構文トランスファ方式統計翻訳への誤り率最小訓練法の導入

今村 賢治, 大熊英男, 隅田英一郎
ATR 音声言語コミュニケーション研究所
{kenji.imamura, hideo.okuma, eiichiro.sumita}@atr.jp

1 はじめに

統計翻訳は、入力された原言語の単語列に対し、最も生成確率の高い目的言語の単語列を、すべての組み合わせから探索することにより翻訳を行う方式である。Brown et al. (1993) では、ソースチャネルモデルを用い、翻訳モデルと言語モデルに分解して翻訳を行っていた。しかし近年、より拡張性の高い log-linear モデル (Och and Ney, 2002) が主流となりつつある。これは、各モデル確率の対数値を特徴関数 (feature function) として定義し、その重み付き線形和でスコアを定義するものである。

Log-linear モデルを用いる場合、重要となるのは以下の 2 点である。

- 特徴関数として何を用いるか
- 特徴関数同士の重みの最適化

前者に関しては、翻訳方式により異なる。たとえば、Brown et al. (1993) の用いたソースチャネルモデルでは、言語モデル、翻訳モデルが特徴関数に対応する。

後者に関しては、自動評価と組み合わせて、開発セットに合わせて重みを最適化する、誤り率最小訓練法 (Och, 2003) がある。

本論文では、構文トランスファ方式統計翻訳を log-linear モデルに適合させ、いくつかの特徴関数を追加する。そして、誤り率最小訓練法を適用し、その効果を調査する。

2 誤り率最小訓練法

2.1 Log-linear モデル

統計翻訳は、与えられた原言語の単語列 $\mathbf{f}$ に対し、最も確率の高い目的言語の単語列 $\mathbf{e}$ を探索する。Brown et al. (1993) の用いたソースチャネルモデルでは、これをベイズ規則を用い、言語モデル、翻訳モデルに分解してモデル化している。

$$\begin{aligned}\hat{\mathbf{e}} &= \underset{\mathbf{e}}{\operatorname{argmax}} P(\mathbf{e}|\mathbf{f}) \\ &= \underset{\mathbf{e}}{\operatorname{argmax}} P(\mathbf{e})P(\mathbf{f}|\mathbf{e}). \end{aligned} \quad (1)$$

一方、log-linear モデル (Och and Ney, 2002) は、事後確率を直接モデル化する。すなわち、

$$P(\mathbf{e}|\mathbf{f}) = \frac{1}{Z(\mathbf{f})} \exp\left(\sum_{m=1}^{M} \lambda_m h_m(\mathbf{e},\mathbf{f})\right). \quad (2)$$

ここで、$Z(\mathbf{f})$ は、正規化用の定数である。したがって、式 (2) を最大化する目的言語の単語列 $\hat{\mathbf{e}}$ は、以下の式で表わされる。

$$\begin{aligned}\hat{\mathbf{e}} &= \underset{\mathbf{e}}{\operatorname{argmax}} P(\mathbf{e}|\mathbf{f}) \\ &= \underset{\mathbf{e}}{\operatorname{argmax}} \left\{\sum_{m=1}^{M} \lambda_m h_m(\mathbf{e},\mathbf{f})\right\}. \end{aligned} \quad (3)$$

ここで、$h_m(\mathbf{e},\mathbf{f})$ は、原言語・目的言語単語列に依存した特徴量を出力する、特徴関数である。λ_m は、特徴関数の重みである。つまり、log-linear モデルを用いた統計翻訳は、各特徴関数値の重み付き和が最大になる目的言語の単語列を探索していることになる。

Log-linear モデルには、以下の利点がある。

- ソースチャネルモデルを特殊ケースとして含んでいる。つまり、特徴関数として言語モデル確率、翻訳モデル確率を用い、

  $$\begin{aligned} h_1(\mathbf{e},\mathbf{f}) &= \log P(\mathbf{e}), \\ h_2(\mathbf{e},\mathbf{f}) &= \log P(\mathbf{f}|\mathbf{e}), \\ \lambda_1 &= \lambda_2 = 1 \end{aligned}$$

  とすると、式 (3) は式 (1) と完全に一致する。
- 上記特徴により、ソースチャネルモデルに比べ拡張性が高い。ソースチャネルモデルの枠組みでは、数学的適合性を保ったまま拡張するのは困難な場合があるが、log-linear モデルでは、特徴関数を追加することにより、容易に拡張できる。たとえば、翻訳文の構文的正しさを検証したい場合、構文木の生成確率を特徴関数として加える。

