概要

近年では，翻訳システムにおいて，機械によって自動的に評価を行う自動評価が盛んになっている．自動評価は，機械的に評価を行うため，開発コストが低いという利点があるという一方で，日英間における翻訳では，人手評価との間で評価が異なる場合があることが知られている．

そこで本研究では，ハイブリッド翻訳とルールベース翻訳を用いて，自動評価と人手評価の相関の調査を行った．結果として，ルールベース翻訳とハイブリッド翻訳の比較において，すべての自動評価と人手評価の結果に差異が生じた．原因としては，ハイブリッド翻訳において，出力文の動詞の誤訳が挙げられる．動詞の誤訳によって，翻訳品質が下がり，人手評価は大きく低下したと考えられる．しかし，自動評価においては，各単語を一定の割合で評価しているため，評価は大きく低下しない．よって評価結果に差異が生じたと考えている．

また，ルールベース翻訳と句に基づく統計翻訳，ルールベース翻訳と階層型統計翻訳の比較においては，自動評価の METEOR，IMPACT，RIBES が人手評価と同じ結果であった．よって，METEOR，IMPACT，RIBES においては他の自動評価法より信頼性があると考えている．

さらに，本研究では新たな評価法として，折り返し翻訳を利用した評価法を提案した．結果として，折り返し翻訳を利用した評価では，英日翻訳にルールベース翻訳を用いた場合に人手評価と同じ結果が得られた．しかし折り返し翻訳が成功した文数が少ないため信頼性は低く，改良の余地がある．

今後は，本実験で用いた 7 つの自動評価法以外においても，人手評価との相関を調査することを考えている．

目 次

図 目 次

表 目 次

第1章　はじめに

機械翻訳システムにおいて，自動評価は効率的な性能評価を行う上で重要である．近年提案されている自動評価法では，BLEU[1] が主流となっている．しかし，BLEU の自動評価と人手評価には差異がある場合が知られている [2][3]．越前谷らは，特許文を用いた自動評価法の調査を行い，自動評価と人手評価の相関関数に大きなばらつきがあることを報告した [4]．しかし，特許文は専門的であり，複雑な文であるので，原因を調査するのは困難である．

そこで本研究では，簡単な日本語の単文 [5] を用いて翻訳実験を行う．そして，自動評価と人手評価の相関を考察する．自動評価として，BLEU[1]，NIST[1]，METEOR[6]，IMPACT[7]，RIBES[8]，TER[9]，WER[9] の 7 種類の自動評価法を用いる．また人手評価として，対比較評価を行う．なお，翻訳システムには，句に基づく統計翻訳，ハイブリッド翻訳，ルールベース翻訳，階層型統計翻訳の 4 種類を用いる．さらに新たな評価手法として，日英翻訳と英日翻訳を組み合わせる “折り返し翻訳を利用した評価方法” を提案し，人手評価との相関を調査する．

結果として自動評価と人手評価の結果には差異が存在した．よって，今回用いた 7 つの自動評価法には問題があると考えている．

また，折り返し翻訳を利用した評価方法では，英日翻訳にルールベース翻訳を用いた場合に人手評価と同じ結果が得られた．しかし折り返し翻訳が成功した文数が少ないため信頼性は低く，改良の余地がある．ここで，本論文の構成を以下に示す．第 2 章において，翻訳システムの概要について説明を行う．第 3 章において，自動評価と人手評価についての説明を行う．第 4 章において，実験環境についての説明を行う．第 5 章において，自動評価と人手評価の結果を示す．第 6 章において，本研究の考察を述べる．第 7 章において，新たな評価手法として，折り返し翻訳を利用した評価について述べる．第 8 章において，結論を述べる．

第2章　翻訳システム

2.1　句に基づく統計翻訳

統計翻訳は，文法構造が近い言語間では翻訳精度が高い傾向にある．しかし，文法構造の異なる言語間では翻訳精度が低い傾向にある．初期の頃は，単語に基づく統計翻訳であったが，近年では句に基づく統計翻訳が主流となっている．句に基づく統計翻訳は，語順の並び替えなどにおいて，単語に基づく統計翻訳よりも優れている．

句に基づく統計翻訳システムの枠組みを図 2.1 に示す．

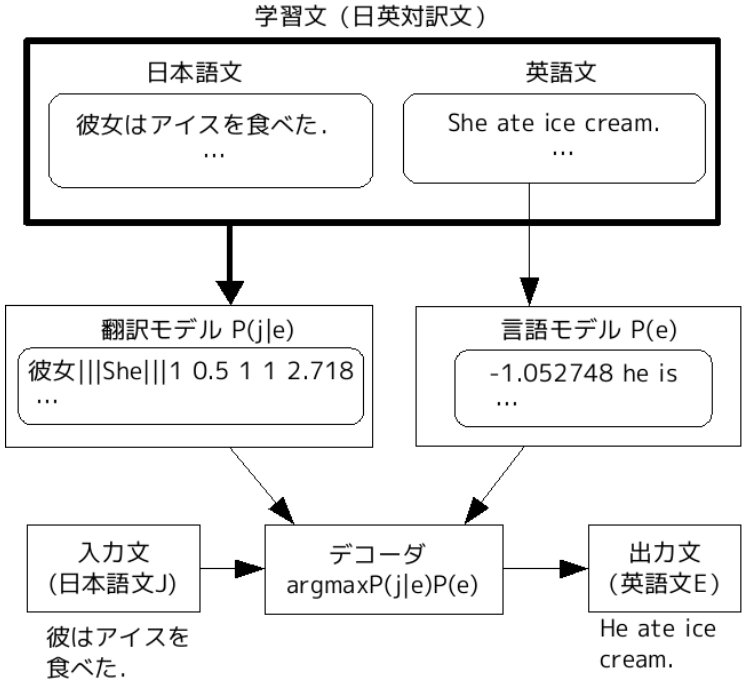

図 2.1　句に基づく統計翻訳システムの枠組

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 を求めることで翻訳を行う．$P(j|e)$ とは e が j に翻訳される確率のことである．式をベイズの定理を用いて以下に示す．

$$
\begin{aligned}
E &= argmax_e P(e|j) && (2.1)\\
&= argmax_e \frac{P(j|e)\,P(e)}{P(j)} && (2.2)\\
&= argmax_e P(j|e)\,P(e) && (2.3)
\end{aligned}
$$

