統計的後編集手法を適用したルールベース翻訳と文レベルの自動品質評価との融合

(株)東芝　研究開発センター　知識メディアラボラトリー

鈴木 博和
Hirokazu Suzuki

1. まえがき

機械翻訳における翻訳品質は日々向上しているが、人手翻訳の品質と比較するとまだまだ不十分であり、人手によるチェック・後編集は不可欠である。翻訳業務において機械翻訳を有効に活用するためには、後編集を如何に効率よく行うかが重要であり、機械翻訳に対して人手によるこの後編集工程の負担が小さくなるような手法の需要は大きい。

ルールベース機械翻訳方式(RBMT)は、統計ベース機械翻訳方式(SMT)と比較して翻訳の品質が安定しているのが特徴であり、後編集作業をルーチン化しやすいという利点がある。しかし、これは同時に欠点にもなりうる。即ち、同じ間違いを繰り返し犯すために、反復的な後編集作業を強いられることを意味する。

この「機械翻訳で頻出する誤りを別の表現に修正する」という作業は SMT と相性が良く、SMT を用いて自動的に後編集する手法が提案されている [Simard, et al., 2007] [Lagarda, et al., 2009] [江原暉将, 2006] [江原暉将, 2008] [村上, ほか, 2010]。これらの手法の特徴は、ある言語から別の言語への翻訳に SMT を用いるのと、全く同じ枠組みで自動後編集モジュールを実現できることにある。自動後編集に SMT を用いる場合、機械翻訳結果を原文側に、それに対応した人手後編集結果（あるいは人手翻訳結果）を訳文側においたパラレルコーパスを用いて SMT の翻訳モデルを学習し、訳文側データを用いて言語モデルを学習する。本研究では、まず NTCIR 7日英特許翻訳タスク [The 7th NTCIR Workshop, 2007/2008]の訓練データ・開発データと句ベース SMT [Koehn, et al., 2003]の枠組みを用いて自動後編集モジュールを実現した。

この SMT による自動後編集は、言語モデル・翻訳モデルのドメインに依存して行われるため、RBMT 結果のドメイン適応と考えることもできる。この場合、RBMT 結果の中には、SMT による自動後編集を行う必要がないものも存在するため、そのような訳文を自動的に検出し、自動後編集を選択的に行えることが好ましい。

本研究では、参照訳なしに機械翻訳の品質を予測する手法として Confidence Estimation [Specia, et al., May 2009] [Specia, et al., 2009]に着目し、そこで有効な手法として認められている PLS 回帰分析手法を用いて翻訳品質予測モデルを構築した。さらにこのモデルを使って自動後編集を選択的に適用した場合の NIST スコアの変化を調べ、本手法の有効性を示した。

2. SMT を用いた自動的後編集

本研究では日英翻訳方向を対象にする。

図 2 は本研究で提案するシステムの概要図を表わす。本節の内容は図 2 の SMT-based Automatic Post-Edit(APE)に相当する。

句ベース SMT として Moses [Koehn, et al., 2007]を用いた。データには NTCIR7 の日英特許翻訳タスクのデータを用い、日英パラレルコーパス約 180 万文対から長文を除いた約 118 万文対の原文に対して日英 RBMT を行い、その翻訳結果と対応訳とのパラレルコーパスを翻訳モデル用訓練データに用いた。言語モデル用データには上記長文を含めた約 180 万文を用いた。

モデル		文数	単語数	
			RBMT 側	Reference 側
翻訳モデル	training	1,184,827	33,719,825	33,356,416
	dev	805	26,277	25,681
言語モデル		1,798,571		59,429,838

表 1：使用したデータ

また、訓練時のパラメータは以下のように設定した：

- 言語モデル：SRILM を用い、3-gram を用いて学習した。ngram-count に指定したオプションは -order 3 -interpolate –kndiscount である。
- 翻訳モデル：オプションに-alignment grow-diag-final-and -reordering msd-bidirectional-fe を指定して学習した。

テストデータには同じく NTCIR7 日英特許翻訳タスクからの 899 文を用いた。図 1 は RBMT 翻訳結果(raw translation)と SMT を用いた自動後編集結果(SMT based APE)の NIST 値 [NIST, 2002]の変化を表わす。