[0]本研究は独立行政法人 情報通信研究機構の研究委託「大規模コーパスベース音声対話翻訳技術の研究開発」により実施したものである。

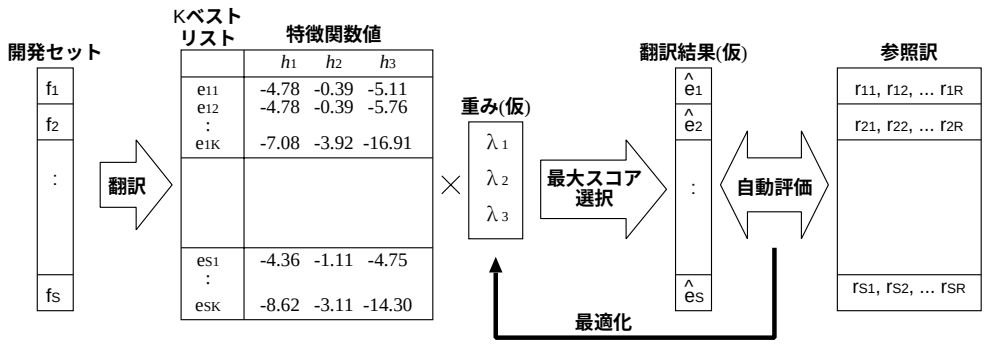

図 1: 誤り率最小訓練法概要

2.2 誤り率最小訓練法

Log-linear モデルは、特徴関数の重みにより、生成される翻訳文が著しく変化するため、重みを最適化する必要がある。その一つの手法が、誤り率最小訓練法 (Och, 2003) である。簡単に言うと、これは機械翻訳の自動評価を用いて、その自動評価値を最大化（逆に言うと、誤り率を最小化）するよう、特徴関数の重みを設定するものである。自動評価には、人間が翻訳した参照訳との類似度を数値化する手法である BLEU (Papineni et al., 2002) や単語誤り率 (mWER, (Nießen et al., 2000)) などが用いられる。

図 1は、誤り率最小訓練法を摸式化したものである。まず翻訳器は、開発セット (テストセットとは異なる) の原文を、ある適当な値に定められた重みを使って翻訳する。この時、最適な 1 文だけを出力するのではなく、スコアの高い K 文を、各特徴関数値とともに出力する。これを K ベストリストと呼ぶ。

誤り率最小訓練は、仮の重みを設定し、特徴関数値の総和を算出することにより、K ベストリストから最大スコアを持つ翻訳文を選択する。すると開発セットの仮の翻訳結果 (1 ベスト) が得られるので、それを自動評価する。そして、仮の重みを変化させて最も自動評価値が高くなる点を探索する。

重みが局所解で停止する場合もあるため、重み (仮) の初期値をランダムに複数個設定し、上記処理を並列的に実行し、最も自動評価値が向上したものを選択することにより、局所解になることを避けている。

3 構文トランスファ方式統計翻訳

構文トランスファ方式は、原言語の構文木を目的言語の構文木に変換することにより、翻訳文を生成する方式である。図 2にその例を示す。入力された原文は、構文解析され、木構造が作られる。そして、そのノード単位に目的言語に変換し、目的言語の木構造を作成する。それを直線化すると、翻訳文が作られる。このような処理を行うためには、３つのルールセットが必要である。すなわち、

- 原言語の構文木を作るための CFG 規則
- 目的言語の構文木を作るための CFG 規則
- 原言語・目的言語間の CFG 規則同士の対応

これを統計モデルで表現するため、Imamura et al. (2005) では、翻訳モデルを原言語→目的言語、目的言語→原言語双方向に適用し、以下の 4 つの確率で表わした。なお、$\mathcal{F}$、$\mathcal{E}$ はそれぞれ原言語、目的言語の構文木、θ, π は、それぞれ $\mathcal{F}$、$\mathcal{E}$ を構成する CFG 規則で、$P(\theta)$, $P(\pi)$ は、CFG 規則の書き換え確率 (左辺が右辺を生成する確率) である。

- 原言語の構文木の生成確率 (原言語木構造モデル)

$$P(\mathcal{F}) = \prod_{\theta:\theta\in\mathcal{F}} P(\theta). \tag{4}$$

- 目的言語の構文木の生成確率 (目的言語木構造モデル)

$$P(\mathcal{E}) = \prod_{\pi:\pi\in\mathcal{E}} P(\pi). \tag{5}$$

- 原言語・目的言語間の CFG 規則の対応確率 (順方向、逆方向。木構造マッピングモデル)

$$P(\mathcal{E}|\mathcal{F}) = \prod_{\substack{\theta:\theta\in\mathcal{F},\\ \pi:\pi\in\mathcal{E}}} P(\pi|\theta), \tag{6}$$

$$P(\mathcal{F}|\mathcal{E}) = \prod_{\substack{\theta:\theta\in\mathcal{F},\\ \pi:\pi\in\mathcal{E}}} P(\theta|\pi). \tag{7}$$

