概要

現在，機械翻訳システムの翻訳品質の評価において，複数の評価手法が提案されている．評価手法には，大きく分けて２つある．人手評価と自動評価である．人手評価は，文全体を人手で評価する．よって大量の文を処理する際，コストが高いデメリットがある．一方，自動評価は，文の参照文と出力文を機械的に比較し評価する．よって大量の文を処理する際，コストが低いメリットがある．しかし，過去の研究で，人手評価と自動評価の評価に差がでると報告されている．そこで本研究では，文一致を使用する新たな自動評価法を提案した．文一致とは，出力文と参照文の完全一致である．参照文は，入力文の対訳の正解文である．提案手法は，文一致を使用することで，人手評価と同様に文全体で自動評価が可能である．

そして，日英方向における単文の，ルールベース翻訳，句に基づく統計翻訳，階層型統計翻訳，2 種類のハイブリッド翻訳について，翻訳実験と人手評価をおこなった．提案手法と人手評価の相関を調査した結果，提案手法と人手評価に差が確認された．よって，提案手法は人手評価に近い自動評価法ではないと判明した．また翻訳実験の際，出力文を自動評価法である BLEU，NIST，METEOR，RIBES のスコアを求め，提案手法との相関係数を調査した．結果，提案手法と他の自動評価法は，高い数値で正に相関していることが判明した．

また，異なる翻訳システムを使用すると，提案手法と他の自動評価法はどのような結果を示すか調査するため，追加実験をおこなった．追加実験は，句に基づく統計翻訳の学習文に，フレーズ対を追加した実験である．入力文は単文，重文複文の 2 種類で実験を行った．結果，入力文に重文複文を使用した場合，提案手法と他の自動評価法は，高い数値で正に相関していることが判明した．

今後は，提案手法にマルチリファレンスと N-best を使用した実験を検討する．

目 次

図 目 次

表 目 次

第1章　はじめに

現在，機械翻訳システムの翻訳品質の評価において，複数の評価手法が提案されている．翻訳品質の評価手法には，人手評価と自動評価がある．しかし，人手評価と自動評価には差がある [1].

松本ら [2] らは，人手評価の自動評価に差がある原因として，以下を報告している．人手評価は，文全体の単語を比較し，動詞のような重要な単語に着目し評価する．一方，自動評価は，出力文と参照文を部分的に比較している．よって，人手評価と自動評価に差が生じる．

そこで本研究では，人手評価と自動評価に差がでるのは，自動評価が出力文と参照文の文全体で評価せず，部分的な単語の比較で評価をおこなうことに原因があると仮定する．そして，出力文と参照文の文一致数で翻訳品質を評価する，新たな自動評価法を提案する．文一致数とは，出力文と参照文において，文を構成する単語が完全に一致する文数である．文一致数の評価により，文全体で評価が可能である．また本研究では，提案手法の有効性を，提案手法と人手評価の相関で調査する．

ここで，本論文の構成を以下に示す．第 2 章で，翻訳システムの説明を行う．第 3 章で，自動評価と人手評価の説明を行う．第 4 章で，人手評価と自動評価の違いについて述べる．第 5 章で，提案手法の説明を行う．第 6 章で，実験環境の説明を行う．第 7 章で，提案手法と人手評価の結果を示す．第 8 章で，本研究の考察を述べる．第 9 章で，追加実験の結果を示す．第 10 章で，結論を述べる．

第2章　翻訳システム

2.1　ルールベース翻訳

本研究では，ルールベース翻訳を”RBMT”と表記する．

RBMT は，人手によって構築された変換規則を元に翻訳を行うシステムである．長所として，規則を厳密に定義するので，規則が存在する翻訳において，精度が高い．しかし，短所として，規則が存在しない翻訳において，精度が低い．さらに人手によって規則を構築するため，開発コストが高い．

一般的な RBMT の手順を以下に示す．

手順 1 辞書の品詞などから原言語の構文解析を行う．

手順 2 目的言語の語順に変換する．

手順 3 再度辞書を参照し，助詞，助動詞などの不足語を補い，目的言語の出力文を生成する．

2.2 句に基づく統計翻訳

本研究では，句に基づく統計翻訳を，”PSMT”と表記する．

統計翻訳は，文法構造が近い言語間で翻訳精度が高い．しかし，文法構造の異なる言語間で翻訳精度が低い．

PSMT の概略を図 2.1 に示す．

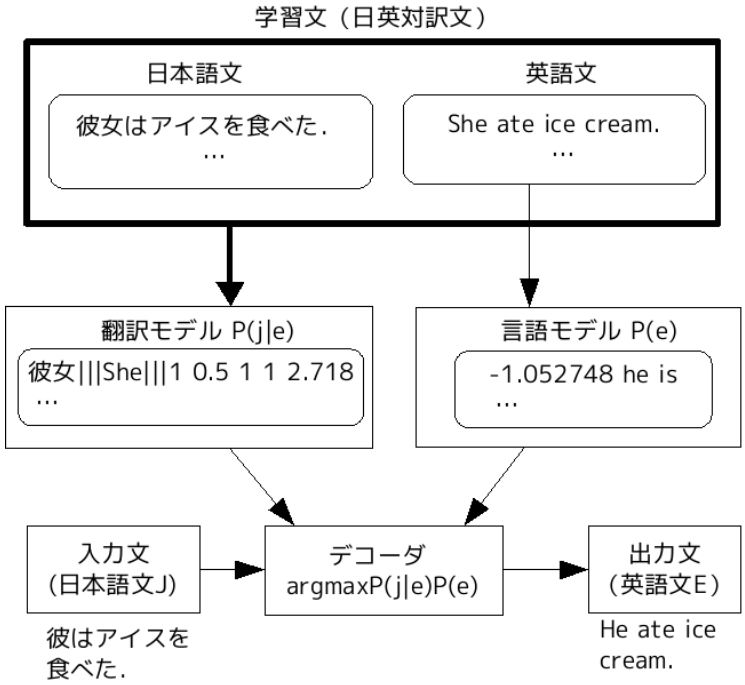

図 2.1 句に基づく統計翻訳

日英統計翻訳システムは，入力文（日本語文 J）が与えられたとき，デコーダを用いて翻訳モデル $P(j|e)$ と言語モデル $P(e)$ の確率を組み合わせ，確率が最大となる英語文 E を求めて翻訳を行う．$P(j|e)$ は e が j に翻訳される確率である．式をベイズの定理を用いて以下に示す．

$$E = argmax_e P(e|j) \tag{2.1}$$

$$= argmax_e \frac{P(j|e)\,P(e)}{P(j)} \tag{2.2}$$

$$= argmax_e P(j|e)\,P(e) \tag{2.3}$$

2.2.1 翻訳モデル

翻訳モデルは，単語または単語列の翻訳確率を組み合わせるモデルである．日英翻訳において，日本語の単語列から英語の単語列へ確率的に翻訳を行うため用いる．また，翻訳モデルは，表 2.1 のようなフレーズレーズテーブルで管理されている．

表 2.1 フレーズテーブルの例

おもしろい 本 ||| interesting book ||| 1 0.157258 1 0.180402 2.718
おもしろい 話 ||| funny stories ||| 0.5 0.0112613 0.2 0.000721527 2.718
おもちゃ 箱 ||| toy box||| 0.125 0.037 0.142 0.201 2.718
タイ 政府 ||| Thai government ||| 0.5 0.2778 0.5 0.093438 2.718

左から，日本語フレーズ，英語フレーズ，フレーズの英日翻訳確率 $P(j|e)$，英日方向の単語翻訳確率の積，フレーズの日英方向の翻訳確率 $P(e|j)$，日英方向の単語翻訳確率の積，フレーズペナルティ(一定) となっている．