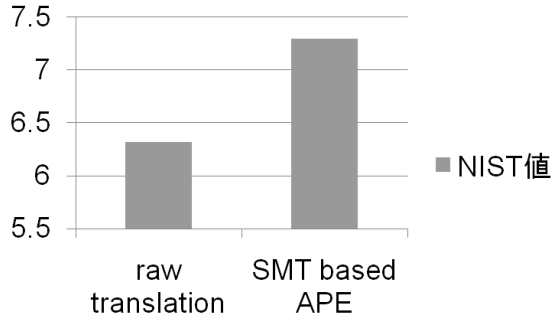

図 1：SMT を用いた自動後編集の効果

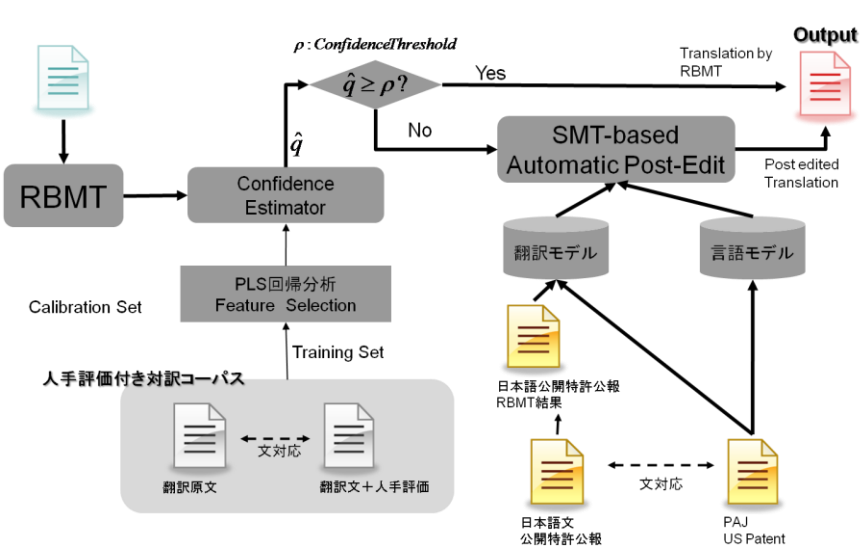

図２：システム概要図

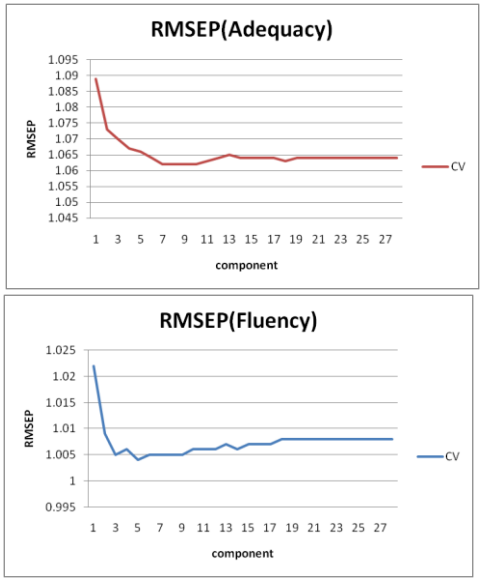

図３：潜在変数の数と RMSEP

自動後編集の結果、NIST 値において約 15%程度の向上が見られる。

3. 参照訳を必要としない自動評価手法

翻訳結果の自動評価に用いられる BLEU や NIST は参照訳を必要とする。テスト用の原文に対して複数の参照訳を準備するのはコストがかかり、またその参照訳が正解のすべてではないという問題がある。このような問題に対し、参照訳を必要とせずに文レベルの翻訳品質の予測を行う confidence estimation(CE)と呼ばれる手法がある。[Blatz, et al., 2003]では、CE を翻訳が'good'か'bad'かを判別する 2 値分類の問題としてとらえている。[Specia, et al., May 2009] [Specia, et al., 2009]では、CE を翻訳品質のスコアが連続値で与えられるような問題と考え、このモデルで良い推定値を与える PLS 回帰分析を用いて、SMT の翻訳品質予測手法を提案している。

本研究では RBMT の翻訳品質を評価するべく feature set を定義し、これを用いて PLS 回帰分析による品質予測モデルを構築する。

3.1.PLS 回帰分析

はじめに本研究で用いる PLS 回帰分析の特徴について述べる。

一般に回帰分析とは、ある変数(説明変数)を使って予測したい変数(目的変数)を説明することであり、特に説明変数が複数の場合は重回帰分析と呼ばれる。人手翻訳のスコアを目的変数とし、翻訳の様々な誤りの数を説明変数として重回帰分析手法を用いて自動評価を行おうとする試みが [Zhu, et al., 2009]で行われている。