2.1.1 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．日英翻訳においては，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に用いる．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(常に一定) となっている．

2.1.2 IBM 翻訳モデル

翻訳モデルの代表例として IBM 翻訳モデルがある．IBM 翻訳モデルは，Model1 から Model5 までの 5 つから構成され，計算が順に複雑となる．IBM 翻訳モデルは，仏英翻訳を想定している．よって，本章では，仏英翻訳を前提に説明を行う．IBM 翻訳モデルでは，フランス語文 f と英語文 e の翻訳モデル $P(f|e)$ を計算するために，アライメント a を用いて，以下の式を考える．

$$P(f \mid e) \;=\; \sum_{a} P(f, a \mid e)$$

IBM モデルでは，英単語は 1:n の対応を持ち，フランス単語は 1:1 の対応を持つと仮定する．なお，フランス単語に適切な対応関係を持つ英単語が存在しないときは，フランス語と英語文の先頭の特殊文字 e_0 を対応させる．

2.1.3 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，言語モデルを用いることで，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．

また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．この低次の確率を改良したスムージングの手法が，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．

2.1.4 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．

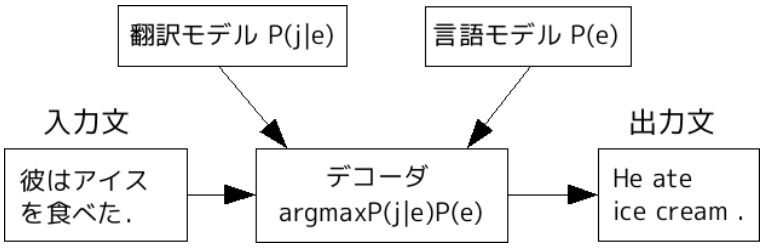

図 2.2 デコーダの動作例

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，適切な英語文を決定させるためには，膨大な計算量が必要となり，膨大な時間が必要となる．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．

2.1.5 パラメータチューニング

パラメータチューニングとは，デコーダで用いるパラメータの最適化を行うことである．一般的には評価関数 (BLEU) を最大にするような翻訳結果が選ばれるように，パラメータ調整を行う．なお，パラメータ調整では，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，重みを変えることで翻訳候補が上位にくるようにパラメータを調整する．

2.2　ハイブリッド翻訳

2.2.1　概要

本研究のハイブリッド翻訳は，前処理としてルールベース翻訳を用いる．さらに後処理として句に基づく統計翻訳を用いる翻訳システムである．また本研究においては，英英統計翻訳を英’英統計統計翻訳と定義する．日英ハイブリッド翻訳の枠組みを図 2.3 に示す.