2.2.2 言語モデル

言語モデルは，単語または単語列に対して，生成確率を付与するモデルである．日英翻訳では，言語モデルを用いて，生成された翻訳候補から英語を選出する．統計翻訳では一般に，N-gram モデルを用いる．N-gram モデルの例を表 2.2 に示す．なお，表 2.2 は，2-gram(2 単語間) である．

表 2.2 N-gram モデルの例

-1.782704 ||| I am ||| -0.04873917
-1.610493 ||| that is ||| -0.01120672
-2.346281 ||| train goes ||| -0.09572452
-1.868116 ||| woman and ||| -0.1343922

表 2.2 において，一番上の行は，左から，“I”の後に “am”が続く確率を常用対数で表した値 “$\log_{10}(P(am|I)) = -1.782704$”，2-gram で表現された単語列 “I am”，バックオフスムージングにより推定された “I”の後に “am”が続く確率を常用対数で表した値 “$\log_{10}(P(am|I)) = -0.04873917$”である．

バックオフスムージングとは，高次の N-gram の値が存在しない場合，低次の N-gram の値から推定する手法である．低次の確率を改良したスムージングの手法は，Kneser-Ney スムージングである．言語モデルの N-gram の作成においては，一般的に Kneser-Ney スムージングが用いられる．

2.2.3 デコーダ

デコーダは翻訳モデル $P(j|e)$ と言語モデル $P(e)$ を組み合わせて，確率が最大となる翻訳候補を探索し，出力する．デコーダの動作例を図 2.2 に示す．

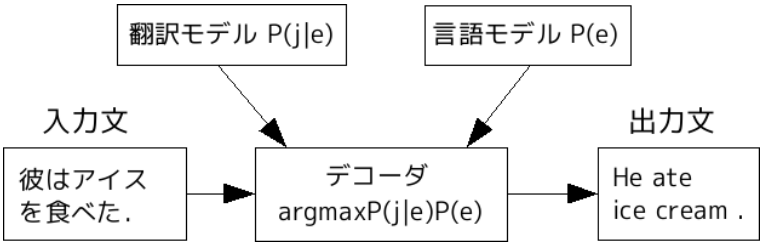

図 2.2 デコーダの動作例

日英翻訳において，$\arg\max_e P(j|e)P(e)$ の確率が最大となる英語文を出力するために，日本語と英語の単語対応を適切な順序で選択する必要がある．しかし，全探索をおこなうには，膨大な計算量と時間が必要となる．そこで，計算量と時間を削減するために，ビームサーチ法を用いる．

2.2.4 パラメータチューニング

パラメータチューニングは，デコーダで用いるパラメータの最適化を行う．一般的に評価関数 (BLEU) を最大にする翻訳結果が選ばれるように，パラメータ調整を行う．なお，パラメータ調整に，試し翻訳を行うデータとして，ディベロップメントデータを用いる．各文に対して上位 100 個程度の翻訳候補を出力し，重みを変えて翻訳候補が上位にくるようにパラメータを調整する．

2.3 階層型統計翻訳

本研究では，階層型統計翻訳を，”HSMT”と表記する．

HSMT とは，階層句を用いて翻訳を行う統計翻訳システムである．句を階層にすることで構文の評価が可能となる．また階層型統計翻訳は，語の並び替えを文脈自由文法で表現する．HSMT の概略を図 2.3 に示す.

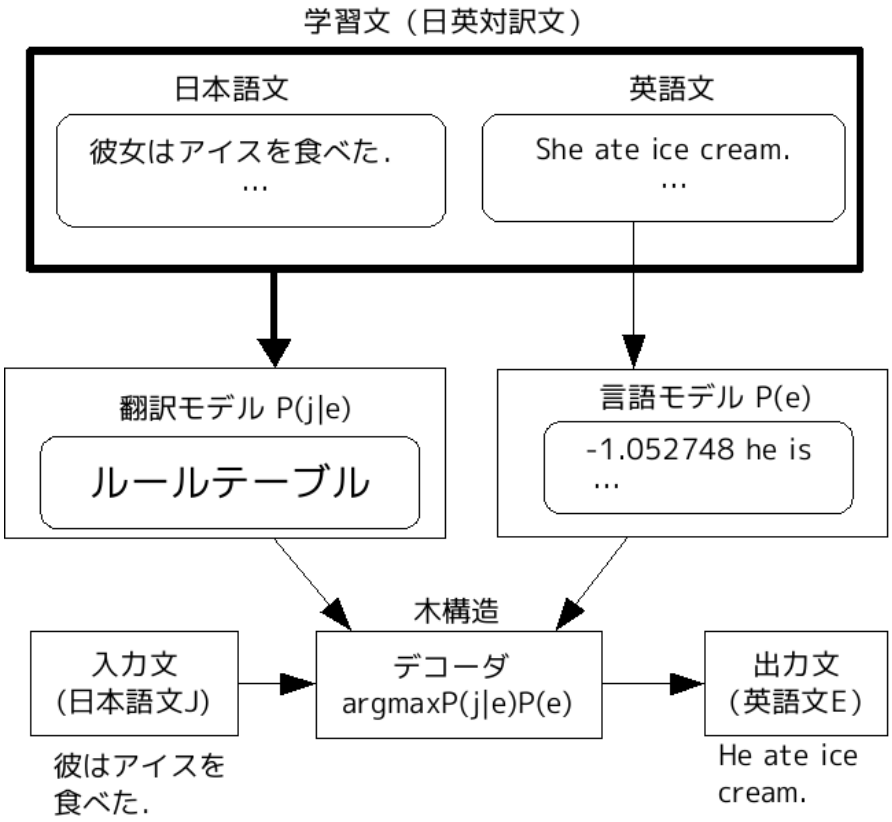

図 2.3 階層型統計翻訳システムの枠組み

HSMT は，翻訳モデルにルールテーブルを用いる点で，PSMT と異なる．さらにデコーダで，木構造を用いて翻訳を行う点も異なる．

2.3.1 翻訳モデル

HSMT の翻訳モデルは，ルールテーブルを用いる．

2.3.2 デコーダ

デコーダは翻訳モデル $P(j|e)$ と言語モデル $P(e)$ を組み合わせて，確率が最大となる翻訳候補を探索し，出力する．

日英翻訳において，$\arg\max_e P(j|e)P(e)$ の確率が最大となる英語文を出力するために，言語モデルと翻訳モデルを用いて翻訳を行う．しかしデコーダにおいて，木構造で翻訳を行うため，適切な英語文を決定させるために，膨大な計算量と時間が必要となる．

HSMT におけるデコーダの動作例を示す．なお，階層句を表 2.3 に示し，デコーダの動作の例を図 2.4 に示す．

表 2.3 階層句の例

X1 found that X2	X1 は X2 だとわかった。
She is X3	彼女が X3 だ
a music teacher	音楽の先生
My mother	私の母

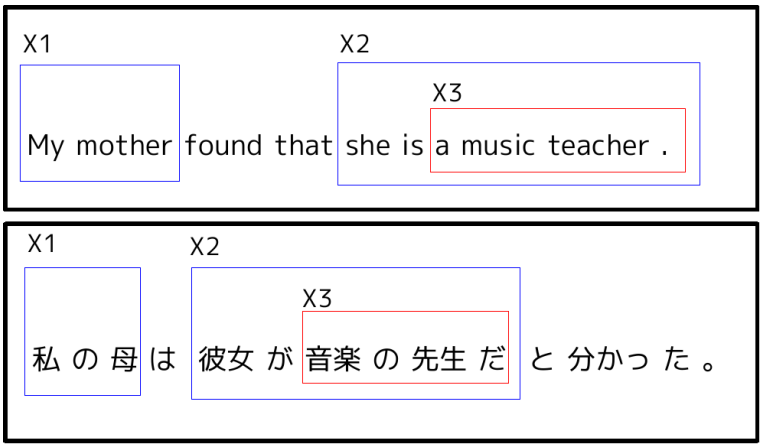

図 2.4 デコーダの動作例

2.4　ハイブリッド翻訳

2.4.1　概要

ハイブリッド翻訳は，前処理にルールベース翻訳を用いる．そして後処理に統計翻訳を使用する翻訳システムである．

本研究では，後処理に句に基づく統計翻訳を用いる場合，“RBMT+PSMT”と表記する．また，後処理に階層型統計翻訳を用いる場合，“RBMT+HSMT”と表記する．本節では，“RBMT+PSMT”を例に挙げてハイブリッド翻訳システムを説明する．

