

統計的後編集の効果的利用に関する検証 Examination on Effective Application of Statistical Automatic Post-Editing

鈴木 博和†
Hirokazu Suzuki

1. まえがき

韓国語と日本語のように言語的に非常に類似した言語であれば、統計ベース翻訳(SMT)も十分な性能が得られることが予想されるが、実際には十分な分量の対訳コーパスを得ることが難しい。

またそのような言語間の翻訳において、言語的類似性を活かし、原言語形態素解析結果の各単語を直接訳語に置き換えるような、単語翻訳をベースにした翻訳手法も考えられる。しかし、韓日翻訳における韓国語のように表記が同一の多品詞語の場合や、一般的には一見出しに対して複数の訳語候補が存在することが多いために、適切な訳語を付与することが難しいという問題もある。

本論文では、小規模な韓日対訳コーパスが存在するときに、(A) 直接韓日 SMT を構築した場合と、(B) 単語ベースの韓日翻訳システムを一旦構築し、その出力に対し自動後編集(Automatic Post-Editing, APE)を行った場合とで、どのような性能差があるかどうかを検証する。

自動後編集は「機械翻訳で頻出する誤りを別の表現に自動的に修正する」というものであるが、この処理は SMT と相性が良く、SMT を用いて自動的に後編集する手法が多く提案されている [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。

本研究では自動後編集モジュールに Moses を用い、機械翻訳結果（韓国語原文に対する単語ベースの韓日翻訳システムの出力）を原文側に、それに対応する参照訳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。

しかし、上記のような自動後編集を適用すると却って翻訳品質が悪化する場合がある。自動後編集の適用前後で翻訳品質の予測が出来れば、どちらか一方を選択することにより、そのような悪影響を抑えることが可能になる。本研究では、Confidence Estimation [Specia, et al., May 2009] [Specia, et al., 2009] [Suzuki, 2011] に着目し、そこで用いられている PLS 回帰分析手法による、参照訳を必要としない翻訳品質予測モデルを導入する。これにより上記 (B) のシステムにおいて、自動後編集の前後で品質変化を予測し、自動後編集の適用可否を判定した場合、翻訳性能がどの程度変化するかを検証を行う。

2. 単語ベース韓日翻訳の構築

まず、始めに上記システム (B) で使用する単語ベース韓日翻訳システムについて説明する。

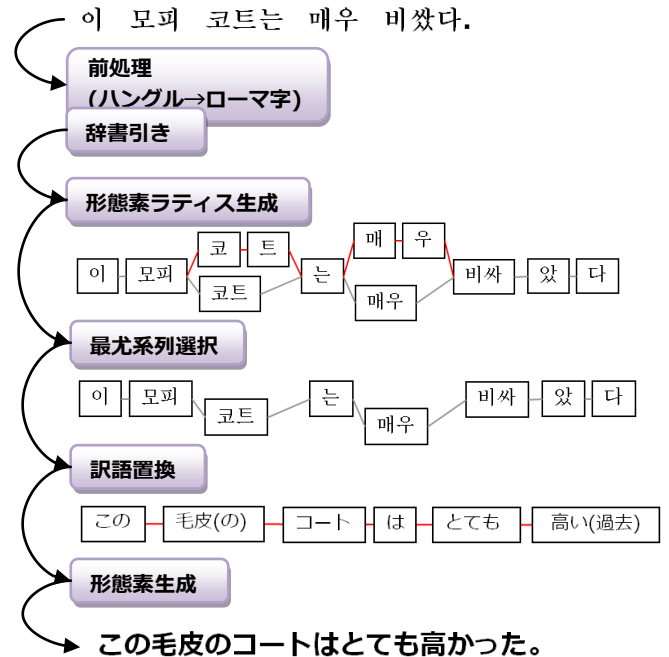


図1 単語ベース韓日翻訳システム概要図

このシステムで最も重要なモジュールは形態素解析モジュールである。今回は開発コストが掛かる規則ベース手法ではなく、短期間で開発が可能な統計的手法(条件付き確率場:Conditional Random Field)[松本, 他, 2004]を用いて韓国語形態素解析モジュールを実現した。図1に示すように前処理のあと辞書引きを行い、そこに付与されている統計的手法で推定したコストに基づいて最もコストの小さい形態素系列(最尤系列)を探索する。その得られた最尤系列中の各形態素を、対応する訳語に置換していくことにより、訳文を生成する。日本語の形態素生成は既に開発済みの英日翻訳用の日本語形態素生成モジュールを再利用する。