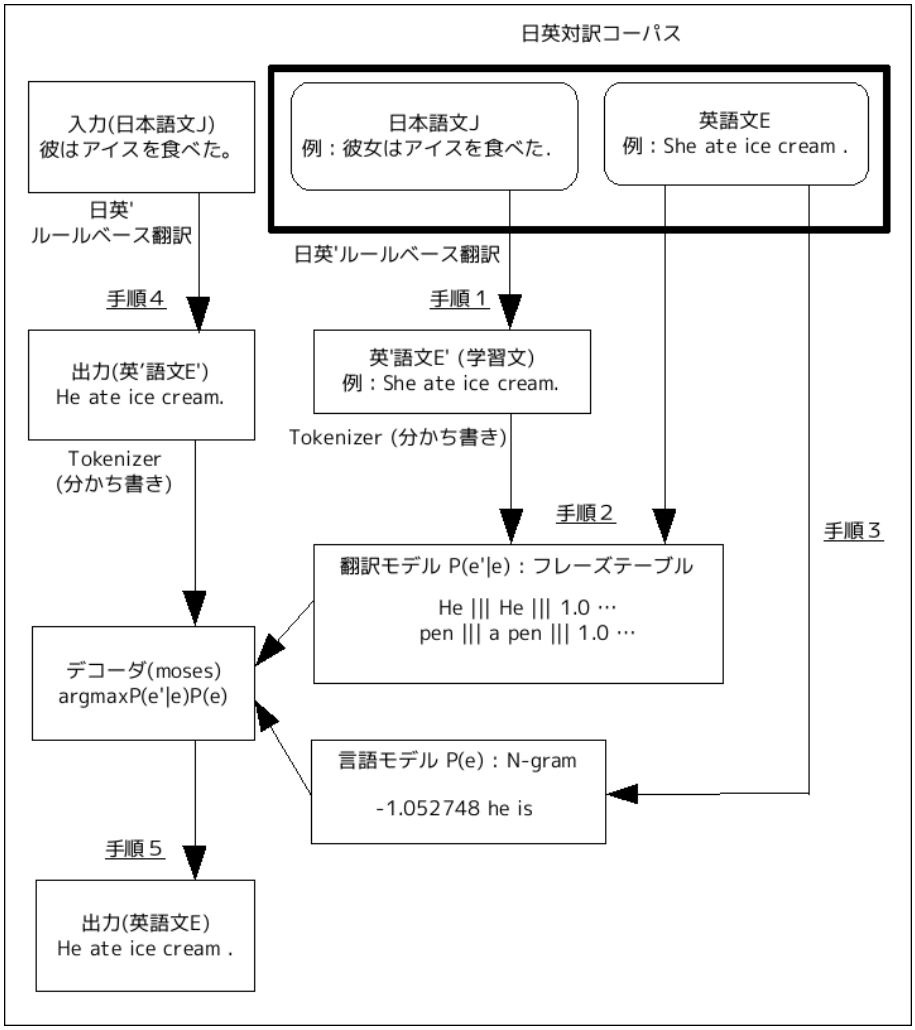

図 2.3　日英ハイブリッド翻訳の枠組

2.2.2 手順

学習の手順

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

手順 4 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 5 手順 4 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 .

2.3 階層型統計翻訳

階層句を用いて翻訳を行う統計翻訳システムであり，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．階層型統計翻訳システムの枠組みを図 2.4 に示す．

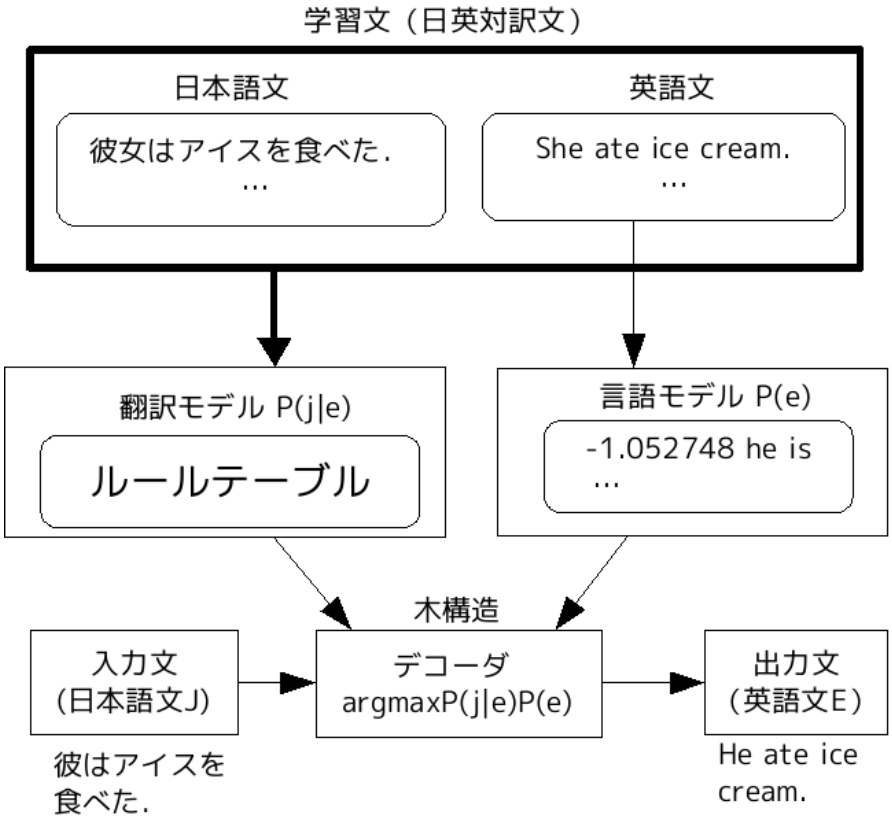

図 2.4 階層型統計翻訳システムの枠組み

句に基づく統計翻訳との違いは，翻訳モデルにルールテーブルを用いることが挙げられる．さらにデコーダにおいて，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

階層型統計翻訳の翻訳モデルにおいて，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，木構造で翻訳を行っているため，適切な英語文を決定させるためには，膨大な計算量が必要となり，膨大な時間が必要となる．

階層型統計翻訳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 2.5 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

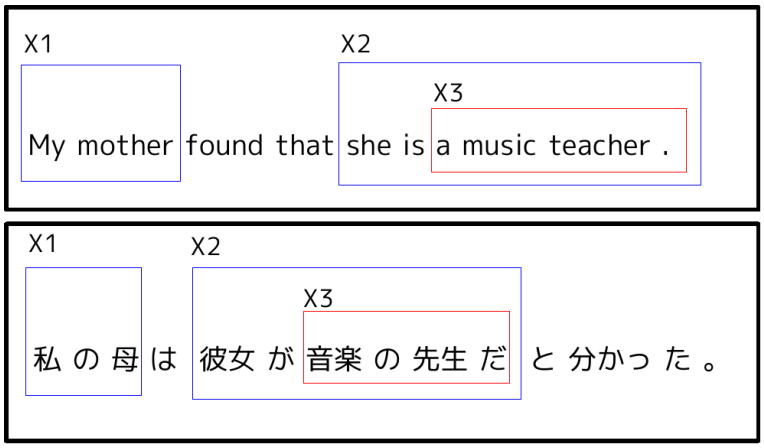

図 2.5 デコーダの動作例

2.4　ルールベース翻訳

ルールベース翻訳は，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．長所としては，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は，精度が高いことが挙げられる．しかし，短所としては，規則が存在しない翻訳においては，精度が低いことが挙げられる．さらに人手によって規則を構築するので，開発コストも高い．

一般的なルールベース翻訳の手順に示す．折り返し翻訳を用いて，日英翻訳におけるハイブリッド翻訳とルールベース翻訳の評価を行う．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．

ルールベース翻訳の例を表 2.4，表 2.5 に示す．

表 2.4　ルールベース翻訳の例 1

入力文	彼は大打撃を受けた。
ルールベース翻訳	He received the great blow .
参照文	He received a hard blow .