しかし重回帰分析は暗黙的に説明変数間が無相関であることを前提としているので、複数の説明変数間で関連性が存在すると正しい予測を行うことができないことが知られている(多重共線性)。

Partial Least Squares(PLS)回帰分析 [Wold, et al., 1984]はこのような共線性が見られる場合に予測モデルを精度よく構築できる優れた方法として知られている。

機械翻訳において参照訳を必要としない自動評価を回帰分析にて行いたい場合、説明変数として使用する feature 間が無相関であるという保証はなく、また無相関となるように feature set を決めることも困難である。従ってこのような場合には PLS 回帰分析を用いると高精度な品質予測モデルを構築できることが期待できる。

3.2. 品質予測モデルの構築

3.2.1. Feature Set

はじめに PLS 回帰分析の入力変数として使用する feature set を表 4 のように定義した。ここではコーパスとして、自動後編集用 SMT の言語モデル学習に用いた単言語コーパスを用いた。

英語形態素解析器として TreeTagger [Schmid, 1994]を、英語構文解析器として Link Grammar Parser[Grinberg, et al., 1999]を用いた。

3.2.2. Trainig Set/Test Set

NTCIR7 日英特許翻訳タスクに参加したシステムから無作為に選択した 8 システムの翻訳結果に対し feature を求め、それを入力変数とし、人手評価結果(Adequacy/Fluency の 5 段階評価)を予測変数として訓練を行い、予測モデルを構築した。また、上記 8 システムを除くシステムから 2 システムを無作為に選択し、その翻訳結果・人手評価結果を評価用のデータセットとして用いた。

3.2.3. 潜在変数

PLS 回帰分析では、入力変数を説明変数として直接回帰を行うわけではなく、入力変数と予測変数から潜在変数(component)を抽出し、この潜在変数を説明変数として回帰を行う。

本研究では潜在変数の数を cross-validation により決定した。用いた指標は RMSEP(Root Mean Squared Prediction Error):

$$RMSEP = \sqrt{1/N \sum_{j=1}^{N} (y_j - y'_j)^2}$$

を用いた（y は実測値、y'は予測値を表わす）。図 3 はその結果を表わし、使用する潜在変数の数は Adequacy 予測時には 8、Fluency 予測時には 5 とした。

3.2.4. 評価用データに対する予測結果

評価用データセットに対するモデルの予測結果は表 2 のようになった。

Adequacy Prediction	Test System1	Test System2
RMSPE	1.07	1.48
Spearman 順位相関係数	0.25	0.41
Fluency Prediction		
RMSEP	0.86	1.28
Spearman 順位相関係数	0.30	0.37

表 2：評価用データに対する予測結果

評価用に用いた両システムに対し、いずれも Fluency の予測の方が RMSEP が小さくなっている。これは用いた feature set が訳文に関する情報のみに基づいていることが原因と考えられる。複数のシステムに対し原文・訳文のみの情報を使って原文中の情報がどれくらい訳文に反映されているかを公平に判断することが困難であったためこのような feature set になっている。

4. 統計的後編集と自動評価手法との融合

前述の PLS 回帰分析モデルを使って RBMT 結果の翻訳品質の自動評価を行い、その結果によって SMT ベースの自動後編集を行うかどうかを決定する。

品質予測値は Adequacy 予測値と Fluency 予測値との調和平均を用い、それが閾値 ρ 以上であれば RBMT 結果をそのまま出力とする。ρ よりも小さい値の場合はドメイン適応させるべき訳文と考え、SMT ベースの自動後編集を適用する。

4.1. 実験

NTCIR7 日英特許翻訳タスクのテストセット(899 文)をシステムへの入力とする。

図 4 は閾値を品質予測値の最小値 1.34 から最大値 5.21 まで 0.01 ずつ増加させた場合の NIST 値(実線)と実際に自動後編集(APE)が適用された文数(破線)の変化を表わす。

