機械翻訳における人手評価と自動評価の考察

松本 拓也　　村上 仁一　　徳久 雅人
鳥取大学 工学部 知能情報工学科
{s082052, murakami, tokuhisa}@ike.tottori-u.ac.jp

1　はじめに

機械翻訳システムにおいて，自動評価法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 の評価と人手評価に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，日常の日本語とは異なるため，原因を調査するのは困難である．

そこで本研究では，日本語の簡単な単文 [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．翻訳システムには，句に基づく統計翻訳，ハイブリッド統計翻訳，ルールベース翻訳，階層句に基づく統計翻訳の 4 種類を用いる．また，自動評価として，7 種類の自動評価法を用いる．

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．

2　自動評価と人手評価

2.1　BLEU(自動評価)

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，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 1 に具体的な例文を示す．なお，表 1 に対応する BLEU スコアを表 2 に示す．

表 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 2　1 文における BLEU の値

出力文 1	BLEU = 0.000
出力文 2	BLEU = 0.367

表 2 より，BLEU スコアにおいて，出力文 1 と出力文 2 を比較すると，出力文 2 が良い評価 となる．

2.2　対比較評価 (人手評価)

人手評価は，翻訳された出力文に対して，人間の判断で翻訳の正確さ，明瞭度を調べることにより，品質を評価する手法である [6]．人手評価の利点として，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，大量の文における評価は極めて難しい．

そこで本研究では，人手評価として，対比較評価のみを行う．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．表 1 の例で対比較評価を行うと，出力文 2 は “the is”と誤った文法を出力しているのに対し，出力文 1 は正しい翻訳である．よって，出力文 1 が精度の高い評価 となる．

2.3　自動評価 (BLEU) と人手評価の差異

表 1 の例より，2.1 節の BLEU(自動評価) と 2.2 節の対比較評価 (人手評価) に差異があることが分かる．

3　実験環境

本研究では，自動評価と人手評価を比較するために，4 つのシステムを用いて実験を行う．

3.1　句に基づく統計翻訳：Phrase

本研究では，句に基づいて翻訳を行う統計翻訳システムとして，“moses[7]” を用いる．

3.1.1　言語モデルの学習

言語モデルの学習には，“SRILM[8]” 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．

3.1.2　翻訳モデルの学習

翻訳モデルの学習には，“train-model.perl[7]” を用いる．

3.1.3　パラメータチューニング

moses のパラメータは，“mert-moses.pl[7]” を用いてチューニングを行う．“distortion-limit”の値は-1(無制限) とする．

3.2　ハイブリッド統計翻訳：Hybrid

本研究のハイブリッド統計翻訳は，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．本研究で行った日英ハイブリッド統計翻訳の手順を以下に示す．

