Levenshtein Distanceを用いたPattern Based SMT

力久剛士† 村上仁一‡ 徳久雅人‡
鳥取大学 工学部 知能情報工学科†
鳥取大学大学院 工学研究科 情報エレクトロニクス専攻‡
{s072063†, murakami‡, tokuhisa‡} @ ike.tottori-u.ac.jp

1 はじめに

パターン翻訳 [1] は，1960 年代に提案された翻訳方法である．人手により作成した，対訳句辞書と対訳文パターン辞書を用いて翻訳を行う．この翻訳方式は入力文が適切な対訳文パターンに適合した場合，翻訳精度の高い出力文が得られる．しかし，対訳句辞書と対訳文パターン辞書の作成は人手で行うため，開発にコストがかかる．

この問題を解決するために江木らは，GIZA++[2] を利用した Pattern Based SMT[3] を提案した．この手法は対訳句辞書と対訳文パターン辞書を自動的に作成する．よって，開発コストを削減することができる．しかし，対訳文パターンに適合しても，人手評価が低い出力文がある．この問題の原因の一つは，不適切な対訳文パターンの選択である．

そこで本研究では，日英 Pattern Based SMT において，対訳文パターンの日本語原文と入力文とのレーベンシュタイン距離 [4](以下 LsD) を求める．この距離を利用して，対訳文パターンの日本語原文と入力文が類似した対訳文パターンの選択を行い，翻訳精度の向上を目指す．

2 Pattern Based SMT[3]

2.1 概要

Pattern Based SMT は，原言語と目的言語の対訳句から成る“ 対訳フレーズ辞書 ”と，対訳文に対して，任意の句を変数化した“ 句に基づく対訳文パターン辞書 ”を統計的手法を用いて自動作成し，翻訳を行う．辞書の自動作成により，開発コストが削減できる．以下に日英翻訳を想定した Pattern Based SMT の手順を示す．

2.2 対訳単語辞書の作成

対訳文と GIZA++を用いて，対訳単語に単語翻訳確率を付与した，“ 対訳単語辞書 ”を作成する．

2.3 単語に基づく対訳文パターンの作成

対訳文と対訳単語の照合を行う．対訳単語と適合した対訳文の単語を変数化して単語に基づく対訳文パターンを作成する．

2.4 対訳フレーズ辞書の作成

対訳文と単語に基づく対訳文パターンを照合し，変数部に対応する対訳フレーズを抽出する．抽出した対訳フレーズに対訳単語辞書を用いて，対訳フレーズ対数確率を付与した，“ 対訳フレーズ辞書 ”を作成する．

2.5 句に基づく対訳文パターン辞書の作成

対訳文と対訳フレーズの照合を行う．対訳フレーズが適合した対訳文のフレーズを変数化して句に基づく対訳文パターンを作成する．その後，句に基づく対訳文パターンの変数化していない部分 (以下字面) と，対訳単語辞書を用いて，対訳文パターン対数確率を付与した，“ 句に基づく対訳文パターン辞書 ”を作成する．句に基づく文パターン辞書作成の例を図 1 に示す．

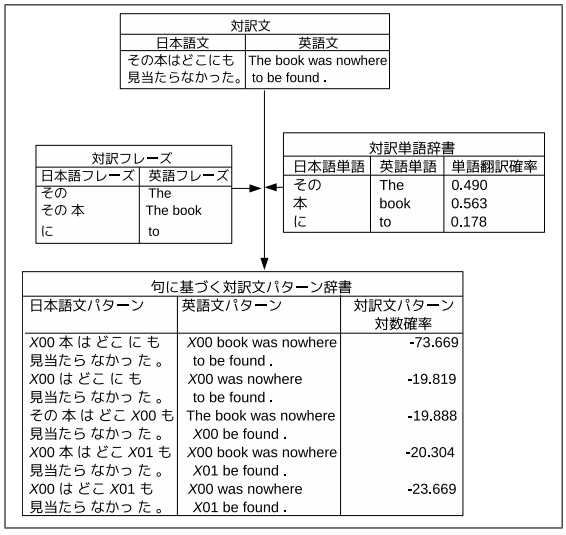

図 1 句に基づく対訳文パターン辞書の作成

2.6 出力文の生成

句に基づく対訳文パターン辞書と対訳フレーズ辞書を利用して出力候補文を生成する．次に，作成した出力候補文から出力文を選択する．出力文の生成方法を以下に示す．

2.6.1 句に基づく日本語文パターンの選択

入力文と，句に基づく日本語文パターンの字面を照合する．字面が多く一致した日本語文パターンを持つ対訳文パターンを優先して選択する．

2.6.2 出力候補文の作成

選択した対訳文パターンにおいて，英語文パターンの変数部に対訳フレーズを用いて英語フレーズを挿入し, 出力候補文を生成する.