本研究では，英英統計翻訳を英’英統計統計翻訳と定義する．RBMT+PSMT の概略を図 5.1 に示す．

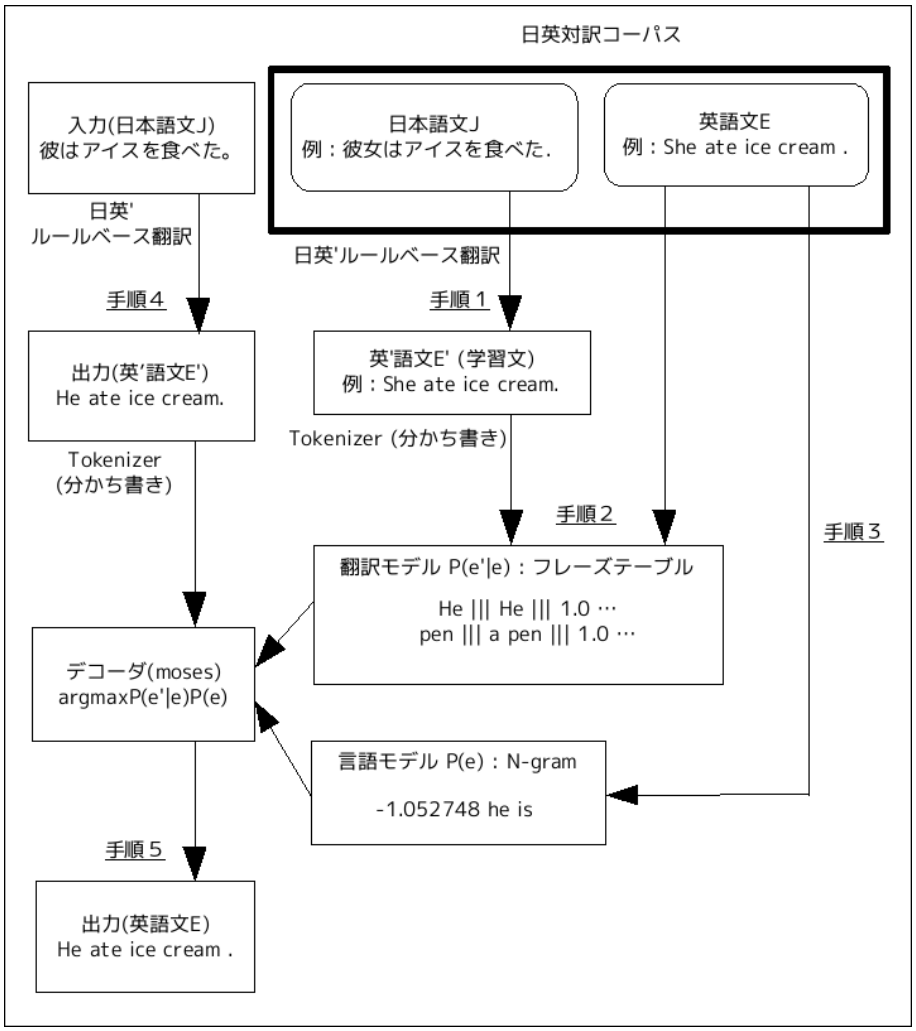

図 2.5　日英ハイブリッド翻訳の枠組

2.4.2 手順

学習の手順を以下に示す

手順 1 ルールベース翻訳を用いて，日英対訳コーパスの日本語文を英’語文に翻訳する．翻訳例を以下に示す．

入力文 (日本語文)	あの 人 の 家 は すぐ 見つかっ た 。
出力文 (英’語文)	That person ’s house was found immediately .
参照文	I soon found that person ’s house .
入力文 (日本語文)	列車 が 着い て いる 。
出力文 (英’語文)	The train has reached .
参照文	The train is in .
入力文 (日本語文)	コーヒー が 飲み たい 。
出力文 (英’語文)	I would like to drink coffee .
参照文	I ’d like some coffee .

手順 2 手順 1 で作成した英’語文と日英対訳コーパスの英語文を用いて，翻訳モデルを作成する．英’英フレーズテーブルの例を以下に示す．

I polished	||| I polished	||| 1 0.0388231 1 0.0748095 2.718
got injured .	||| was fatally wounded .	||| 1 0.0479841 1 2.72564e-05 2.718
is dancing	||| dancing	||| 0.037037 0.00659718 0.166667 0.333333 2.718

手順 3 日英対訳コーパスの英語文を用いて，言語モデルを作成する．N-gram モデルの例を以下に示す．

-3.425136	His computer
-3.494154	our TV
-0.1251315	due to

翻訳の手順を以下に示す

手順 1 ルールベース翻訳を用いて，テスト文の日本語文を英’語文に翻訳する．翻訳例を以下に示す.

入力文 (日本語文)	ウイスキー を 1 杯 もらお う 。
出力文 (英’語文)	I will get whiskey one cup .
参照文	I ’ll have a whiskey .
入力文 (日本語文)	この 理論 は くずれる だろ う 。
出力文 (英’語文)	This theory will collapse.
参照文	This theory won ’t hold water .
入力文 (日本語文)	手続き は 個人 でし て 下さい 。
出力文 (英’語文)	Procedure is an individual and please give it to me .
参照文	Carry out the procedure by yourselves , please .

手順 2 手順 1 で作成した英’語文を入力文として，英’英統計翻訳を行う．なお，翻訳モデル，言語モデルは手順 2，手順 3 で作成されたものを使用する．翻訳例を以下に示す.

入力文 (英’語文)	I will get whiskey one cup .
出力文 (英語文)	Let ’s get whiskey a cup .
参照文	I ’ll have a whiskey .
入力文 (英’語文)	This theory will collapse.
出力文 (英語文)	This theory will fail .
参照文	This theory won ’t hold water .
入力文 (英’語文)	Procedure is an individual and please give it to me .
出力文 (英語文)	There is an individual Please give it to me .
参照文	Carry out the procedure by yourselves , please .

2.5 各システムの翻訳例

RBMT，PSMT，HSMT，RBMT+PSMT，RBMT+HSMT の翻訳例を表 2.4，表 2.5，表 2.6 に示す.

表 2.4 各システムの翻訳例 1

入力文	電気 コンロ の コイル が 焼き 切れ た 。
RBMT	The coil of the electric cooker was able to be burned off .
PSMT	The The buckets or of electricity .
HSMT	The electric The of the cooking stove .
RBMT+PSMT	The company of the electric cooker was burned out .
RBMT+HSMT	The heavy coil in the electric cooker was burned out .
参照文	The heater coil is burnt out .

表 2.5 各システムの翻訳例 2

入力文	もっと 右 へ 寄っ て ください 。
RBMT	Please come to visit the right more .
PSMT	more to the right .
HSMT	more to the right .
RBMT+PSMT	Please come visit to the right .
RBMT+HSMT	Please come visit to the right .
参照文	Please move over more to the right .

表 2.6 各システムの翻訳例 3

入力文	彼 の 考え方 は 極端 すぎる 。
RBMT	His view is too going too far.
PSMT	His way of thinking is too the extreme .
HSMT	His way of thinking is too the extreme .
RBMT+PSMT	His opinion is too going too far .
RBMT+HSMT	His way of thinking is too going too far .
参照文	His way of thinking goes too far .