この方式の問題点は、訳し分け規則が存在しないために、複数の訳語候補が存在するときに訳語選択ができない点にある。しかし、訳し分け規則がないということは常に同じ翻訳誤りを行うということになり、この性質は APE と非常に相性が良いといえる。

3. SMT および APE 用データセット

次に、システム (A) の SMT およびシステム (B) の APE モジュールを構築する。使用する句ベース SMT には Moses [Koehn, et al., 2007]を用いた。

† (株) 東芝 研究開発センター 知識メディアラボ
トリー

中央日報¹および東亜日報²から収集にした韓日対訳文 177,572 文の内、158,179 文を訓練用データ、17,500 文を開発用データ、1,893 文をテスト用データ（テストセット 1）として用いた。

またオープン評価用データとして、MBC 対訳ニュース記事³500 文（テストセット 2）を用いた。

システム（B）の APE モジュールの訓練・評価用には、上記のデータの原文側を前節で述べた単語ベース韓日翻訳システムの出力に置き換えたデータを用いた。

4. APE の効果

表 1 は各データセットに対するシステム（A）の結果、およびシステム（B）の APE 適用前後の結果である。

テストセット 1		
	NIST	BLEU
System A	9.8439	0.5728
System B (Before APE)	8.9346	0.4460
System B (After APE)	11.7149	0.6829
テストセット 2		
	NIST	BLEU
System A	6.1661	0.3335
System B (Before APE)	9.0260	0.5241
System B (After APE)	8.9944	0.5374

表 1 システム（A）とシステム（B）の翻訳結果

まず、テストセット 1 では、単語ベース翻訳結果に対して APE を適用したものが最も性能が良かった。システム（A）はシステム（B）に於ける単語ベース韓日翻訳単体よりも性能が良い。したがって、十分な分量のコーパスを準備できないとしても、訓練時と翻訳時のドメインが合致していれば、コストを抑えて機械翻訳システムを構築可能なことがわかる。しかし、多少コストを掛けても単語ベース韓日翻訳を構築し、それに対して APE を適用したほうが、同じ分量のコーパスでもはるかに効率的に高精度な翻訳システムを実現できることがわかる。

一方、テストセット 2 ではシステム（A）の結果が最も悪い。システム（B）の単語ベース翻訳単体は安定した性能を持っており、APE を適用しても大きな性能の悪化はなかった（NIST 値は微減するが BLEU 値は向上）。

このことから、「単語ベース翻訳システム+APE」という構成は同じ分量のコーパスを SMT よりも効率的に利用することができ、ドメインが適合する場合に極めて高い性能を有することが分る。これはドメイン適応の観点でも非常に有利な特徴といえる。

表 2 はテストセット 2 に存在する参照訳と APE 適用前後の訳文の例である。

	訳文
参 照	江南を出発して 1 時間余りで代表チームを乗せた車が到着し、祭りの幕が上がります。
訳	

APE 適用前	江南を出発するの 1 時間余りに代表チームをやいた車両が到着して祭りの幕が上がります。
APE 適用後	江南（カンナム）を出発するの、1 時間後に代表チームを乗せた車両が到着し、祭りの幕が上がります。

表 2 テストセット 2 における APE 適用の例

訓練に用いたコーパスの特徴に依存した後処理として「江南→江南（カンナム）」のように読みのような補助情報が付与されるケースが多い。このような後編集処理が NIST 値に影響していると思われる。

次に、APE が効果的に機能しているのかを検証するために人手評価を行った。評価はテストセット 1 および 2 から抽出した各 50 文に対し、APE により訳文が改善したか否かを判断することによって行った。その結果、テストセット 1 の場合は 50 文中 47 文改善（94%）であり、テストセット 2 の場合は 50 文中 29 文改善（58%）であった。テストセット 2 に関しては翻訳結果に対して APE を適用するかどうかは慎重に判断する必要がある。

そこで、APE の適用前後で翻訳品質の変化を予測し、その変化量に応じて、APE 適用前の訳文を出力するか、APE 適用後の訳文を出力するかを判断する識別器を用意する。そして、そのような訳出の制御によりどのように性能が変化するかの検証を行う。

3. PLS 回帰分析を用いた翻訳品質予測

上記の制御を行うために翻訳品質を予測する場合、BLEU や NIST のように参照訳を用いることができない。したがって参照訳を必要としない品質評価手法が必要となる。ここでは、[Suzuki, 2011]の手法と同様に PLS 回帰分析によって翻訳品質予測モデルを構築する。