図 4 より、閾値が 2~3.5 付近では NIST 値の変化が大きく、閾値が 4 を超える辺りから NIST 値の変化は緩やかになっていることが分かる。

5. 考察と今後の課題

品質予測値が 4 以上であれば RBMT の翻訳結果を自動後編集する必要がない可能性が高い。表 3 はその一例である。SRC は入力原文、REF は参照訳、RAW は RBMT 結果、APE は自動後編集結果を表わす。

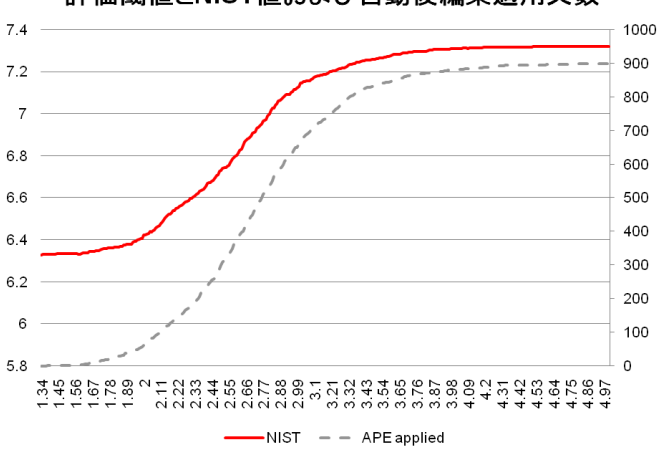

図 4：閾値と NIST 値の変化

原文・訳文	RBMT の品質予測値	品質予測値の差分
SRC：なお、トップ 3 t は、仮想の存在である。 REF：the top 3t has an imaginary existence . RAW：3 t of tops are existence of imagination . APE：3 t of the frames are presence or absence of virtual .	4.80	-1.83
SRC：バリ取り作業にロボットを利用することは従来より公知の技術である。 REF：the use of a robot for deburring work is a known prior art . RAW：it is technology better known than before to use a robot for barricade picking work . APE：it is better known than before to use a robot for deburring work .	2.52	+0.13

表 3：例

上記において第 1 の例の場合、RBMT 結果の品質予測値 4.8 で自動後編集後には品質予測値は 1.83 減少してしまう。実際の自動後編集後の訳文を見てもその悪化が確認できる。

一方、第 2 の例の場合、RBMT 結果の品質予測値は 2.52 であり、自動後編集後には品質予測値が 0.13 上昇している。訳文を見ると、「バリ取り作業」が"barricade picking work"から"deburring work"に変更されており効果が確認できる。

本研究の手法によれば、SMT による自動後編集を RBMT 結果に対して選択的に適用することができるため、湧き出し語などの SMT 固有の問題の低減が期待できる。ただしこの場合、最適な閾値をどのように設定するかが問題となる。 [Specia, et al., 2009]では予測の信頼度が所定

の値を満たすように 2 分探索によって閾値を決定する手法が提案されており、本手法でも最適な閾値導出手法を適用する必要がある。

また、今回 PLS 回帰分析の際採用した feature set には訳文側の情報しか用いていない。Adequacy の予測精度を向上させるためには単語対応情報などを使って、原文・訳文の双方の情報を feature set に入れる必要がある。

参考文献

Blatz, J., Fitzgerald, E., Foster, G., Gandrabur, S., Goutte, C., Kulesza, A., et al. (2003). Confidence estimation for machine translation. *Technical report, Johns Hopkins Univ.*

Koehn, P., Federico, M., Cowan, B., Zens, R., Dyer, C., Bojar, O., et al. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. *Proceedings of the ACL 2007 Demo and Poster Sessions* , pages 177-180.

Koehn, P., Och, F. J., & Marcu, D. (2003). Statistical Phrase-based Translation. *Proceedings of NAACL HLT 2003* , pages 127-133.

Lagarda, A.-L., Alabau, V., Casacuberta, F., Silva, R., & Diaz-de-Liano, E. (June 2009). Statistical Post-Editing of a Rule-Based Machine Translation System. *Proceedings of NAACL HLT 2009, ACL* , pages 217-220.

NIST: Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurence Statistics. (2002). 参照先: ftp://jaguar.ncsl.nist.gov/mt/mt2001/mt-eval-02-jan-public.pdf

Simard, M., Goutte, C., & Isabelle, P. (April 2007). Statistical Phrase-based Post-editing. *Proceedings of NAACL HLT 2007, ACL* , pages 508-515.