第3章　評価手法

3.1　人手評価

人手評価は，利点として，文法や意味を正確に評価可能である．しかし欠点として，時間と人件費が膨大にかかるため，大量の文の評価は難しい．

人手評価には，様々な評価方法がある．大きく分けて2種類ある．絶対評価と相対評価である．絶対評価には，了解度と正確さの観点から9段階で評価を行う手法，Adequacy(意味が伝わっているか) と Fluency(読みやすさ) の観点からそれぞれ5段階で評価を行う手法，さらに10点満点で評価を行う手法などがある．相対評価には，2つの翻訳システムの翻訳結果を比較して，翻訳品質が高いほうを良い評価とする，対比較評価などがある．

例として，Adequacy と Fluency の5段階評価の評価基準を表3.1に示す．

表 3.1　Adequacy と Fluency の評価基準

ランク	Adequacy	Fluency
5	入力文の意味が全て伝わっている．	かなり読みやすい．
4	入力文の意味がほとんど伝わっている．	少し読みやすい．
3	入力文の意味はかなり伝わっている．	ほとんど変わらない．
2	入力文の意味が少し伝わっていない．	少し読みにくい．
1	入力文の意味が全く伝わっていない．	かなり読みにくい．

3.2 自動評価

3.2.1 BLEU

BLEU[3] は，機械翻訳システムの自動評価において，現在主流な評価法である．BLEU は，N-gram 適合率で評価を行う．実験では 4-gram を用いる．BLEU は 0 から 1 のスコアを算出し，スコアが大きい方が良い評価である．BLEU の計算式を以下に示す．

$$BLEU = BPexpW_n \sum_{n=1}^{N} (log_e P_n) \tag{3.1}$$

$$W_n = \frac{1}{N} \tag{3.2}$$

$$P_n = \frac{\sum_i \text{出力文中} i \text{と参照文} i \text{で一致した} N-gram \text{数}}{\sum_i \text{出力文中} i \text{の中の全} N-gram \text{数}} \tag{3.3}$$

ここで，BP は短い翻訳文が高い評価にならないように補正を行うパラメータである．また W_n は N-gram の重みである．具体的な計算例を以下に示す．

例

日本語文：お先 に 失礼 し ます 。
参照文：Excuse me , I must be going now .
出力文：Excuse me , but I mest be going now .

計算方法

参照文と出力文の N-gram より計算を行うと

$$P_1 = \frac{9}{10}, P_2 = \frac{7}{9}, P_3 = \frac{5}{8}, P_4 = \frac{3}{7}, W_1 = 1, W_2 = \frac{1}{2}, W_3 = \frac{1}{3}, W_4 = \frac{1}{4} \tag{3.4}$$

これらのスコアを計算式に代入すると

$$BLEU \text{ スコア} = e^{W_4(logP_1+logP_2+logP_3+logP_4)} \tag{3.5}$$

$$= e^{\frac{1}{4}(log\frac{9}{10}+log\frac{7}{9}+log\frac{5}{8}+log\frac{3}{7})} \tag{3.6}$$

$$= 0.6580 \tag{3.7}$$

また BLEU は，英語とフランス語などの文法構造が近い言語間において，人手評価と評価が一致する場合が多い．しかし，英語と日本語などの文法構造が異なる言語間において，人手評価と評価が一致しない場合がある．原因として，BLEU は部分的な単語列の一致数を調べ，スコアを求めていることが挙げられる．そのため，参照文との比較に

おいて，同一の単語列を局所的に含む出力文が高いスコアを算出する．したがって，出力文において，文法的な誤りが存在しても高いスコアを算出してしまう．表 3.2 に具体的な例文を示す．なお，表 3.2 に対応する BLEU スコアを表 3.3 に示す．

表 3.2 翻訳例

入力文	その 機械 の 構造 に は 欠陥 が ある 。
出力文 1	The structure of the machine has a defect .
出力文 2	The structure of **the is** a fault in the machine .
参照文	There is a fault in the machine 's construction .

表 3.3 1 文における BLEU スコア

出力文 1	BLEU = 0.000
出力文 2	BLEU = 0.367

表3.3 より，出力文 1 と出力文 2 を比較すると，1 文における BLEU スコアは，出力文 2 が良い評価となる．しかし出力文 2 は “the is” と出力されているので，文法的に誤っている．

3.2.2 NIST

NIST[3] は，BLEU と同様に N-gram 適合率で評価を行う．情報量で重み付けしている点が異なる．また，実験では 5-gram を用いる．NIST は 0 から∞のスコアを出力し，スコアが大きい方が良い評価である．NIST の計算式を以下に示す．

$$NIST = \sum_{n=1}^{N} \frac{\sum_{i} \left(\sum_{\text{出力文 } i \text{ と参照文 } i \text{ に共通する } w_i \cdots w_n} Info\left(w_1 \cdots w_n\right) \right)}{\sum_{i} \text{出力文 } i \text{ の中の全 } N-gram \text{ 数}} \tag{3.8}$$

$$Info\left(w_i \cdots w_n\right) = \log_2 \frac{\text{評価コーパス中の } w_1 \cdots w_{n-1} \text{数}}{\text{評価コーパス中の } w_1 \cdots w_n \text{数}} \tag{3.9}$$

3.2.3 METEOR

METEOR[4] は，単語属性が正しい場合に高いスコアを出す．実験では *uni*-gram を用いる．METEOR は 0 から 1 までのスコアを出力し，スコアの大きい方が評価が良い評価である．計算式を以下に示す．

$$F\text{値} = \frac{P \times R}{\alpha \times P + (1-\alpha) \times R} \tag{3.10}$$

$$Pen = \gamma \times (\frac{c}{m})^{\beta} \tag{3.11}$$

$$METEOR = F \times (1 - Pen) \tag{3.12}$$

METEOR は F 値，ペナルティ関数 Pen を用いて計算される．F 値は適合率 P と再現率 R の調和平均で求められる．そしてペナルティ関数 Pen において，m は参照文と出力文の間で一致した単語数を示す．また c は，一致した単語を対象として，参照文と一致する単語列を 1 つのまとまりに統合した際のまとまりの数を示す．したがって，参照文と出力文が同一文である場合は c=1 となる．なお α，β，γ の値はパラメータである．具体的な計算例を以下に示す．

例

日本語文：お先 に 失礼 し ます 。
参照文：Excuse me , I must be going now .
出力文：Excuse me , but I mest be going now .

計算方法

参照文 B と出力文 A，A と B の重複部分 C とする．またパラメータ $\alpha = 0.8, \beta = 2.5, \gamma = 0.4$ とする．

$$\text{適合率}\, P = \frac{C}{A} = \frac{9}{10} \tag{3.13}$$

$$\text{再現率}\, R = \frac{C}{B} = \frac{9}{9} \tag{3.14}$$

$$F\text{値} = \frac{P * R}{\alpha * P + (1-\alpha) * R} = \frac{45}{46} \tag{3.15}$$

$$\text{ペナルティ関数}\, Pen = \gamma * (\frac{c}{m})^{\beta} = 0.4 * (\frac{2}{9})^{2.5} = 0.00931169\cdots \tag{3.16}$$

$$METEOR\text{スコア} = F * (1 - Pen) \tag{3.17}$$

$$= \frac{45}{46} * (1 - 0.0093) \tag{3.18}$$

$$= 0.9692 \tag{3.19}$$

3.2.4 RIBES

RIBES[5] は，参照文と出力文との間で，共通単語の出現順序を順位相関係数で評価を行う評価法である．計算式を以下に示す．

$$RIBES = NSR \times P^{\alpha} \tag{3.20}$$

$$RIBES = NKT \times P^{\alpha} \tag{3.21}$$

$$NSR = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n} \tag{3.22}$$

$$NKT = \frac{\sum_{i=1}^{n-1} K_i - \sum_{i=1}^{n-1} L_i}{\frac{n(n-1)}{2}} \tag{3.23}$$

ここで，NSR はスピアマンの順位相関係数であり，NKT はケンドールの順位相関係数である．P は共通単語が少ない場合のペナルティである．また α はペナルティに対する重みとして使用され，$0 \leqq \alpha \leqq 1$ の値である．n は文の単語数であり，d_i は参照文の i 番目の単語と出力文の i 番目の単語の語順の差分である．
K_i は，以下の 2 つの単語列の共起数である．

- 出力文における，出力文の i 番目の単語以降の単語列．
- 参照文における，出力文の i 番目の単語以降の単語列．

L_i は，以下の 2 つの単語列の共起数である．

- 出力文における，出力文の i 番目の単語以降の単語列．
- 参照文における，出力文の i 番目の単語以前の単語列．

RIBES は，単語の出現順を順位相関係数を用いて評価することで，文全体の語順に着目することができる．なお，RIBES は 0 から 1 のスコアを出力し，スコアが大きい方が良い評価である．具体的な計算例を次のページに示す．

例

日本語文：雨 に 濡れた ため ， 彼 は 風邪 を 引い た 。
参照文：He caught a cold because he got soaked in the rain .
出力文：He got soaked in the rain because he caught a cold .