表 2.5　ルールベース翻訳の例 2

入力文	我々のチームが試合に勝った。
ルールベース翻訳	Our team won the game .
参照文	Our team won the game .

2.5 各システムの翻訳例

ルールベース翻訳，ハイブリッド翻訳，句に基づく統計翻訳，階層句に基づく統計翻訳の翻訳例を表 2.6，表 2.7，表 2.8 に示す.

表 2.6 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 。
ルールベース翻訳	The coil of the electric cooker was able to be burned off .
ハイブリッド翻訳	The company of the electric cooker was burned out .
句に基づく統計翻訳	The The buckets or of electricity .
階層型統計翻訳	The electric The of the cooking stove .
参照文	The heater coil is burnt out .

表 2.7 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 。
ルールベース翻訳	Please come to visit the right more .
ハイブリッド翻訳	Please come visit to the right .
句に基づく統計翻訳	more to the right .
階層型統計翻訳	more to the right .
参照文	Please move over more to the right .

表 2.8 各システムの翻訳例 3

入力文	彼 の 考え方 は 極端 すぎる 。
ルールベース翻訳	His view is too going too far.
ハイブリッド翻訳	His opinion is too going too far .
句に基づく統計翻訳	His way of thinking is too the extreme .
階層型統計翻訳	His way of thinking is too the extreme .
参照文	His way of thinking goes too far .

第3章　評価

3.1　自動評価

3.1.1　BLEU

BLEU は，機械翻訳システムの自動評価において，現在主流となっている評価法である．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 のスコアを出力し，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

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

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

$$= 0.6580 \tag{3.7}$$

また BLEU は，英語とフランス語のような文法構造が近い言語間において，人手評価と一致する場合が多い．しかし，英語と日本語のような文法構造が異なる言語間においては，人手評価と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ることにより，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 3.1 に具体的な例文を示す．なお，表 3.1 に対応する BLEU スコアを表 3.2 に示す．

表 3.1 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 3.2 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3.2 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，文法的な誤りを含んでいる．

3.1.2 NIST

NIST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．

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

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

3.1.3 METEOR

METEOR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．

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

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

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

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

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

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

$$= 0.9692 \tag{3.19}$$

3.1.4 IMPACT

IMPACT は，名詞句のチャンクを用いて評価を行う手法である．参照文と出力文において，対応する名詞句を用いて，一致する単語列を正確に決定する．さらに，名詞句の出現順に関して類似性を決定する．IMPACT は再現率と適合率から F 値を求め，F 値を IMPACT のスコアとしている．計算式を以下に示す．

$$R = \left(\frac{\sum_{i=0}^{RN} \left(\alpha^i \sum_{c \in LCS} length\,(c)^\beta \right)}{m^\beta} \right)^{\frac{1}{\beta}} \tag{3.20}$$

$$P = \left(\frac{\sum_{i=0}^{RN} \left(\alpha^i \sum_{c \in LCS} length\,(c)^\beta \right)}{n^\beta} \right)^{\frac{1}{\beta}} \tag{3.21}$$

$$score = \frac{(1+\gamma^2)\,RP}{R+\gamma^2 P} \tag{3.22}$$

$$\gamma = \frac{P}{R} \tag{3.23}$$

ここで，LCS とは最長共通部分列であり，RN は LCS の決定プロセスの繰り返し数を示す．そして，i は RN に対するカウンターで，α と β はパラメータである．また m は参照文の単語数，n は出力文の単語数，$length(c)$ は共通部分の単語数を示す．なお，IMPACT はスコアが大きい方が良い評価である．

3.1.5 RIBES

RIBES 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．

$$RIBES = NSR \times P^\alpha \tag{3.24}$$

$$RIBES = NKT \times P^\alpha \tag{3.25}$$

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．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．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．

3.1.6 TER

TERは，Translation Edit Rateの略で翻訳の誤り率を求める評価法である．計算式を以下に示す．

$$TER = \frac{\sum_i (\text{挿入語数}\, i + \text{置換語数}\, i + \text{削除語数}\, i + \text{シフト語数}\, i)}{\sum_i (\text{参照文}\, i \,\text{の平均単語数})} \tag{3.26}$$

分子は参照文と出力文の比較における編集操作数のことである．TERの編集操作は挿入，置換，削除，シフトの4種類の編集を行うことである．なお，TERはスコアが小さい方が良い評価である．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

例では挿入語数= 1 より，分子の編集操作数=1 である．また分母は参照文の平均単語数=9 である．

$$\begin{aligned} TER\,\text{スコア} &= \frac{\sum_i (\text{挿入語数}\, i + \text{置換語数}\, i + \text{削除語数}\, i + \text{シフト語数}\, i)}{\sum_i (\text{参照文}\, i \,\text{の平均単語数})} && (3.27)\\ &= \frac{1}{9} && (3.28)\\ &= 0.1111 && (3.29) \end{aligned}$$

3.1.7 WER

WERは，Word Error Rateの略で単語の誤り率を求める評価法である．以下に計算式を示す．

$$WER = \frac{\sum_i (\text{挿入語数}\, i + \text{置換語数}\, i + \text{削除語数}\, i)}{\sum_i (\text{参照文}\, i \,\text{の平均単語数})} \tag{3.30}$$

分子は参照文と出力文の比較における編集操作数のことである．WERの編集操作は挿入，置換，削除の3種類の編集を行うことである．なお，WERはスコアが小さい方が良い評価である．具体的な計算例を以下に示す．

例

日本語文： 事態は正常に戻っています。
参照文： Things are back to normal .
出力文： Things returned to normal .