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．

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．

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．

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．

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．

日英ハイブリッド統計翻訳の枠組みを図 1 に示す．

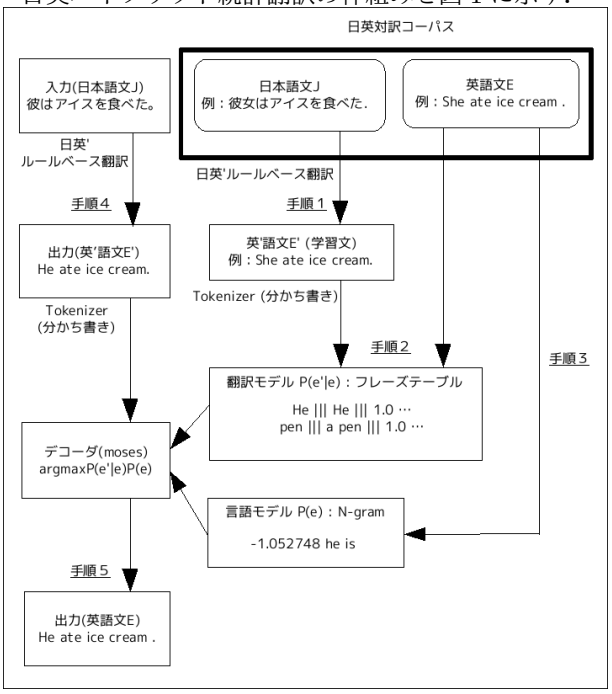

図 1　日英ハイブリッド統計翻訳の枠組

3.3　ルールベース翻訳：Rule

ルールベース翻訳は，人手で変換規則を構築して翻訳を行うシステムである．本研究では，最新の試作ルールベース翻訳システムを用いる．

3.4　階層句に基づく統計翻訳：Hierarchy

階層句を用いて翻訳を行う統計翻訳システムである．本研究では，“moses” を用いる．

3.4.1　各システムの翻訳例

本研究では翻訳システムの記述書式を表 3 に示す．そして各システムの翻訳例を表 4 に示す．

表 3　翻訳システムの記述書式

ルールベース翻訳	Rule
ハイブリッド統計翻訳	Hybrid
句に基づく統計翻訳	Phrase
階層句に基づく統計翻訳	Hierarchy

表 4　各システムの翻訳例

入力文	電気 コンロ の コイル が 焼き 切れ た 。
Rule	The coil of the electric cooker was able to be burned off .
Hybrid	The company of the electric cooker was burned out .
Phrase	The The buckets or of electricity .
Hierarchy	The electric The of the cooking stove .
参照文	The heater coil is burnt out .

表 4 の例では，Rule(ルールベース翻訳) が最も良い翻訳である．

3.5　評価方法

本研究では，自動評価法として BLEU[1]，NIST[1]，METEOR[10]，IMPACT[11]，RIBES[12]，TER[13] および WER[13] を用いる．また，人手評価として対比較評価を用いる．

BLEU，NIST は N-gram 適合率により評価を行う．METEOR は単語の出現頻度により評価を行う．IMPACT は，名詞句のチャンクを用いて，参照文との類似性を求めることで評価を行う．RIBES は，単語の出現順序を順位相関係数を用いて評価を行う．これらの自動評価法は値が大きい方が良い評価である．それに対して，MTER，WER は Error Rate であるので，値が小さい方が良い評価である．

3.6　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 5 のように用いる．

表 5　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[9]”を用いて，形態素解析を行う．また，英語文に対して，“tokenizer.perl[7]”を用いて，分かち書きを行う．表 6 に単文コーパスの例を示す．

表 6　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 .

4　実験結果

4.1　自動評価

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，ハイブリッド統計翻訳，ルールベース翻訳および階層句に基づく統計翻訳を用いる．それぞれの自動評価の結果を表 7 に示す．

表 7　自動評価結果

	Rule	Hybrid	Phrase	Hierarchy
BLEU	0.1320	**0.1798**	0.1341	0.1352
NIST	4.8260	**5.5426**	4.9239	4.9628
METEOR	0.4724	**0.7601**	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 7 の結果より，すべての自動評価において，ハイブリッド統計翻訳が最良の値 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．

4.2 人手評価

本研究では，ルールベース翻訳に対して，その他 3 つの翻訳システム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，ハイブリッド統計翻訳，句に基づく統計翻訳および階層句に基づく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 8 に以下に示す．さらに人手評価の結果を表 9 に示す．

表 8 評価基準

Rule ○	ルールベース翻訳の方が優れている
Hybrid ○	ハイブリッド統計翻訳が ルールベース翻訳より優れている
Phrase ○	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
Hierarchy ○	階層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
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 9 人手評価結果

Rule ○	Hybrid ○	差なし	同一出力
23	5	59	13
Rule ○	Phrase ○	差なし	同一出力
34	3	63	1
Rule ○	Hierarchy ○	差なし	同一出力
30	3	66	1

表 9 の結果より，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．

4.3 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 7 と人手評価の表 9 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句に基づく統計翻訳において，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．

4.4 自動評価と人手評価に差異がある翻訳例

表 10 に，ハイブリッド統計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 Phrase と Hierarchy についても表 11，表 12 に示す．なお，() 内は BLEU の値を示す．

表 10 Rule と Hybrid の対比較評価例

入力文	両者 の 間 に 商談 が 成立 し た 。
Rule (0.000)	The business talk was materialized among both .
Hybrid (0.3564)	The concluded negotiations between the two .
参照文	A bargain was arranged between the two .
人手評価	Rule ○

表 11 Rule と Phrase 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 。
Rule(0.4317)	Her long hair was wavy to the wind .
Phrase(0.4468)	Her long hair in the wind .
参照文	Her long hair was streaming in the wind .
人手評価	Rule ○

表 12 Rule と Hierarchy 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 。
Rule (0.0000)	This is a magazine which all families can read interestingly .
Hierarchy (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	Rule ○