計算方法

$$
\begin{aligned}
d_1 &= 1-8=-7, d_2=2-9=-7, d_3=3-10=-7, \\
d_4 &= 4-11=-7, d_5=5-7=-2, d_6=6-1=5, \\
d_7 &= 7-2=5, d_8=8-3=5, d_9=9-4=5, \\
d_{10} &= 10-5=5, d_{11}=11-6=5
\end{aligned}
$$

$$
NSR = 1-\frac{6(6*(-5)^2+(-2)^2+4*(-7)^2)}{11^3-11} = -0.59 \tag{3.24}
$$

$$
RIBES = \frac{-0.59+1}{2} = 0.2050 \tag{3.25}
$$

$$
\begin{aligned}
K_1 &= 5, K_2=4, K_3=3, K_4=2, K_5=1, \\
K_6 &= 0, K_7=0, K_8=3, K_9=2, K_{10}=1
\end{aligned}
$$

$$
\begin{aligned}
L_1 &= 5, L_2=5, L_3=5, L_4=5, L_5=5, \\
L_6 &= 5, L_7=4, L_8=0, L_9=0, L_{10}=0
\end{aligned}
$$

$$
NKT = \frac{21-34}{\frac{11*10}{2}} = -0.23 \tag{3.26}
$$

$$
RIBES = \frac{-0.23+1}{2} = 0.3850 \tag{3.27}
$$

第4章　人手評価と自動評価の違い

4.1　先行研究

松本ら [2] は，ルールベース翻訳とハイブリッド翻訳を用いて，人手評価と自動評価について考察した．結果，ルールベース翻訳とハイブリット翻訳の比較で，すべての人手評価と自動評価の結果に差が生じた．原因として，出力文の動詞の語訳を挙げている．動詞の語訳によって翻訳品質が下がり，人手評価が低下した．一方，自動評価は，動詞などの重要な単語でも，一定の割合で評価しているため，評価は低下しない．また人手評価は文全体の単語に着目し，評価する．一方，自動評価は，出力文と参照文を比較し，単語単位で均一に評価する．よって自動評価と人手評価に差が生じたと結論づけた．

表 7.1 に先行研究の自動評価の結果を示す．太字の数値は各自動評価でもっとも高い数値を表している．また表 4.3 に，先行研究の人手評価の結果を示す．表 4.3 の評価基準は表 4.2 に示す．

表 4.1　自動評価結果

	RBMT	PSMT	HSMT	RBMT+PSMT
BLEU	0.1320	0.1341	0.1352	**0.1798**
NIST	4.8260	4.9239	4.9628	**5.5426**
METEOR	0.4724	0.4544	0.4551	**0.5078**
RIBES	0.7281	0.7114	0.7198	**0.7540**

表 4.2 評価基準

ルールベース翻訳○	ルールベース翻訳の方が優れている
ハイブリッド翻訳○	ハイブリッド翻訳が ルールベース翻訳より優れている
句に基づく統計翻訳○	句に基づく統計翻訳が ルールベース翻訳より優れている
階層型統計翻訳○	階層型統計翻訳が ルールベース翻訳より優れている
差なし	意味に差がない or 共に意味が不明瞭である
同一出力	出力文が完全に同じ文である

表 4.3 人手評価結果

ルールベース翻訳○	ハイブリッド翻訳○	差なし	同一出力
23	5	59	13
ルールベース翻訳○	句に基づく統計翻訳○	差なし	同一出力
34	3	63	1
ルールベース翻訳○	階層型統計翻訳○	差なし	同一出力
30	3	66	1

表 4.1 の自動評価は，ハイブリッド翻訳の時，もっとも高い評価をしている．しかし，表 4.3 の人手評価は，ハイブリッド翻訳より，ルールベース翻訳が高い評価をしている．よって，先行研究で人手評価と自動評価の差が確認された．

4.2 問題点

先行研究は，人手評価と自動評価に差があると報告した．人手評価は，参照文を必ずしも必要とせず，文全体に対し，重要な単語の意味や文法が正しいか評価を行う．一方，自動評価は，参照文を必ず必要とし，単語または単語列に対し，出力文と参照文を比較して一致する割合で評価を行う．

表 4.4 の例文で，人手評価と自動評価にどのように差が生じるか解説する．

表 4.4 例文

入力文	ホラー映画を見るのは楽しくありません。
出力文	It is fun to see the horror movie .
参照文	It is not fun to watch the horror movie .

表 4.4 において，人手評価は，入力文と出力文を比較する．出力文の意味は，“ホラー映画を見るのは楽しいです。”であり，入力文の “ホラー映画を見るのは楽しくありません。”と意味が逆転しているので，評価は低い．

一方，表 4.4 において，自動評価は，出力文と参照文を比較する．その時，出力文と参照文の，アンダーラインが引かれている “It is”，“fun to”，“the horror movie .”を比較し，文の大部分が一致しているので，評価は高い．

よって，人手評価と自動評価に差が生じる．

第5章　提案手法

5.1　本研究の前提

本研究では，人手評価と自動評価に差が出るのは，自動評価が出力文と参照文の文全体で評価せず，単語もしくは単語列の比較で評価することに原因があると仮定する．

そこで本研究では，出力文と参照文の文全体で評価する自動評価法を提案する．

5.2　提案手法

提案手法は，文一致数で翻訳品質を評価する，新たな自動評価法である．文一致とは，出力文と参照文の完全一致である．文一致数とは，出力文と参照文が文一致した数である．本研究では，各翻訳システムにおいて，10,000 文の出力文と 10,000 文の参照文を比較し，文一致数で評価する．提案手法は文一致を使用するため，文全体で自動評価が可能である．

文一致の例を図 5.1 に示す．図 5.1 中の出力文と参照文の，太字で示した”The moon is rising .”が文一致している．

出力文
・・・・・・・
A meal is done .
The moon is rising .
Years passed away unawares .
The report is wrong .
Thread tangled .
・・・・・・・

参照文
・・・・・・・
A meal is ready .
The moon is rising .
Years slipped by .
The report is in error .
The thread got tangled .
・・・・・・・

図 5.1　文一致の例

5.3 提案手法の評価

提案手法の評価は，人手評価との相関係数で行う．相関係数が正の相関を示し，かつ高い数値の場合，提案手法と人手評価は差が小さいと言える．

第6章　実験環境

6.1　翻訳システム

本研究の実験に使用した翻訳システムの実験環境を，以下で説明する．

6.1.1　ルールベース翻訳

ルールベース翻訳には，東芝の Tauras[7] を使用する．

6.1.2　句に基づく統計翻訳

6.1.2.1　翻訳モデルの学習

翻訳モデルの学習に，“train-model.perl[8]” を使用する．

6.1.2.2　言語モデルの学習

言語モデルの学習に，“SRILM[9]” の “ngram-count” を使用する．本研究では，N-gram モデルは 5-gram とする．またスムージングに，“Kneser-Ney discount”を使用する．