これらの確率値は、確率文脈自由文法における内側確率に相当するため、ボトムアップパーザの枠組みを拡張することにより翻訳を行うことができる。また、これらの対数値は、そのまま log-linear モデルの特徴関数値となる ($h_1 \sim h_4$)。

なお、言語モデルについては単語 trigram を使用している (h_5)。

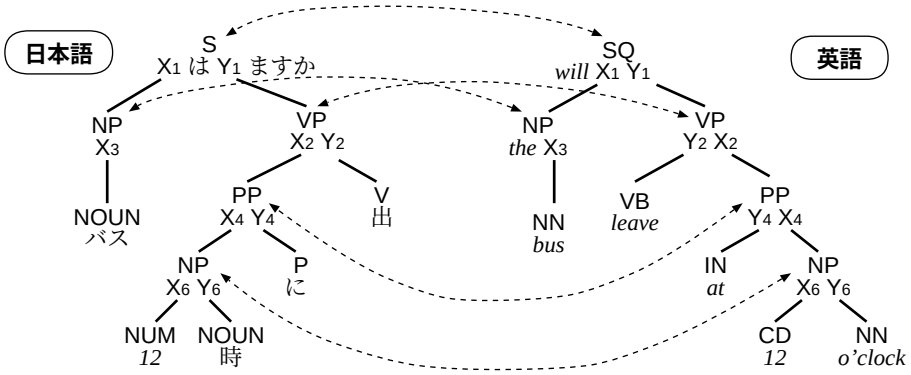

図 2: 構文トランスファ方式による翻訳例

4 句に基づく統計翻訳に用いられる特徴関数の導入

構文トランスファ方式は、基本的には単語翻訳を構文情報を用いて文を組み立てる方式であるが、長単位句 (複数単語列) を単位としても翻訳文を作ることができる。Imamura et al. (2005) は、長単位句と単語を同等に用い、両者の翻訳結果を統計モデルで選択することにより、構文トランスファ方式統計翻訳を実現した。したがって、Imamura et al. (2005) の方式は句に基づく統計翻訳 (Phrase-based SMT) の特徴もあわせ持ち、そこで用いられている特徴関数を導入することが可能である。本稿では、句の統計翻訳の一つである (Zens and Ney, 2004) を参照し、以下の 4 関数の追加を試みる。

- 単語の翻訳確率（順方向および逆方向の 2 関数）。ここではIBM Model 1 を想定する。ただし、CFG 規則単位でデコードを行えるよう、対応する原言語規則・目的言語規則内に含まれる単語だけで対応をとる。順方向の単語翻訳確率は以下の式で表わす。逆方向の単語翻訳確率は、この原言語と目的言語を入れ替えたものである (h_7)。

$$h_6(\mathbf{e},\mathbf{f}) = \sum_{\substack{\theta:\theta\in\mathcal{F},\\ \pi:\pi\in\mathcal{E}}} \sum_{e:e\in\pi} \log\left(\sum_{f:f\in\theta} p(e|f)\right). \quad (8)$$

  木構造マッピングモデル確率は低頻度規則では高確率 (つまり確率 1.0) になる場合が多く、単語の翻訳確率はこれを平滑化する効果がある。

- 単語数、および句数のペナルティ。いずれも、目的言語の単語数、句数を直接特徴関数値として使用する。単語数のペナルティの重みが正の場合、長い（単語数が多い）翻訳文、負の場合は短い翻訳文が生成される。句数のペナルティは、重みが正の場合、より細かい単位で木構造マッピングが行われ、負の場合は大きな単位でマッピングが行われる。

$$h_8(\mathbf{e},\mathbf{f}) = \text{目的言語の単語数}, \quad (9)$$
$$h_9(\mathbf{e},\mathbf{f}) = \text{構文木のノード数}. \quad (10)$$

表 1: コーパスサイズ

セット名	項目	日本語	英語
訓練	文数	152,170	
	単語数	1,178,682	1,103,655
開発セット	文数	500	
	単語数	4,019	3,731
テストセット	文数	506	-
	単語数	4,066	-