2.6.3 出力文の選択

対訳文パターン対数確率と出力候補文の作成に用いた対訳フレーズ対数確率と言語モデル (tri-gram) を用いて，翻訳文の翻訳対数確率を計算する．出力候補文の翻訳対数確率が最も高い出力候補文を“ 出力文 ”として出力する．出力文生成の例を図 2 に示す.

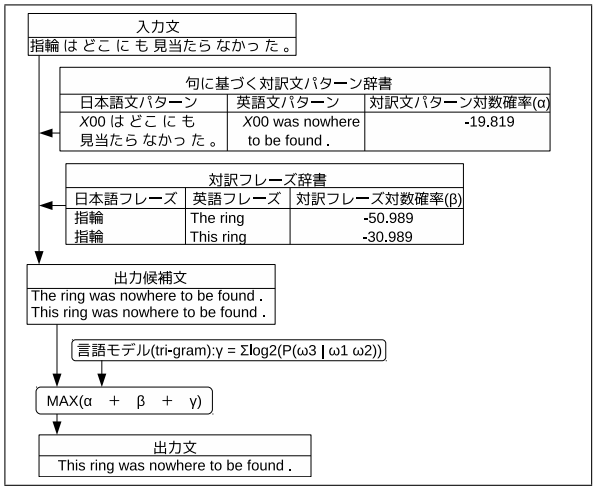

図 2 英語出力文生成の例

2.7 Pattern Based SMT の問題点

Pattern Based SMT の出力文には，人手評価の低い出力文がある．誤り解析を行った結果，不適切な対訳文パターンの選択が一つの原因であった．人手評価の低い入力文に対して，適切な対訳文パターンを使用した結果，人手評価の高い出力文が得られた．不適切な対訳文パターンを使用した翻訳の例を表 1 に，適切な対訳文パターンを使用した翻訳の例を表 2 に示す.

表 1 不適切な文パターンを使用した翻訳の例

入力文	指輪 は どこ に も 見当たら なかっ た 。
参照文	The ring was nowhere to be found .
日本語文パターン	$X00$ は $X01$ に $X02$ なかっ た 。
英語文パターン	It took $X00$ forever to $X02$ $X01$.
日本語文パターンの原文	彼 は なかなか 仕事 に 取りかから なかっ た 。
英語文パターンの原文	It took him forever to get down to work .
出力文	It took the ring forever to be found anywhere .

表 1 の例では，日本語文パターンの原文は入力文とは類似しておらず，生成した出力文に対する人手評価は低い.

表 2 の例では，表 1 と異なり，日本語文パターンの原文は入力文と類似している．よって表 2 は，生成した出力文に対する人手評価は表 1 の例と比較して高い.

表 2 適切な文パターンを使用した翻訳の例

入力文	指輪 は どこ に も 見当たら なかっ た 。
参照文	The ring was nowhere to be found .
日本語文パターン	$X00$ は どこ に も 見当たら なかっ た 。
英語文パターン	$X00$ was nowhere to be found .
日本語文パターンの原文	その 本 は どこ に も 見当たら なかっ た 。
英語文パターンの原文	The book was nowhere to be found .
出力文	This ring was nowhere to be found .

3 提案手法

3.1 概要

Pattern Based SMT において人手評価が低い原因の一つは，入力文に対して不適切な対訳文パターンを選択することである．人手評価が低い入力文に対して対訳文パターンの日本語原文と入力文が類似した対訳文パターンを与えた結果，人手評価が高い出力候補文が選択された.

よって，対訳文パターンの日本語原文と入力文の類似した対訳文パターンを選択するために，入力文と日本語文パターンの原文との類似度を利用する．類似度は LsD を用いて求め，出力候補文を選択する際に，対訳文パターン対数確率 (α) の代わりに使用する.

3.2 レーベンシュタイン距離 (Levenshtein Distance)[4]

二つの文字列がどの程度異なっているかを示す距離として,LsD がある.LsD とは, 一つの文字列をもう一つの文字列にするための編集回数である. 編集は, 挿入, 削除, 置換の三つがある．本研究では， 削除編集 D，置換編集 S と単語数 N を用いて入力文と日本語文パターンの原文との類似度を式 (1) の計算式により求める.

$$\text{類似度} = \frac{N - D - S}{N} \tag{1}$$

3.3 実験手順

本研究では，日英翻訳を行う．提案手法は出力文の生成において，入力文と日本語文パターンの原文との類似度を求める．提案手法による出力候補文の生成の例を図 3 に示す.

