与えるクラスタを選択する際のモデルとしては、クラスタ依存の日本語単語 3-gram をそのまま用いている。

表 2: 評価実験評価コーパス

言語	文数	総単語数	異なり単語数
日本語	1,524	13,686	2,023
英語	1,524	12,364	1,723

一方、選択されたクラスタに対して、翻訳時に用いられる英語の言語モデルとしては、クラスタ依存の英語単語 3-gram と全ての英語学習データを用いて作成したクラスタ非依存単語 3-gram を線形補間 [4] したものを用いた。この理由は、クラスタ非依存モデルでは、学習データの量がクラスタ依存モデルに比べ、必然的に少なくなり、データスパースネスの問題を引き起こす可能性があるためである。

5.3 クラスタ数と性能の関係

まず、対訳文対に対して行うクラスタリングの際の、クラスタ数を変化させた時の提案法の性能評価を行った。評価基準は翻訳目的言語である英語の評価セットに対するパープレキシティである。この時、対象の英語言語モデルにおける、クラスタ依存モデルとクラスタ非依存モデルの線形補間係数は０．５で固定した。

変化させたクラスタ数は、５、１０、２０である。結果を図１に示す。図１において左側の軸目盛り"Perplexity"がパープレキシティであり、クラスタ数を変化させた時の値が点線で示されている。クラスタ数が１の場合は適応を行っていない場合、すなわちベースラインを示している。また、右側の軸目盛り"Reduction Rate"は適応を行うことによってエントロピーが減少した評価セット文の割合をしめしている。すなわち、この値が８９であれば、評価セット文１、５２４文のうち、１、３５７文が適応によってエントロピーが減少し、残りの１６７文では逆に増加したことを示している。

図１に示される通り、クラスタ数の増加と共に適応後の言語モデルのパープレキシティは減少している。その値はクラスタ数２０の場合で、ベースラインの２６．９から１８．６に減少しており、割合では約３１％となっている。また、この時エントロピーが減少した評価セット文の割合も８９％と高く、特定の文に限って効果が現れているわけではないことを示している。

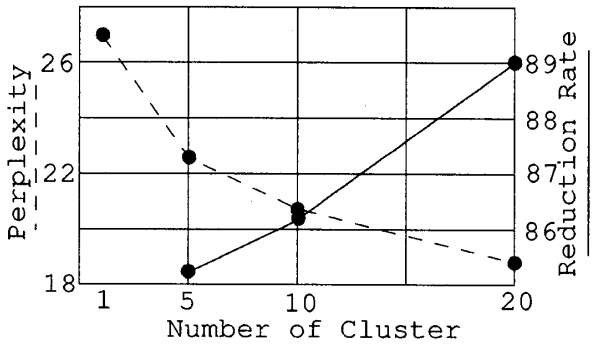

図 1: クラスタ数と性能の関係

5.4 補間係数と性能の関係

続いて、クラスタ依存モデルとクラスタ非依存モデルの線形補間係数を変化させた場合の性能変化を調べた。この時のクラスタ数は前節の評価で最もパープレキシティの低かった２０に固定した。変化させた補間係数はクラスタ依存モデルに対して０．０５、および０．１から０．９まで０．１刻みである (０の場合はクラスタ非依存モデルのみを使うことを意味している)。このうち、０．５から０．９のあいだの結果を図２に示す。左右の軸は図１と同様である。

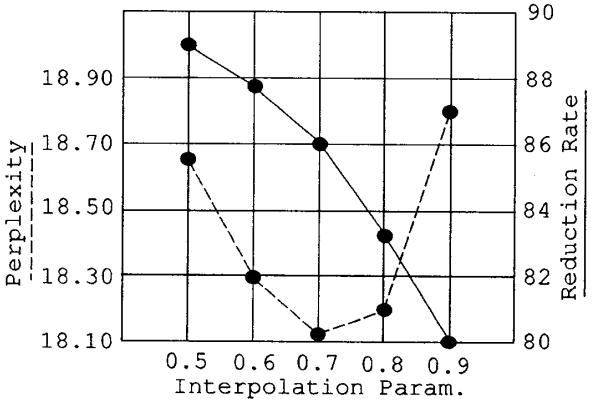

図 2: 補間係数と性能の関係

図２に示されるように、補間係数０．７でパープレキシティは最小値１８．１を示しており、これはベースラインの２６．９に対して約３３％の減少となっている。しかしながら、エントロピー減少文の割合は補間係数が大きくなるに従って減少している。この割合は図には示されていないが、補間係数０．１で最小となり、その値は９２％である。またこの時のパープレキシティは２３．１であった。エントロピー減少文の割合が減る理由としては、補間係数が大きくなると共に、線型補間されたモデルのクラスタ依存性が高まるが、逆にカバレッジが減少するためと考えられる。

6 まとめ

本稿では、統計翻訳のための言語モデルのオンラインタスク適応に関する提案を行った。一般のタスク適応では適応対象のタスクは既知の条件で行われるが、事前にタスクが与えられない場合も十分に考えられる。このような場合、タスクの推定と適応を同時に行う、オンラインタスク適応が有効である。

提案手法では、このオンラインタスク適応のために、対訳文対に対するクラスタリングを利用する方法を用いる。この対訳文対のクラスタは言語モデル適応におけるタスクとして見なされる。このタスク推定の方法としては、入力された翻訳原言語文に対して最も低いパープレキシティを与えるクラスタを選択するという方法をとる。そして、最終的に翻訳目的言語に対する、選択されたクラスタ依存言語モデルを用いることで性能向上を図る。

旅行対話文を対象とした日英翻訳の実験では、翻訳目的言語である英語の言語モデルに対し、パープレキシティで約３３％の削減をはかることができ、提案手法の有効性を確認することができた。また、本稿では、言語モデルのオンラインタスク適応のみを試みたが、本手法はそのまま翻訳モデルのオンラインタスク適応にも応用が可能であると考えられ、その場合、さらなる性能の向上が期待できる。

参考文献

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

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

[3] 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.

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