5 実験

5.1 実験条件

今回、日英翻訳を対象に実験を行った。訓練に使用したコーパスは、旅行会話に頻繁に用いられる文を集めたコーパス (BTEC) である。コーパスサイズを表 1に示す。開発セットには、Internaltional Workshop on Spoken Language Translation で用いられた 2004 年のテストセットを用い、テストセットには同会議の 2005 年テストセットを用いた。

また、誤り率最小訓練に用いた自動評価指標は BLEU で、1000 ベストリストから最適化を行った。最適化結果は、初期重みによって異なる場合があるため、ランダムに 10 種類を用い、2 回実験を行った。

5.2 実験結果

自動評価として、BLEU, NIST (Doddington, 2002), mWER の結果を表 2に示す。なお、使用した参

表 2: 自動評価による翻訳品質評価

特徴関数	最適化	回	BLEU	NIST	mWER
構文 (4 関数)+ 言語モデル	なし	-	0.595	8.68	0.333
構文 (4 関数)+ 言語モデル	あり	1 回目	0.615	8.85	0.311
	あり	2 回目	0.609	8.72	0.322
構文 (4 関数)+ 句 (4 関数)+ 言語モデル	あり	1 回目	**0.627**	**9.22**	**0.298**
	あり	2 回目	0.621	8.46	0.311

表 3: 主観評価による翻訳品質評価

特徴関数	最適化	回	A	AB	ABC
構文 (4 関数)+ 言語モデル	なし	-	276 (54.5%)	329 (65.0%)	373 (73.7%)
構文 (4 関数)+ 言語モデル	あり	1 回目	289 (57.1%)	**342 (67.6%)**	376 (74.3%)
	あり	2 回目	287 (56.7%)	336 (66.4%)	371 (73.3%)
構文 (4 関数)+ 句 (4 関数)+ 言語モデル	あり	1 回目	**294 (58.1%)**	341 (67.4%)	**386 (76.3%)**
	あり	2 回目	280 (55.3%)	331 (65.4%)	371 (73.3%)

照訳数は 1 原文あたり 16 である。また、主観評価として、 A(完全訳), B (部分訳), C(理解可能訳), D(不可訳) の 4 段階で評価した結果を表 3に示す。 mWER のみ、低い値ほど良好な翻訳であることを表わす。

まず、自動評価結果に着目すると、誤り率最小訓練を行うことにより、 BLEU, mWER は評価値が向上した。 log-linear モデル + 誤り率最小訓練は、構文トランスファ方式でも品質を向上させることができる。句に基づく統計翻訳の特徴関数を追加することにより、自動評価値はさらに向上した。

しかし、 2 回目の訓練では句の特徴関数を追加することにより、 NIST スコアが最適化なしに比べて若干低下した。これは、 BLEU で最適化を行ったこと、および重みが局所解で停止したことが原因である。特に、局所解で停止した場合、翻訳結果が不安定になることがある。このことは、主観評価結果を見ても分かる。 1 回目の訓練では A, AB, ABC ランクいずれも最適化前に比べ、品質が向上したが、 2 回目の訓練では、 ABC ランクが若干低下した。このように、誤り率最小訓練法は、初期重みの設定により、翻訳品質がぶれることがある。初期重み数を増やすなどして、大域的に最適な重みを設定すれば、訳質は安定すると考えられる。

6 まとめ

本稿では、構文トランスファ方式統計翻訳を log-linear モデルに適応させ、誤り率最小訓練を行った。その際、句に基づく統計翻訳で用いられている特徴関数を追加した。その結果、自動評価の観点では翻訳品質が向上した。主観評価の観点では、ほぼ向上したが、初期重みの設定により、品質向上しなかった場合もあった。誤り率最小訓練法は、局所解で停止する可能性があり、大域最適解に近付けることにより、安定した品質向上が望める。

参考文献

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the HLT Conference*.

Kenji Imamura, Hideo Okuma, and Eiichiro Sumita. 2005. Practical approach to syntax-based statistical machine translation. In *Proceedings of Machine Translation Summit X*, pages 267–274.

Sonja Nießen, Franz Josef Och, Gregor Leusch, and Hermann Ney. 2000. An evaluation tool for machine translation: Fast evaluation for MT research. In *Proceedings of LREC 2000*, pages 39–46.

Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of ACL 2002*, pages 295–302.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the ACL 2003*, pages 160–167.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL 2002*, pages 311–318.

Richard Zens and Hermann Ney. 2004. Improvements in phrase-based statistical machine translation. In *HLT-NAACL 2004: Main Proceedings*, pages 257–264.