計算方法

例では置換語数数=1，削除語数=1より，分子の編集操作数=2である．また分母は参照文の平均単語数=6である．

$$\begin{aligned} WER\,\text{スコア} &= \frac{\sum_i (\text{挿入語数}\, i + \text{置換語数}\, i + \text{削除語数}\, i)}{\sum_i (\text{参照文}\, i \,\text{の平均単語数})} && (3.31)\\ &= \frac{2}{6} && (3.32)\\ &= 0.3333 && (3.33) \end{aligned}$$

3.2 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ことが挙げられる．しかし欠点として，時間と人件費が膨大にかかることが挙げられ，大量の文の評価は極めて難しい．本実験では，対比較評価を行う．対比較評価とは，各出力文を比較することで評価を行う評価法である．対比較評価の判断基準については 5.2 節で詳しく説明する．

また人手評価には，他にも様々な評価方法がある．例えば，了解度と正確さの観点から 9 段階で評価を行う方法，理解容易性と忠実度の観点から 5 段階で評価を行う方法，さらに 10 点満点で評価を行う方法などがある．例として 10 点満点法 [10] を表 3.3 に示す．

表 3.3 10 点満点評価法

得点	評価点の付与基準
10 点	英語らしく明解で完全に理解できる． 用語，語形，構文に誤ったところがない．
9 点	もう少し英語らしい適切な言い方があるが，他は上記に同じ．
8 点	明解でほぼ完全に理解できる． しかし，あまり重要でない点で文法やスタイルに不適切さがあり， おかしな言葉使いがあるが，訂正は容易．
7 点	概して明瞭で理解できるが，スタイル，用語，構文が上記より若干貧弱．
6 点	言いたいことか大体すぐ分かる． しかし，スタイル，用語，表現選択のまずさ，翻訳もれの言葉， 文法的に誤った配置などがあり，包括的な理解が妨げられる． ポストエディットのできる限界．
5 点	良く考えると概要はほぼ分かる． 用語のまずさ，奇怪な構文，訳し漏れの言葉があり，正確さを欠く．
4 点	分かるような気がするが，実際には分からぬとも言える． 仮装行列のような訳．用語，構文，表現が全般的におかしく， 重要語の訳しもれがある．
3 点	全般的に理解不能． 意味がないように見えるが，よく考えてみると言いたいこと についての仮説ができる．部分的には分かるところがある．
2 点	部分的にも全体的にも理解不能だが，言いたいことが匂う．
1 点	殆ど絶望的だが，完全に無意味だとは言い切れない．
0 点	完全に理解不能． いくら考えても言っていることがさっぱり分からない． （アポートや訳文出力の無いものはこのランク）

第4章　実験環境

4.1　翻訳モデルの学習

翻訳モデルの学習には，“train-model.perl[11]” を用いる．

4.2　言語モデルの学習

言語モデルの学習には，“SRILM[12]” の “ngram-count” を用いる．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”を用いる．

4.3　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[11]” を用いてチューニングを行う．また，Moses[11] の設定ファイル “moses.ini”の修正も行う．“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のことである．

4.4 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 182,899 文 [5] から表 4.1 のように用いる．

表 4.1 実験に使用する文

英語学習文	100,000 文
日本語学習文	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，日本語文に対して，“MeCab[13]”を用いて，形態素解析を行う．また，英語文に対して，“tokenizer.perl[11]”を用いて，分かち書きを行う．表 4.2 に単文コーパスの例を示す．

表 4.2 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 .

4.5 評価方法

4.5.1 自動評価

本研究では，自動評価法としてBLEU[1]，NIST[1]，METEOR[6]，IMPACT[7]，RIBES[8]，TER[9] およびWER[9] を用いる．

4.5.2 人手評価

本研究では，人手評価として，ルールベース翻訳を基準する対比較評価を行う．対比較評価を行う対象を以下に示す．それぞれにおいて対比較評価を行う．表 4.3 に対比較評価のを行う翻訳を示す．

表 4.3 対比較評価の比較対象

ルールベース翻訳	ハイブリッド翻訳
ルールベース翻訳	句に基づく統計翻訳
ルールベース翻訳	階層型統計翻訳

第5章　翻訳実験

5.1　自動評価結果

テスト文を用いて，日英翻訳を行う．翻訳システムとして，句に基づく統計翻訳，ハイブリッド翻訳，ルールベース翻訳および階層型統計翻訳を用いる．それぞれの自動評価の結果を表 5.1 に示す．

表 5.1　自動評価結果

	ルールベース翻訳	ハイブリッド翻訳	句に基づく統計翻訳	階層型統計翻訳
BLEU	0.1320	**0.1798**	0.1341	0.1352
NIST	4.8260	**5.5426**	4.9239	4.9628
METEOR	0.4724	**0.5078**	0.4544	0.4551
IMPACT	0.4477	**0.4854**	0.4411	0.4476
RIBES	0.7281	**0.7540**	0.7114	0.7198
TER	0.7154	**0.6526**	0.7002	0.6834
WER	0.7393	**0.6776**	0.7296	0.7087