6.1.2.3　デコーダ

デコーダは”Moses[8]”を使用する．

6.1.2.4　パラメータチューニング

デコーダの Moses において，パラメータは，“mert-moses.pl[8]” を使用し，チューニングを行う．また，Moses[8] の設定ファイル “moses.ini”の修正を行う．“distortion-limit” の値は，パラメータチューニングで変更されない．よって，手作業で “distortion-limit” の値を，-1(無制限) に変更する．“distortion-limit”はフレーズの並び替えにおける制約である．

6.1.3 階層型統計翻訳

6.1.3.1 翻訳モデルの学習

翻訳モデルの学習に，“train-model.perl[8]” を使用する．

6.1.3.2 言語モデルの学習

言語モデルの学習に，“SRILM[9]” の “ngram-count” を使用する．本研究では，N-gram モデルは 5-gram とする．またスムージングに，“Kneser-Ney discount”を使用する．

6.1.3.3 デコーダ

デコーダは”Moses[8]”を使用する．

6.1.4 ハイブリッド翻訳

ハイブリッド翻訳は，以下の 2 種類を使用する．

- RBMT+PSMT
- RBMT+HSMT

6.2 実験データ

実験には，辞書の例文より抽出した単文コーパス [6] から表 6.1 に示す文数を使用する．

表 6.1 実験に使用する文

日本語学習文	100,000 文
英語学習文	100,000 文
テスト文	10,000 文
ディベロップメント文	1,000 文

また統計翻訳の前処理として，日本語文に対し，“Mecab[10]”を使用し，形態素解析を行う．また，英語文に対し，“tokenizer.perl[8]”を使用し，分かち書きを行う．表 6.2 に単文コーパスの例を示す．

表 6.2 単文コーパスの例

日本語文	私 は 家 の 外 に 出 た 。
英語文	I went outside the house .
日本語文	私 は 山 に 登っ た 。
英語文	I climbed a mountain .
日本語文	私 は 雷 を 恐れる 。
英語文	I have a horror of thunder .

6.3 評価方法

6.3.1 人手評価

人手評価は，入力文と出力文と参照文を比較し，翻訳品質を 1 から 5 の数値でランクづけする手法で行う．ランク 1 がもっとも悪い評価で，ランク 5 がもっとも良い評価である．評価対象は，各翻訳システムの 10,000 文の出力文から，ランダムに 100 文抽出した文である．評価基準を表 6.3 に，評価の例を表 6.4 から表 6.8 に示す．

表 6.3 人手評価の評価基準

ランク	ランク付与基準
5	文法が正しく，言いたいことがすぐわかる． ネイティブレベルの文で，重要な情報の欠落はない．
4	文法が正しく，言いたいことがすぐわかる． 重要な情報の欠落はない．
3	文法に誤りがあるが，言いたいことがすぐわかる． 重要な情報の欠落はない．
2	文法に誤りがあるが，言いたいことがすぐわからない． 重要な情報の欠落はない．
1	文法に誤りがあり，言いたいことも分からない． 重要な情報が欠落している．

表 6.4 ランク 5 の例

	評価文	人手評価
入力文	その 技術 は 元来 軍事 用 に 開発 さ れ た 。	5
出力文	The technology was originally developed for military affairs .	
参照文	The technology was originally developed for military use .	

表 6.5 ランク 4 の例

	評価文	人手評価
入力文	陪審 の 評決 が 出 た 。	4
出力文	The jury’s verdict came out .	
参照文	The jury has reached a verdict .	

表 6.6 ランク 3 の例

	評価文	人手評価
入力文	彼ら は 秘密 の うち に 結婚 し た 。	3
出力文	They got married to the inside of the secret .	
参照文	They were married in private .	

表 6.7 ランク 2 の例

	評価文	人手評価
入力文	彼 は 歓迎 会 の 会場 を テーブル から テーブル へ と 歩き回っ た 。	2
出力文	He trod to the table of the welcome from the table .	
参照文	He circulated from table to table at the reception .	

表 6.8 ランク 1 の例

	評価文	人手評価
入力文	この 町 に は あまり 見る 所 が ない 。	1
出力文	There 's no doesn 't have much as this town .	
参照文	There are not many sights to see in this town .	

6.3.2 自動評価

本研究では，自動評価に BLEU[3]，NIST[3]，METEOR[4] ，RIBES[5] を使用する．

第7章　実験結果

7.1　提案手法と人手評価

単文コーパスを使用し，日英翻訳を行った．各翻訳システムにおける，提案手法と人手評価の結果と相関係数を，表 7.1 に示す．表 7.1 の”人手評価”の欄は，ランクづけした 100 文の平均値である．また，各評価手法においてもっとも高い数値を太字で，もっとも低い数値をアンダーラインで示す．

表 7.1　提案手法と人手評価

翻訳システム	提案手法	人手評価
RBMT	<u>157</u>	**4.15**
PSMT	178	<u>2.42</u>
HSMT	187	2.65
RBMT+PSMT	**323**	3.44
RBMT+HSMT	304	3.50
提案手法と人手評価の相関係数		0.1819

表 7.1 において，提案手法は，RBMT+PSMT の時，もっとも高い数値であった．また，RBMT の時，もっとも低い数値であった．一方，人手評価は，RBMT の時，もっとも高い数値であった．RBMT において，提案手法と人手評価に差がでた．

また，提案手法と人手評価の相関係数は”0.1819”であった．正の相関を示しているが，数値が非常に低い．よって，提案手法と人手評価には差があり，提案手法は有効性がない．

第8章　考察

8.1　提案手法と人手評価の差

7.1 節で，提案手法と人手評価に差があることが判明した．どのように差が生じたか調査するため，以下に人手評価をした例文を挙げる．

提案手法がもっとも高い評価をした RBMT+PSMT の例文と NIST の数値を，表 8.1，表 8.2 に示す．また，提案手法がもっとも低い評価をした RBMT の例文と NIST の数値を，表 8.3，表 8.4 に示す．

表 8.1　RBMT+PSMT の例

	評価文	人手評価
入力文	彼女 は まだ 口紅 を つける 年 で は ない 。	1
出力文	She is not a year is still on in on lipstick .	
参照文	She is not old enough to use lipstick .	

表 8.2　自動評価の数値 1

BLEU	0.0000
NIST	1.3208

表 8.1 の出力文は，”year”の次に”is”があり，また”on in on”という意味不明な単語列が出力されている．文法に誤りがあり，理解不能なので，評価は 1 である．

表 8.3 RBMT の例

	評価文	人手評価
入力文	この ネクタイ と お 洋服 と は よく 合います 。	5
出力文	This necktie and clothes suit well .	
参照文	This tie and your suit go well together .	

表 8.4 自動評価の数値 2

BLEU	0.0000
NIST	1.7349

表 8.3 の出力文は，参照文と表現の違いはあるが，文法的に正しく，理解が容易にできるので，評価は 5 である．

8.2 提案手法の問題点

本節では，提案手法の問題点を，表 8.5 の例文で考察する．

表 8.5 提案手法の問題点

	評価文	人手評価
入力文	彼の妹はとてもかわいい。	5
出力文	His younger sister is so cute .	
参照文	His younger sister is so pretty .	