3.4 LsD による類似度の付与

入力文と対訳文パターンの日本語原文との LsD を求める．次に，LsD より“ 類似度 ”を求める．最後に，対訳文パターン対数確率と出力候補文の生成に用いた対訳フレーズ対数確率と言語モデルと“ 類似度 ”を用い

て，出力候補文の翻訳対数確率を計算する．出力候補文の翻訳対数確率が最も高い翻訳文を“ 出力文 ”として出力する．

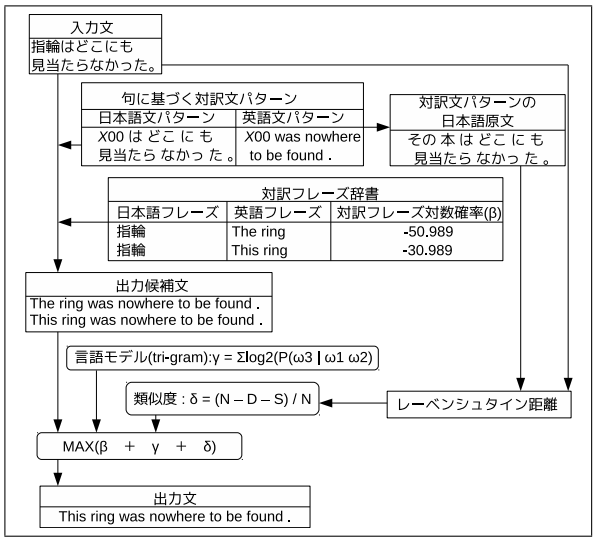

図 3 提案手法における出力文の生成の例

4 実験

4.1 実験データ

対訳文および翻訳実験に用いるテスト文として電子辞書から抽出した単文データを用いる [5]. なお，単文データは，日本語文が単文であるが，英語文は単文とは限らず，重文・複文が含まれる．コーパスの内訳を表 3 に示す.

表 3 実験データ

対訳文	100,000 文対
入力文	100 文

4.2 実験結果

提案手法を用いて，翻訳実験を行った．実験の結果，入力文 100 文中，出力文 81 文を得た.

5 評価

提案手法とベースラインの対比較評価を行った．ベースラインには，2 節の Pattern Based SMT を用いる．結果を表 4 に示す.

表 4 人手による評価

提案手法○	ベースライン○	差なし	同一出力
8	7	18	48

提案手法○の例を表 5，ベースライン○の例を表 6 に示す.

表 5 提案手法○の例

入力文		彼 は 仕事 で 京都 へ 行っ た 。
参照文		He went to Kyoto on business .
提案手法	日 本 語 文 パターン	彼 は 仕事 で $X00$ 行っ た 。
	英語文パターン	He went $X00$ on business .
	対訳文パターンの日本語原文	彼 は 仕事 で ロンドン へ 行っ た 。
	対訳文パターンの英語原文	He went to London on business .
	出力文	He went to Kyoto on business
ベースライン	日 本 語 文 パターン	彼 は $X00$ で $X01$ 行っ た 。
	英語文パターン	He went $X01$ by $X00$.
	対訳文パターンの日本語原文	彼 は シベリア 経由 で ヨーロッパ へ 行っ た 。
	対訳文パターンの英語原文	He went to Europe by via Siberia .
	出力文	He went to Kyoto by work

表 6 ベースライン○の例

入力文		彼女 の 心臓 は どきどき し て い た 。
参照文		Her heart thudded .
提案手法	日 本 語 文 パターン	彼女 の $X00$ $X01$ し て い た 。
	英語文パターン	$X00$ $X01$
	対訳文パターンの日本語原文	彼女 の 目 は 疲労 で ピクピク し て い た 。
	対訳文パターンの英語原文	Her eyes twitched with fatigue .
	出力文	The heart beats
ベースライン	日 本 語 文 パターン	彼女 の $X00$ は $X01$ し て い た 。
	英語文パターン	Her $X00$ was $X01$.
	対訳文パターンの日本語原文	彼女 の 声 は はつらつ と し て い た 。
	対訳文パターンの英語原文	Her voice was fresh as springtime .
	出力文	Her heart was pounding .

表 5 の例では，ベースラインで選択された対訳文パターンの日本語原文と比べて，提案手法の方が入力文と類似した対訳文パターンが選択がされた．しかし，表 6 の例では，入力文と類似した日本語文パターンが選択されなかった.