表 5.1 の結果より，すべての自動評価において，ハイブリッド翻訳が最良の値 を示した．しかし，ルールベース翻訳は，BLEU，NIST，TER および WER において，最悪の値を示した．

また，METEOR，IMPACT，RIBES は，句に基づく統計翻訳において，最悪の値を示した．

5.2 人手評価結果

本研究では，ルールベース翻訳に対して，ハイブリッド翻訳，句に基づく統計翻訳，階層型統計翻訳をそれぞれ比較することで，対比較評価を行う．手順としては，まず日英翻訳に対して，ハイブリッド翻訳，句に基づく統計翻訳および階層型統計翻訳の出力文からランダムに各 100 文抽出する．次に抽出した 100 文に対して，1 文毎に対比較評価を行う．なお，評価基準を表 5.2 に以下に示す．さらに人手評価の結果を表 5.3 に示す．

表 5.2 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 5.3 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 5.3 の結果より，すべての人手評価において ルールベース翻訳が最良 であることが示された．

5.3 自動評価と人手評価の比較のまとめ

実験結果より，自動評価の表 5.1 と人手評価の表 5.3 を比較すると，自動評価と人手評価の差異が示された．したがって，本研究で用いたすべての自動評価法に問題があると考えている．

また，自動評価法の METEOR，RIBES は，ルールベース翻訳と句に基づく統計翻訳および階層型統計翻訳において，人手評価と同様の結果となっている．よって METEOR，RIBES は，その他の自動評価法より信頼性があると考えている．

5.4 自動評価と人手評価に差異がある翻訳例

表 5.4 に，ハイブリッド翻訳とルールベース翻訳において，自動評価と人手評価の差異が確認できた例を示す．また句に基づく統計翻訳と階層型統計翻訳についても表 5.5，表 5.6 に示す．なお，() 内は BLEU のスコアを示す．

表 5.4 ルールベース翻訳とハイブリッド翻訳の対比較評価例

入力文	両者 の 間 に 商談 が 成立 し た 。
ルールベース翻訳 (0.000)	The business talk was materialized among both .
ハイブリッド翻訳 (0.3564)	The concluded negotiations between the two .
参照文	A bargain was arranged between the two .
人手評価	ルールベース翻訳◯

表 5.4 より，ハイブリッド翻訳とルールベース翻訳を比較すると，BLEU の値はハイブリッド翻訳が高い．しかし，ハイブリッド翻訳の出力文は，先頭に “The concluded” となっていて，The の後に主語となる単語がなく，文法的に誤りなので，人手評価ではルールベース翻訳が良いと判断した．

表 5.5 ルールベース翻訳と句に基づく統計翻訳の対比較評価例

入力文	彼女 の 長い 髪 は 風 に 波打っ て い た 。
ルールベース翻訳 (0.4317)	Her long hair was wavy to the wind .
句に基づく統計翻訳 (0.4468)	Her long hair in the wind .
参照文	Her long hair was streaming in the wind .
人手評価	ルールベース翻訳◯

表 5.5 より，mose とルールベース翻訳を比較すると，BLEU の値は句に基づく統計翻訳が高い．しかし句に基づく統計翻訳の出力文には，“波打っ て い た ”を表す動詞が存在しないので，人手評価ではルールベース翻訳のほうが良いと判断した．

表 5.6 ルールベース翻訳と階層型統計翻訳の対比較評価例

入力文	これ は 家族 みんな が 興味深く 読める 雑誌 です 。
ルールベース翻訳 (0.0000)	This is a magazine which all families can read interestingly .
階層型統計翻訳 (0.3124)	This is a family can read all 興味深く magazine .
参照文	This is a family interest magazine .
人手評価	ルールベース翻訳◯

表 5.6 より，階層型統計翻訳とルールベース翻訳を比較すると，BLEU の値は階層型統計翻訳が高い．しかし，階層型統計翻訳の出力文は，未知語が含まれ，翻訳品質が劣っているので，人手評価ではルールベース翻訳が良いと判断した．

第6章　考察

6.1　自動評価と人手評価の差異の原因

6.1.1　未知語の影響

人手評価において，未知語の影響により，翻訳品質が低下している文が多かった．よって，各翻訳システムの出力文において，未知語が含まれる文数を調査した．結果を表6.1に示す．また例を表6.2に示す．

表 6.1　未知語の含む文の数

ルールベース翻訳	ハイブリッド翻訳	句に基づく統計翻訳	階層型統計翻訳
307	433	3079	2889

表 6.2　動詞の誤訳の例 2

入力文	このページはインキがにじんでいた。
ルールベース翻訳	As for this page, ike was blurred .
句に基づく統計翻訳	The インキ にじん in this page .

表6.1より，句に基づく統計翻訳と階層型統計翻訳は他の2つのシステムに比べて，未知語を含んでいる文が2400文以上多く出力されている．一方，ルールベース翻訳とハイブリッド翻訳の比較では，未知語を含んでいる文の差は100文程度しかなく，大きな差異はみられない．しかし，ルールベース翻訳とハイブリッド翻訳の対比較評価では大きな差異が存在している．したがって，自動評価の問題として，未知語以外に原因があると考えている．

6.1.2 動詞の誤訳の問題

差異が存在した原因として，単語の重要度の違いが挙げられる．自動評価は，各単語を同じ割合で評価している．しかし人手評価は，各単語を同じ割合で評価していない．例えば，動詞は文全体の意味に与える影響は大きい．しかし助詞は文全体の意味に与える影響は小さい．よって翻訳文に動詞の誤訳が含まれると文質は低下し，人手評価において，評価も低下する．したがって，自動評価と人手評価では，各単語の重要度が異なっているので，評価結果に差異が存在したと考えている．動詞の誤訳の例として，動詞の欠落の例を表 6.3，表 6.5 に示す．

表 6.3 動詞の誤訳の例

入力文	話題が転じて教育問題の話になった。
ルールベース翻訳	Subject changed and it **became** a talk of the educational problem .
ハイブリッド翻訳	Subject changed and the story of educational problems .