Simard, M., Ueffing, N., Isabelle, P., & Kuhn, R. (June 2007). Rule-based Translation With Statistical Phrase-based Post-editing. *Proceedings of the second Workshop on Statistical Machine Translation, ACL* , pages 203-206.

Specia, L., Cancedda, N., Turchi, M., & Cristianini, N. (May 2009). Estimating the Sentence-Level Quality of Machine Translation Systems. *Proceedings of the 13th Annual Conference of the EAMT* , pages 28-35.

Specia, L., Saunders, C., Turchi, M., Wang, Z., & Shawe-Taylor, J. (2009). Improving the Confidence of Machine Translation Quality Estimates. *MT Summit XII* .

The 7th NTCIR Workshop. (2007/2008). (NII) 参照先: http://research.nii.ac.jp/ntcir/ntcir-ws7/ws-en.html

江原暉将. (2006). 規則方式機械翻訳と統計的後編集を組み合わせた特許文の日英機械翻訳. 平成 17 年度 AAMT/Japio 特許翻訳研究会報告書, pages 40-44.

江原暉将. (2008). 句レベルの統計的後編集と翻訳精度の評価. 平成 19 年度 AAMT/Japio 特許翻訳研究会報告書, pages 2-11.

村上仁一, 徳久雅人. (2010). ルールベース翻訳と統計翻訳を結合した特許翻訳. 第 1 回特許情報シンポジウム,AAMT/Japio 特許翻訳研究会, pages 46-53.

Wold, S., Ruhe, A., Wold, H., & Dunn, W. J. (1984). The covariance problem in linear regression. the partial least squares(pls) approach to generalized inverses. *SIAM Journal on Scientific Computing* , pages 5:735-743.

Zhu, X., Yang, M., Wang, L., Wang, J., & Li, S. (2009). A Quantitative Analysis of Linguistic Factors in Human Translation Evaluation. *2nd International Symposium on Knowledge Acquisition and Modeling* , pages 410-413.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Tree. *Proceedings of International Conference on New Methods in Language Processing* .

Grinberg, D., Lafferty, J., & Sleator, D. (1999). A robust parsing algorithm for link grammars. *Proceedings of the 4th International Workshop on Parsing Technologies* .

Feature ID	Feature Description
1	訳文の 1-gram 頻度の四分位範囲/1-gram 総数
2	(訳文の 1-gram 頻度の第 3 四分位-第 2 四分位)/(訳文の 1-gram 頻度の第 2 四分位-第 1 四分位)
3	訳文 1-gram 頻度の 50 パーセンタイル/1-gram 総頻度
4	訳文の 2-gram 頻度の四分位範囲/2-gram 総数
5	(訳文の 2-gram 頻度の第 3 四分位-第 2 四分位)/(訳文の 2-gram 頻度の第 2 四分位-第 1 四分位)
6	訳文 2-gram 頻度の 50 パーセンタイル/2-gram 総頻度
7	訳文の 3-gram 頻度の四分位範囲/3-gram 総数
8	(訳文の 3-gram 頻度の第 3 四分位-第 2 四分位)/(訳文の 3-gram 頻度の第 2 四分位-第 1 四分位)
9	訳文 3-gram 頻度の 50 パーセンタイル/3-gram 総頻度
10	訳文中の非隣接単語間の MI(mutual information)の合計値/単語ペア数
11	訳文中の非隣接単語間の Dice 係数の合計値/単語ペア数
12	訳文の 2-gram 言語モデル確率/2-gram 数
13	訳文の 3-gram 言語モデル確率/3-gram 数
14	訳文の backward 2-gram 言語モデル確率/2-gram 数
15	訳文の backward 3-gram 言語モデル確率/3-gram 数
16	訳文の品詞 2-gram 言語モデル確率/品詞 2-gram 数
17	訳文の品詞 3-gram 言語モデル確率/3-gram 数
18	訳文の品詞 backward 2-gram 言語モデル確率/品詞 2-gram 数
19	訳文の品詞 backward 3-gram 言語モデル確率/品詞 3-gram 数
20	名詞数/単語数
21	動詞数/単語数
22	形容詞数/単語数
23	副詞数/単語数
24	Link Grammar Parser の有効 linkage 数
25	Link Grammar Parser の無効 linkage 数
26	Link Grammar Parser の null count/単語数
27	Link Grammar Parser の名詞句数
28	Link Grammar Parser の動詞句数

表 4：Feature set