3.1 PLS 回帰分析

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。

3.2 feature set

まず、入力変数として用いる feature set を表 3 のように設計した。

Feature
2-gram 言語モデル確率
3-gram 言語モデル確率
Backward 2-gram 言語モデル確率

¹ <http://japanese.joins.com//>

² <http://www.donga.com/>

³ MBC の韓国語ニュース 2000-2009 年, MBC 放送報道局 (著), ISBN: 978-4757418646

Backward 3-gram 言語モデル確率
1-gram IQR ¹ / sample range
2-gram IQR / sample range
3-gram IQR / sample range
(原文単語数-訳文単語数) ² /原文単語数*訳文単語数
訳文内容語数/訳文形態素数
訳文機能語数/訳文形態素数
(内容語数-機能語数) ² /内容語数*機能語数
内容語の相互情報量
内容語の Dice 係数
特定品詞列 ² の 3-gram log 確率
頻度 0 の 2-gram 数/2-gram 数
頻度 0 の 3-gram 数/2-gram 数
(原文形態素数-訳文形態素数) ² /原文形態素数*訳文形態素数
(原文内容語数-訳文内容語数) ² /原文内容語数*訳文内容語数
(原文機能語数-訳文機能語数) ² /原文機能語数*訳文機能語数
(原文名詞数-訳文名詞数) ² /原文名詞数*訳文名詞数
(原文動詞数-訳文動詞数) ² /原文動詞数*訳文動詞数

表 3 feature set

韓国語形態素解析は Mach³を用い、日本語形態素解析には Mecab を用いた。

3.3 学習データと目的変数

東亜日報の韓国語記事の機械翻訳結果に対し、5 人の評価者が Adequacy/Fluency5 段階評価を行ったもの各 150 文のうち、PLS 回帰分析モデルの構築に使用する学習データは 100 文、残りの 50 文を評価用に用いる。

目的変数には、5 人の Adequacy 評価の中央値、Fluency 評価の中央値の調和平均値を用いる。また一般に、人手での後編集は機械訳文に対する、挿入・削除・置換・シフトにより行われるため、目的変数に TER の値も用いることにした。Adequacy・Fluency 実測中央値の調和平均値と TER 実測値との Spearman 順位相関係数は-0.466 であった。これは中程度の逆相関関係であり TER と Adequacy・Fluency 評価の間にはある程度関係性が認められることがわかる。

3.4 翻訳品質予測モデルの評価

PLS 回帰分析による翻訳品質予測モデルの評価は表 4 のようになった。

Adequacy/Fluency 調和平均値予測	
RMSPE	1.014
Spearman 順位相関係数	0.454
TER 予測	
RMSEP	1.027
Spearman 順位相関係数	0.392

表 4 翻訳品質予測モデルの評価

¹ IQR : Inter Quartile Range

² (名詞,助詞,名詞), (名詞,助詞,動詞), (動詞,助詞,動詞)

³ <http://cs.sungshin.ac.kr/~shim/demo/mach.html>

RMSEP(Root Mean Squared Prediction Error)は以下の式で表される (y は実測値、y' は予測値を表わす) :

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

3.5 APE 適用評価基準

次に、APE 結果を採用するか否かを判断するための評価基準を定義する。ここでは、Adequacy・Fluency の調和平均予測値を pred(Adequacy,Fluency)、TER 予測値を pred(TER)とし、以下で評価式を定義する：

$$Quality = \alpha \cdot \Delta pred(Adequacy, Fluency) + \beta \cdot \Delta pred(TER)$$

$\Delta pred$ は APE 適用前後での予測値の差分を表す。

上記評価結果を見ると Adequacy・Fluency 調和平均値予測のほうが TER 予測よりも若干性能がよい。したがって、評価式の α 、 β はそれぞれ経験的に $\alpha=0.7$ 、 $\beta=-0.3$ とした。

APE の採用可否は閾値 θ を用いて、以下の式で決定する：

Quality $\geq \theta$: APE 後の結果を出力

Quality $< \theta$: APE 前の結果を出力

3.6 閾値による性能変化

上記閾値 θ を変化させ、APE の適用を制御した場合に翻訳性能がどのように変化するかを調べる。

	APE 適用文数	APE 非適用文数	文数
APE を適用すべき文	(1)	(2)	29
APE を適用すべきでない文	(3)	(4)	21
計			50

閾値を変化させたときに、以下のように Precision と Recall を以下のように計算する：

$$Precision = ((1) + (4)) / 50$$

$$Recall = (1) / 29$$

閾値による変化は図 2 の様になった。閾値が最小値のときは、すべての訳文に対して APE が適用される。この時の F 値は 0.734 となっている。閾値を最小値から最大値まで変化させると、APE が適用される文数は減少するため、Recall は減少し続けるが、APE を適用すべきでない文に対して APE を適用しない文数は増えるため Precision は若干上昇する。閾値が-0.012 になったときに F 値は最高となる(0.772)。その後 F 値は減少し続ける。