表 6.4 1 文における自動評価結果

	BLEU	NIST	METEOR	IMPACT	RIBES	TER	WER
ルールベース翻訳	0.000	0.702	0.404	0.280	0.639	1.429	1.429
ハイブリッド翻訳	0.000	**1.248**	**0.600**	**0.485**	**0.760**	**0.857**	**0.857**

表 6.3 では，人手評価において，ルールベース翻訳が良いと判断できる．しかし表 6.4 では，自動評価において，ハイブリッド翻訳の方が良い．

表 6.5 動詞の誤訳の例 2

入力文	父は犬小屋を大きく作り替えた。
ルールベース翻訳	The father remade the doghouse greatly .
ハイブリッド翻訳	My father made the doghouse .

表 6.6 1 文における自動評価結果

	BLEU	NIST	METEOR	IMPACT	RIBES	TER	WER
ルールベース翻訳	0.0000	1.3879	0.3745	0.4416	0.7863	0.6667	0.6667
ハイブリッド翻訳	0.000	1.0566	0.4589	0.4508	0.7999	0.6667	0.6667

表 6.5 では，人手評価において，ルールベース翻訳が良いと判断できる．しかし表 6.6 では，自動評価において，ハイブリッド翻訳の方が良いと示している評価法が多い．

第7章　折り返し翻訳を利用した評価

7.1　折り返し翻訳を利用した評価の概要

本研究で用いた７つの自動評価法は，参照文を用いることで評価をしている．しかし，参照文の作成・入手は容易ではない．したがって本研究では，参照文を必要としない評価方法として，折り返し翻訳を利用した評価を提案する．折り返し翻訳を利用した評価の枠組みを図 7.1 に示す．

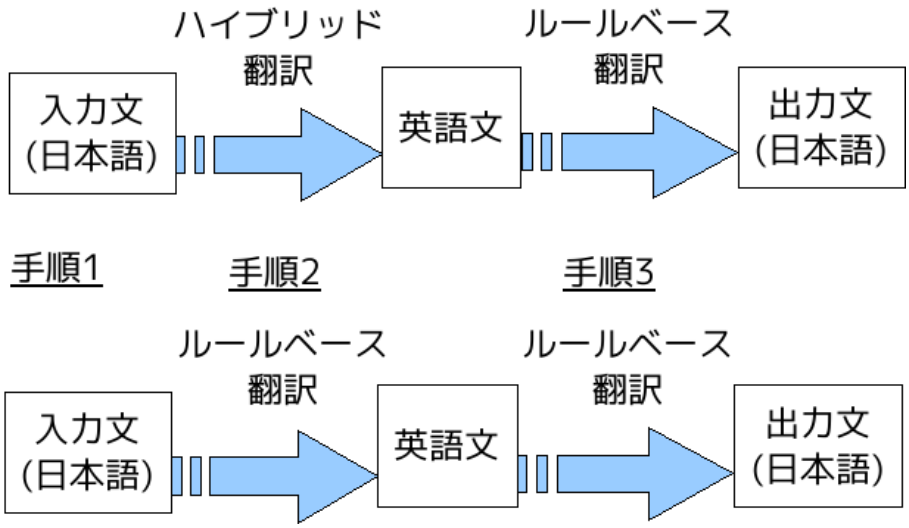

図 7.1　折り返し翻訳を利用した評価の枠組み

7.2　折り返し翻訳を利用した評価の手順

折り返し翻訳を用いて，日英翻訳におけるハイブリッド翻訳とルールベース翻訳の評価を行う．手順を以下に示す．

手順 1 入力文として，日本語文 10,000 文を準備する．

手順 2 ハイブリッド翻訳とルールベース翻訳を用いて，入力文の日英翻訳をそれぞれ行う．

手順 3 手順 2 で翻訳されたそれぞれの英語文に対して英日翻訳を行う．

英日翻訳にルールベース翻訳 を使用する．

手順 4 手順 1 の入力文 (日本語) と，手順 3 の出力文 (日本語) を比較する．

手順 5 手順 4 で完全に同一である日本語の文数を調査する．

なお，翻訳の過程で，未知語が含まれる文は，手順 5 においてカウントをしない．

手順 6 英日翻訳に統計翻訳 を用いた場合も行う．手順 1，手順 2，手順 4，手順 5 についても同様に行う．

7.3　折り返し翻訳を利用した評価の結果

テスト文を用いて，折り返し翻訳を利用した評価を行う．英日翻訳には，共に ルールベース翻訳 を用いる．折り返し翻訳を利用した評価結果を表 7.1 に示す．表 7.1 の翻訳システムは日英翻訳に用いた翻訳システムを示す．

7.3.1　英日翻訳にルールベース翻訳を用いた場合

表 7.1　日本語文が完全一致した文数

ルールベース翻訳	ハイブリッド翻訳
130	89

表 7.1 の結果より，折り返し翻訳を利用した評価では，ルールベース翻訳 が表 5.3 の人手評価と同様の結果となった．

7.3.2　英日翻訳に統計翻訳を用いた結果

テスト文を用いて，折り返し翻訳を利用した評価を行う．英日翻訳には，共に 統計翻訳 を用いる．折り返し翻訳を利用した評価結果を表 7.1 に示す．表 7.2 の翻訳システムは日英翻訳に用いた翻訳システムを示す．