5 考察

5.1 自動評価と人手評価の差異の原因

5.1.1 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 13 に示す．

表 13 未知語の含む文の数

Rule	Hybrid	Phrase	Hierarcy
307	433	3079	2889

表 13 より，Phrase と Hierarchy 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，Rule と Hybrid 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，Rule と Hybrid 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．

5.1.2 単語の重要度の違いの問題

日本語文において，文の意味は，各単語が連鎖することで表現されている．たが，助詞は文の意味に与える影響が少ない．それに対し，動詞は，文の意味に与える影響が大きい．自動評価では，各単語は同じ割合で評価されている．そのため，動詞と助詞どちらかが誤った場合でも評価値は変わらない．しかし人手評価においては，動詞が誤った場合と助詞が誤った場合を比較すると，動詞が誤った場合の方が文質は低下する．つまり，動詞が誤った場合に，人手評価が悪くなってしまう．よって自動評価と人手評価の結果に差異があると考えている．

5.2 折り返し翻訳を利用した評価手法の提案

5.2.1 目的

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 [14] を利用した評価手法を提案する．

5.2.2 折り返し翻訳を利用した評価手法の手順

折り返し翻訳を用いて，日英翻訳におけるハイブリッド統計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 ハイブリッド統計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．なお，英日翻訳としてルールベース翻訳 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．

5.2.3 実験結果

テスト文を用いて，折り返し翻訳を利用した評価を行う．なお，日英翻訳にはハイブリッド統計翻訳と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 14 に示す．表 14 の日英 Rule は，日英翻訳にルールベース翻訳を用いた場合を示す．また，日英 Hybrid は，日英翻訳にハイブリッド統計翻訳を用いた場合を示す．

表 14 日本語文が完全一致した文数

日英 Rule	日英 Hybrid
130	89

表 14 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 9 の人手評価と同様の結果となることが示された．

5.2.4 折り返し翻訳を利用した評価の例

折り返し翻訳が成功した例を表 15 に，失敗した例を表 16 に示す．

表 15 折り返し翻訳の成功例

日本語文	彼女は長期休暇をとる。
英語文 (日英 Rule)	She takes a long leave.
日本語文 (英日 Rule)	彼女は長期休暇をとる。
日本語文	私は山に登った。
英語文 (日英 Hybrid)	I climbed the mountain.
日本語文 (英日 Rule)	私は山に登った。

表 16 折り返し翻訳の失敗例

日本語文	私は家の外に出た。
英語文 (日英 Rule)	I went out of the house.
日本語文 (英日 Rule)	私は家から出ていった。
日本語文	もっと右へ寄ってください。
英語文 (日英 Hybrid)	Please come visit to the right .
日本語文 (英日 Rule)	右へ遊びに来てください。

5.2.5 折り返し翻訳を利用した評価と人手評価

折り返し翻訳を利用した評価では，人手評価と同様の結果が得られた．よって，折り返し翻訳を利用した提案手法の有効性が示された．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，130 文と 89 文しか存在しないことが挙げられる．さらに，5.2.2 節の手順 3 において，英日翻訳に句に基づく統計翻訳を用いた実験も行った．しかし人手評価との間に差異が生じる結果となった．よって，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．

6 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．その結果，自動評価と人手評価に差異が生じた．したがって，すべての自動評価法に問題があると考えている．

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．

参考文献

[1] George Doddington, “Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics”,Proceeding HLT ’02 Proceedings of the second international conference on Human Language Technology Research 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 特許翻訳研究会 報告書, pp.2-13, March, 2009.

[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] 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, 9(4), pp. 569-579, 1994.

[7] 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.

[8] SRILM(The SRI Language Modeling Toolkit) : srilm.tgz http://www.speech.sri.com/projects/srilm/

[9] MeCab : http://mecab.sourceforge.net/

[10] METEOR, The METEOR Automatic Machine Translation Evaluation System http://www-2.cs.cmu.edu/~alavie/METEOR/

[11] 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 (ACL 2010), pp.108-117, 2010.

[12] RIBES: Rank-based Intuitive Bilingual Evaluation Measure http://www.kecl.ntt.co.jp/icl/lirg/ribes/

[13] Richard Schwartz, Linnea Micciulla, John Makhoul. “A Study of Translation Edit Rate with Targeted Human Annotation”, AMTA, pp. 223-231, 2006.

[14] Harold Somers, “Round-Trip Translation:What Is It Good For?”, Proceedings of the Australasian Language Technology , pp.127-133, 2005.