表 8.5 において，人手評価は，入力文と出力文を比較し，単語の意味と文法が正しいので，高い評価をする．しかし，提案手法は，出力文と参照文を比較し，文一致していないので，低い評価をする．

出力文中の”cute”は，参照文中の”pretty”と，ほとんど同じ意味である．しかし，違う表現をされている．よって，提案手法では，翻訳品質が高い表 8.5 の出力文でも，低い評価をしてしまうという問題点がある．

本節で示した提案手法の問題点は，マルチリファレンスや N-best を使用することで解決する可能性がある．

8.3 提案手法と自動評価

本節では，提案手法と自動評価法との関係性を考察する．表 8.6 で，各翻訳システムにおける提案手法と自動評価の結果を示す．表 8.6 中で，各自動評価法でもっとも高い数値を太字で示す．また，提案手法と自動評価の相関係数を，表 8.7 で示す．

表 8.6 提案手法と自動評価

	提案手法	BLEU	NIST	METEOR	RIBES
RBMT	157	0.1296	4.7750	0.4695	0.7253
PSMT	178	0.1345	4.8199	0.4520	0.7115
HSMT	187	0.1417	4.9065	0.4573	0.7198
RBMT+PSMT	**323**	**0.1727**	5.3639	**0.4979**	0.7455
RBMT+HSMT	304	0.1698	**5.4070**	**0.4979**	**0.7479**

表 8.6 から，各自動評価法は，すべてハイブリッド翻訳時にもっとも高い数値であった．

表 8.7 提案手法との相関係数

BLEU	NIST	METEOR	RIBES
0.9950	0.9888	0.9104	0.9073

BLEU，NIST，METEOR，RIBES は，すべて 0.9 以上の数値で提案手法と正に相関した．特に BLEU との相関が大きかった．

8.4　今後の課題

8.2 節の提案手法の問題を解決する手法として，マルチリファレンスや N-best を使用する手法が挙げられる．本節では，マルチリファレンスと N-best について説明する．

8.4.1　マルチリファレンス

マルチリファレンスは，参照文の候補文が複数あることである．概略を図 8.1 に示す．参照文が増えることで，文一致しやすくなる．

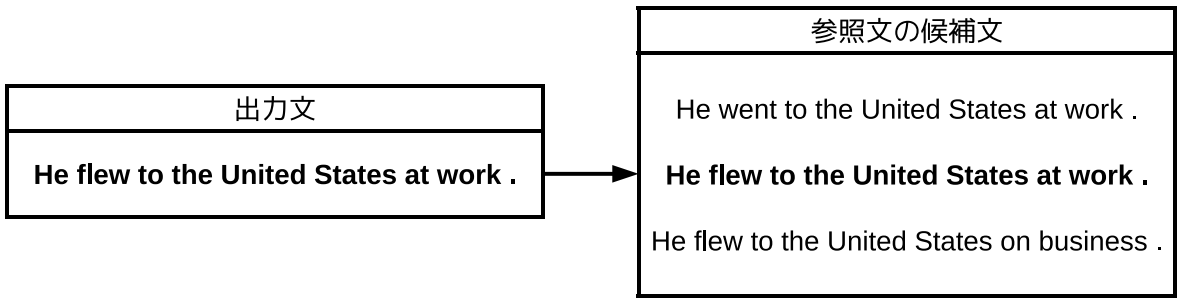

図 8.1　マルチリファレンスの例

8.4.2　N-best

N-best は，出力文の候補文が複数あることである．概略を図 8.2 に示す．出力文が増えることで，文一致しやすくなる．

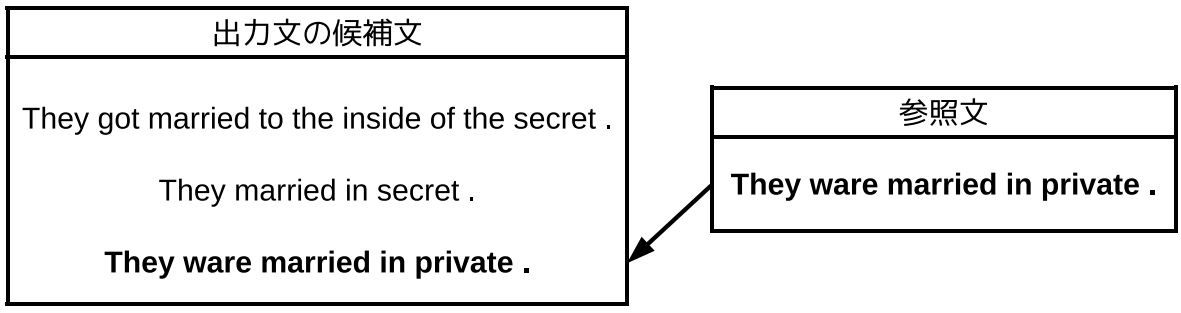

図 8.2　N-best の例

第9章　追加実験

9.1　実験

本実験では，日英方向と英日方向の PSMT において，学習データにフレーズ対を追加した実験を行う．実験は単文と重文複文それぞれで行う．追加するフレーズ対は，人手で作成された対訳フレーズ辞書の，英辞郎 [12] と鳥バンク [13] の 2 種類を使用する．そして，リオーダリングした場合と，リオーダリングしない場合についてそれぞれ実験を行う．したがって，合計 8 種類の翻訳実験を行う．さらに，それぞれの翻訳結果において，提案手法と自動評価との相関を調査する．

9.2　実験環境

9.2.1　翻訳モデルの学習

翻訳モデルの学習に，”train-model.perl を使用する．

9.2.2　言語モデルの学習

言語モデルの学習に，”SRILM の”ngram-count”を使用する．本実験では，”N-gram”モデルは”5-gram”とする．またスムージングに，”Kneser-Ney discount”を使用する．

9.2.3 英辞郎

英辞郎は，EDP(Electronic Dictionary Project) がアップデートし続けている英和・和英辞書である．英辞郎のデータには対訳フレーズ対の他に翻訳例や注釈など，本来の文に出てこない”〜”等の記号が含まれる．本実験では，英辞郎のクリーニングを行い，必要な英語と日本語のフレーズ対のみにした 1,366,575 フレーズ対を使用する．表 9.1 にクリーニング後の英辞郎のフレーズ対の例を示す．

表 9.1 クリーニング後の英辞郎のフレーズ対の例

日本語文	から 出て くる
英語文	come out from
日本語文	の 結果 として 生じる
英語文	come out from
日本語文	に 関する 情報 を 得る
英語文	obtain infomation on

9.2.4 鳥バンク

鳥バンクは，自然言語処理のための言語知識ベースを収録したデータバンクである．日本語の重文と複文を対象とする “意味類型パターン辞書”が収録されている．本実験では，鳥バンクから抽出した 698,472 フレーズ対を用いる．表 9.2 に鳥バンクのフレーズ対の例を示す．

表 9.2 鳥バンクのフレーズ対の例

日本語文	コート の すそ
英語文	the edge of my coat
日本語文	偉大 な 学者
英語文	become a great scholar
日本語文	カメラ を 買う
英語文	buy a camera

9.2.5 実験データ

本実験は，学習データとして辞書の例文より抽出した単文と重文複文の対訳文対を使用する．表 9.3 に，実験に英辞郎と鳥バンクのフレーズ対を追加しない場合の文数を示す．

表 9.3 コーパスを追加しない場合に使用する文数

学習データ (単文)	100,000 文
学習データ (重文複文)	100,000 文
テスト文 (単文)	10,000 文
テスト文 (重文複文)	10,000 文

9.3 実験結果

9.3.1 リオーダリングなし

本節では，リオーダリングなしの実験結果を示す．表 9.4 に，各実験と実験結果表の対応を示す．表 9.5 に，単文における日英統計翻訳の結果を示す．

表 9.4 対応表

表 9.5	単文における日英統計翻訳
表 9.6	単文における英日統計翻訳
表 9.7	重文複文における日英統計翻訳
表 9.8	重文複文における英日統計翻訳

表 9.5 単文 日英

	提案手法	BLEU	NIST	METEOR	RIBES
PSMT	206	0.1232	4.5103	0.4913	0.6986
PSMT+英辞郎	**213**	**0.1398**	**5.0113**	**0.5183**	**0.7181**
PSMT+鳥バンク	208	0.1390	4.9317	0.5132	0.7119
PSMT+鳥バンク+英辞郎	209	0.1385	4.9732	0.5133	0.7134
提案手法との相関係数		0.7190	0.7685	0.8095	0.8769