表 7.2　日本語文が完全一致した文数

ルールベース翻訳	ハイブリッド翻訳
130	**252**

表 7.2 の結果より，折り返し翻訳を利用した評価では，ルールベース翻訳 が表 5.3 の人手評価と異なる結果となった．

7.4 折り返し翻訳の例

折り返し翻訳が成功した例を表 7.3，表 7.4 に示す．さらに，失敗した例を表 7.5，表 7.6 に示す．

表 7.3 折り返し翻訳の成功例

日本語文 (入力文)	彼女は長期休暇をとる。
英語文	She takes a long leave.
日本語文 (出力文)	彼女は長期休暇をとる。

表 7.4 折り返し翻訳の成功例

日本語文 (入力文)	私は山に登った。
英語文	I climbed the mountain.
日本語文 (出力文)	私は山に登った。

表 7.5 折り返し翻訳の失敗例

日本語文 (入力文)	私は家の外に出た。
英語文	I went out of the house.
日本語文 (出力文)	私は家から出ていった。

表 7.6 折り返し翻訳の失敗例

日本語文 (入力文)	もっと右へ寄ってください。
英語文	Please come visit to the right .
日本語文 (出力文)	右へ遊びに来てください。

7.5 折り返し翻訳を利用した評価の考察

折り返し翻訳を利用した評価では，英日翻訳にルールベース翻訳を用いた場合に人手評価と同様の結果が得られた．しかし，問題点として，折り返し翻訳により同一となった文が，10,000 文に対して，英日にルールベース翻訳を用いた場合で 130 文と 89 文，英日に統計翻訳を用いた場合で 130 文と 252 文しか存在しないことが挙げられる．よって，全体での折り返し成功率は 1%〜3%前後という結果となった．さらに，英日翻訳に統計翻訳を用いた実験においては，人手評価との間に差異が生じる結果となった．したがって，折り返し翻訳を利用した評価法は，信頼性が低く，改良の余地が多いと考えている．

7.6 追加実験

追加実験として英日翻訳において，日英翻訳で用いたシステムとは異なるルールベース翻訳システムを使用して折り返し翻訳を行う．英日翻訳で用いるシステムは “翻訳の王様” と呼ばれるルールベース翻訳システムである．表 7.7 に折り返し翻訳の成功した文数を示す．なお，表 7.7 には日英翻訳で用いたシステムを記載する．

表 7.7 翻訳の王様を用いた折り返し翻訳の一致文数

ルールベース翻訳	ハイブリッド翻訳	句に基づく統計翻訳	階層型統計翻訳
2	4	3	3

表 7.7 より，折り返し翻訳に “翻訳の王様” を用いた実験において，成功文数が 1 万文中 4 文以下と極めて低い結果となった．よって信頼性は低いと考えている．

第8章　おわりに

本研究では，単文を用いて，自動評価と人手評価を行い，結果を比較した．結果として，ルールベース翻訳とハイブリッド翻訳の比較において，すべての自動評価と人手評価の結果に差異があることを確認した．したがって，今回用いた 7 種類の自動評価法に問題があると考えている．

しかし METEOR，IMPACT，RIBES については，ルールベース翻訳と句に基づく統計翻訳の比較，さらにルールベース翻訳と階層句に基づく統計翻訳の比較において，自動評価と人手評価の結果が同じになった．よって METEOR，IMPACT，RIBES は他の自動評価法より信頼性があると考えている．

また折り返し翻訳を利用した評価では，英日翻訳にルールベース翻訳を用いた場合に人手評価と同じ結果となった．しかしながら，折り返し成功率が 1〜3%前後であり，信頼性は低いと考えている．よって，折り返し翻訳を利用した評価法は，今後改良の余地がある．

したがって今後は，さらに様々な自動評価法を検討し，人手評価と同様の結果が得られる評価法を調査していきたい．

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

[2] 福田智大, 村上仁一, 徳久雅人, 村上仁一, “ルールベース翻訳を前処理に用いた統計翻訳”, 言語処理学会第 16 回年次大会, PB2-12, pp.676-679, 2010.

[3] 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”, 言語処理学会第 17 回年次大会, D5-5, pp.1127-1130, 2011.

[4] 越前谷博, 下畑さより, 内山将夫, 宇津呂武仁, 江原暉将, 藤井敦, 山本幹夫, 神門典子, “NTCIR-7 データを用いた機械翻訳自動評価基準のメタ評価”, AAMT/Japio 特許研究会, pp.2-13, 2008.

[5] 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.

[6] 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.

[7] Hiroshi Echizen-ya, Kenji Araki, “Automatic Evaluation Method for Machine Translation using Noun-Phrase Chunking”, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp.108-117, 2010.

[8] 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.

[9] Richard Schwartz, Linnea Micciulla, John Makhoul. “A Study of Translation Edit Rate with Targeted Human Annotation”, AMTA, 2006.

[10] 池原悟, 白井諭, 小倉健太郎, “言語表現体系の違いに着目した日英機械翻訳機能試験項目の構成”, 人工知能学会, 1994.

[11] 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.

[12] SRILM, The SRI Language Modeling Toolkit
http://www.speech.sri.com/projects/srilm/

[13] MeCab, http://mecab.sourceforge.net/

[14] Harold Somers, “Round-Trip Translation:What Is It Good For?”