6 考察

6.1 提案手法の有効性

人手評価の結果，提案手法の有効性が確認できなかった．提案手法○の数とベースライン○の数がほぼ同等であったため，入力文と対訳文パターンの日本語原文との類似度の効果と，対訳文パターン対数確率の効果が同様であると考えられる．

よって，提案手法は，対訳文パターンの選択において，対訳単語辞書を用いない新たな対訳文パターンの選択手法であると言える．

6.2 字面の対応が取れていない対訳文パターンの使用

対訳文パターンの選択の際に，類似度を用いることにより，対訳文パターンの日本語原文と入力分が類似した対訳文パターンの選択をすることができた．しかし，評価を行った結果，ベースラインと比較して，精度は向上しなかった．そこで，ベースライン○と評価された 7 文に対して，誤り解析を行った．誤り解析の結果，選択した対訳文パターンにおいて，対訳文パターンの日本語原文と入力文は類似していたが，日本語文パターンの字面と英語文パターンの字面の対応が取れていない対訳文パターンがあることがわかった．字面の対応の取れていない対訳文パターンの例を表 7 に示す．

表 7 字面の対応が取られていない日英文パターンの例

入力文		彼女 は 思いがけない 質問 に まごつい た よう だっ た 。
参照文		She seemed to be embarrassed at the unexpected question .
提案手法	日本語文パターン	*X*00 *X*01 *X*02 に *X*03 ようだっ た 。
	英語文パターン	I came across a very helpful person when I *X*03 *X*01 *X*02 *X*00 .
	対訳文パターンの日本語原文	地獄 で 仏 に あっ た ようだっ た 。
	対訳文パターンの英語原文	I came across a very helpful person when I was in dire trouble .
	出力文	I came across a very helpful person when I was upset at the unexpected question to her .
ベースライン	日本語文パターン	*X*00 *X*01 *X*02 に *X*03 た *X*04 だっ た 。
	英語文パターン	*X*00 was *X*03 *X*04 *X*01 *X*02 .
	対訳文パターンの日本語原文	それ は 浮き彫り に し た キューピッド の 像だっ た 。
	対訳文パターンの英語原文	It was a figure of Cupid in relief .
	出力文	She was perplexed by such a stroke of questions .

表 7 において，日本語文パターンの字面と英語文パターンの字面の数が著しく異なる対訳文パターンを，対訳文パターンの選択の際に除外することにより，人手による対比較評価が改善すると考える．また，本研究では，対訳フレーズ対数確率と対訳文パターン対数確率，類似度が等しくなるように重みを設定した．この重みを最適化することで人手評価が向上すると考える．

7 まとめ

本研究では，Pattern Based SMT において，不適切な対訳文パターンの選択を抑制するため，対訳文パターンを選択する際に，対訳文パターン対数確率の代わりとして，入力文と対訳文パターンの日本語原文との LsD を使用することを提案した．人手による評価をした結果，出力文 81 文中，提案手法○ 8 文，ベースライン○ 7 文であり，提案手法の有効性は確認できなかった．また，誤り解析を行った結果，日本語文パターンの字面と英語文パターンの字面において，対応が取れていない対訳文パターンがあることがわかった．よって，日本語文パターンの字面と英語文パターンの字面の数が著しく異なる対訳文パターンを，除外することにより，翻訳精度が向上すると考える．また，今後対訳フレーズ対数確率と対訳文パターン対数確率と類似度の重みを最適化し，実験を行う．

参考文献

[1] 渡辺日出雄, 武田浩一, “ パターンベース翻訳システム PalmTree ”, 情報処理学会第 55 回全国大会講演論文集, pp80-81, 1997.

[2] Franz Josef Och, Hermann Ney, “ A Systematic Comparison of Various Statistical Alignment Models ”, Computational Linguistics, 29(1), pp.299-314, 1996.

[3] 江木孝史, 村上仁一, 徳久雅人, “ 句に基づく対訳文パターンの自動作成と統計的手法を用いた英日パターン翻訳 ”, 自然言語処理学会第 20 回年次大会予稿集, pp.951-954, 2014.

[4] Vladimir Iosifovich Levenshtein, “ Binary codes capable of correcting deletions, insertions, and reversals ”, Soviet Physics Doklady, 10(8), pp.707-710, 1966.

[5] 村上仁一, 藤波進, “ 日本語と英語の対訳文対の収集と著作権の考察 ”, 第一回コーパス日本語学ワークショップ予稿集, pp.119-130, 2012.