5. まとめと今後の課題

韓国語と日本語のように言語的に非常に類似した言語対でかつ十分な分量の対訳コーパスが利用できない場合は、初期開発コストはかかるが、一旦単語ベース翻訳システムを構築し、その結果に対して APE を実行する翻訳システムのほうが、対訳コーパスから直接 SMT を実現するよりもはるかに効率の良い翻訳システムを構築可能であることを示した。また、PLS 回帰分析により APE 適用前後での翻訳品質の変化を予測し、その変化量に応じて

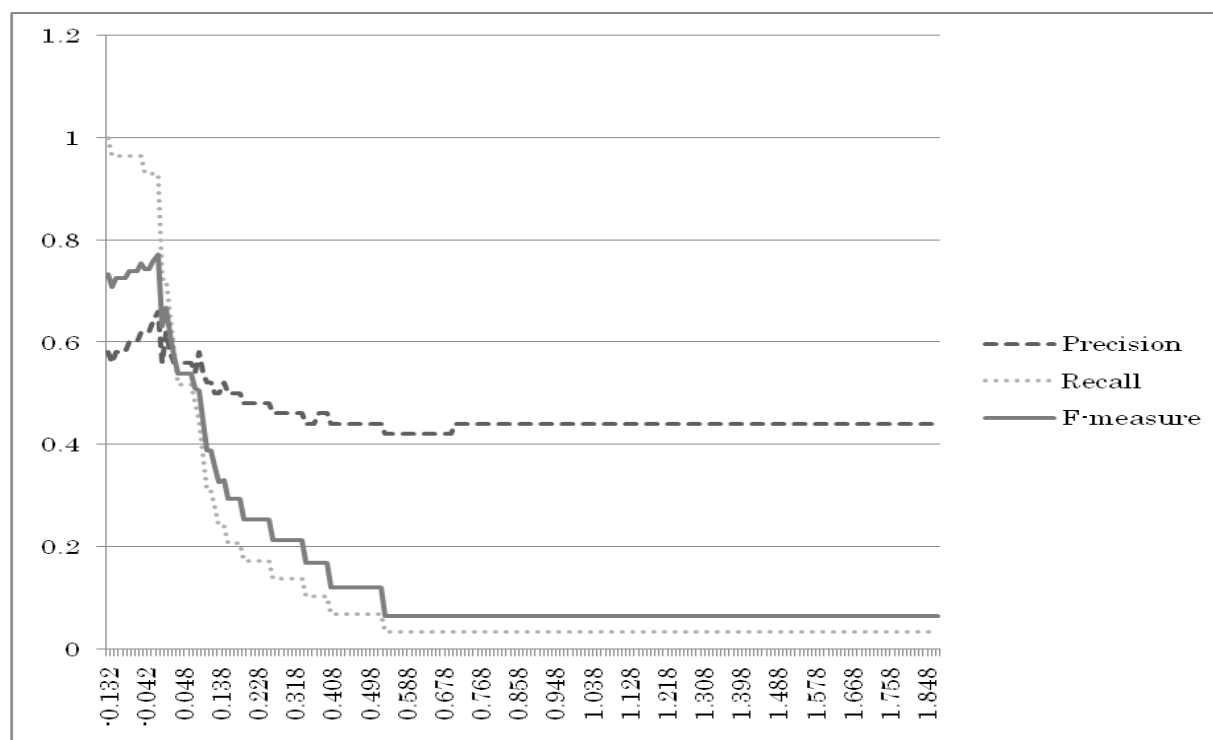


図2 翻訳性能の変化

APE 適用を制御すれば、ある程度の性能向上が図れることがわかった。

今後の課題として、

- 湧き出し語などの抑制
- より優れた翻訳品質予測モデルの構築
- 言語的性質を考慮した feature set 設計
- APE モジュールを単語ベース翻訳システム「内部」に導入し、句レベルでの置換と単語レベルでの置換を行うハイブリッド化などが考えられる。

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-Occurrence 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*.

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*.

Suzuki H. (2011). Automatic Post-Editing based on SMT and its selective application by Sentence-Level Automatic Quality Evaluation. *MT Summit XIII*.

工藤 拓, 山本 薫, 松本 裕治. (2004). Conditional Random Fields を用いた日本語形態素解析. 情報処理学会自然言語処理研究会 SIGNAL-161.