表 9.6 単文 英日

	提案手法	BLEU	NIST	RIBES
PSMT	168	0.1640	4.4104	0.6267
PSMT+英辞郎	**210**	**0.1837**	4.8253	**0.6499**
PSMT+鳥バンク	173	0.1735	4.7802	0.6423
PSMT+鳥バンク+英辞郎	189	0.1752	**4.8271**	0.6405
提案手法との相関係数		0.9232	0.6649	0.8176

表 9.7 重文複文 日英

	提案手法	BLEU	NIST	METEOR	RIBES
PSMT	28	0.0888	3.8093	0.4358	0.6450
PSMT+英辞郎	47	0.1068	4.3133	0.4631	0.6645
PSMT+鳥バンク	187	**0.2191**	**6.1100**	**0.5527**	**0.7112**
PSMT+鳥バンク+英辞郎	**190**	0.2179	6.0554	0.5517	0.7095
提案手法との相関係数		0.9996	0.9958	0.9951	0.9876

表 9.8 重文複文 英日

	提案手法	BLEU	NIST	RIBES
PSMT	19	0.1294	4.0328	0.5662
PSMT+英辞郎	29	0.1375	4.2647	0.5787
PSMT+鳥バンク	**110**	0.2304	**5.9541**	0.6482
PSMT+鳥バンク+英辞郎	109	**0.2307**	5.9400	**0.6483**
提案手法との相関係数		0.9997	1.0000	0.9994

表 9.7 と表 9.8 において，提案手法と他の自動評価法は，すべて 0.9 以上の数値で正に大きく相関した.

9.3.2 リオーダリングあり

本節では，リオーダリングありの実験結果を示す．表 9.9 に，各実験と実験結果表の対応を示す．

表 9.9 対応表

表 9.10	単文における日英統計翻訳
表 9.11	単文における英日統計翻訳
表 9.12	重文複文における日英統計翻訳
表 9.13	重文複文における英日統計翻訳

表 9.10 単文 日英

	提案手法	BLEU	NIST	METEOR	RIBES
PSMT	210	0.1265	4.5581	0.4945	0.7027
PSMT+英辞郎	220	0.1450	5.1004	0.5239	**0.7229**
PSMT+鳥バンク	215	**0.1502**	**5.1726**	**0.5297**	0.7159
PSMT+鳥バンク+英辞郎	**222**	0.1426	5.0447	0.5174	0.7178
提案手法との相関係数		0.6257	0.7206	0.6206	0.8788

表 9.11 単文 英日

	提案手法	BLEU	NIST	RIBES
PSMT	167	0.1686	4.4662	0.6341
PSMT+英辞郎	**214**	**0.1870**	4.8756	**0.6552**
PSMT+鳥バンク	176	0.1808	**4.9475**	0.6486
PSMT+鳥バンク+英辞郎	194	0.1799	4.8906	0.6473
提案手法との相関係数		0.8583	0.5609	0.8469

表 9.12 重文複文 日英

	提案手法	BLEU	NIST	METEOR	RIBES
PSMT	29	0.0914	3.8531	0.4387	0.6497
PSMT+英辞郎	51	0.1110	4.3594	0.4660	0.6686
PSMT+鳥バンク	**210**	**0.2692**	**7.0049**	**0.5913**	**0.7194**
PSMT+鳥バンク+英辞郎	202	0.2315	6.2661	0.5602	0.7174
提案手法との相関係数		0.9899	0.9849	0.9883	0.9917

表 9.13 重文複文 英日

	提案手法	BLEU	NIST	RIBES
PSMT	19	0.1332	4.0821	0.5740
PSMT+英辞郎	29	0.1397	4.2929	0.5832
PSMT+鳥バンク	109	**0.2464**	6.0813	**0.6564**
PSMT+鳥バンク+英辞郎	**113**	0.2410	**6.0950**	0.6560
提案手法との相関係数		0.9970	0.9996	0.9994

リオーダリングなしの実験と同様に，表 9.12 と表 9.13 において，提案手法と他の自動評価法は，すべて 0.9 以上の数値で正に大きく相関した．

9.4 追加実験の考察

9.3.1 項と 9.3.2 項から，重文複文コーパスを使用した結果，日英方向・英日方向の翻訳に関わらず，提案手法と他の自動評価法は全て，“0.98”以上の数値で正に相関した．

第10章　おわりに

本研究では，人手評価と自動評価の評価に差があるのは，自動評価が出力文と参照文の部分的な単語の比較で評価しているためと仮定した．そこで，文一致数で翻訳品質を評価する自動評価法を提案した．しかし，提案手法と人手評価との相関が小さかった．したがって，提案手法の有効性はないと判明した．また，提案手法と他の自動評価法の相関係数を調査した結果，提案手法と他の自動評価法は全て “0.9” 以上の数値で正に相関していた．

今後は，提案手法にマルチリファレンスと N-best を使用した実験を検討する．

謝辞

最後に，1 年間に渡ってご指導いただきました鳥取大学工学部知能情報工学科計算機 C 研究室の村田真樹教授，村上仁一准教授，徳久雅人講師そして計算機工学講座 C 研究室の方々に心から御礼申し上げます．また，よくアドバイスをくれた岡崎響君と，参考にさせていただいた論文の著者の方々に深く感謝致します．

参考文献

[1] 東江恵介, 出羽達也, 村上仁一, “日英方向におけるハイブリッド翻訳とルールベース翻訳の人手評価”, 言語処

[2] 松本 拓也, 村上仁一, 徳久雅人, ” 機械翻訳における人手評価と自動評価の考察”, 言語処理学会 18 年次大会, E2-8, pp.505-508, 2012. 理学会第 17 回年次大会, D5-5, pp.1127-1130, 2011.

[3] BLEU, NIST, “Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics”, 2002.

[4] Alon Lavie, Abhaya Agrwal, “METEOR: An Automatic Metric for MT Evaluation with High Level of Correlation with Human Judgments”, Proceedings of the ACL 2007 Workshop on Statistical Machine Translation, 2007.

[5] 平尾努, 磯崎秀樹, Kevin Duh, 須藤克仁, 塚田元, 永田昌明, “RIBES: 順位相関に基づく翻訳の自動評価法”, 言語処理学会第 17 年次大会, D5-2, pp.1111-1114, 2011.

[6] 村上仁一, 徳久雅人, “日英対訳データベースの作成のための 1 考察”, 言語処理学会第 17 回年次大会, D4-5, pp.979-982, 2011.

[7] Shinya Amano, Hideki Hirakawa, Yoshinao Tsutsumi: “TAURAS: The Toshiba machine translation system”, Manuscr Program MT Summit, pp.15–23, 1987.

[8] Philipp Koehn, Marcello Federico, Brooke Cowan, Richard Zens, Chris Dyer, Alexandra Constantin, Evan Herbst,“Moses: Open Source Toolkit for Statistical Machine Translation”, Proceedings of the ACL 2007 Demo and Poster Sessions, pp.177-180, 2007.

[9] Andreas Stolcke: “SRILM - am Extensible Language Modeling Toolkit”, 7th International Conference on Spoken Language Processing, pp.901–904, 2002.

[10] Taku Kudo, Kaoru Yamamoto, Yuji Matsumoto: “Applying Conditional Random Fields to Japanese Morphological Analysis”, Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004), pp.230–237, 2004.

[11] 日野 聡子, 村上 仁一, 徳久雅人, 村田真樹, “統計翻訳における英辞郎を利用したパラレルコーパスの効果”, 言語処理学会第 17 回年次大会, P2-29, pp400-403, 2011.

[12] 英辞郎, http://www.alc.co.jp/

[13] 鳥バンク, http://unicorn.ike.tottori-u.ac.jp/toribank/, 2007.