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

† 東江 恵介　　†† 出羽 達也　　† 村上 仁一

† 鳥取大学大学院 工学研究科 情報エレクトロニクス専攻
†† 東芝 研究開発センター 知識メディアラボラトリー

† s062041@ike.tottori-u.ac.jp , murakami@tottori-u.ac.jp
†† tatsuya.izuha@toshiba.co.jp

1　はじめに

現在，機械翻訳の分野において，対訳データから自動的に翻訳規則を生成し翻訳を行う，統計翻訳が注目されている．また，ルールベース翻訳と統計翻訳を組み合わせて翻訳を行う，ハイブリッド型機械翻訳も盛んに行われている [1][2].

しかし，その評価方法は自動評価で行う場合が多い．また，人手評価を行う場合，多くのハイブリッド翻訳の研究では，統計翻訳とハイブリッド翻訳が評価対象であり，ルールベース翻訳とハイブリッド型機械翻訳を評価対象とした，ハイブリッド翻訳の研究は少ない．

そこで，本調査では，ルールベース翻訳とハイブリッド翻訳の人手評価と自動評価を行い，翻訳精度を調査する．

2　日英ハイブリッド翻訳システム

本調査で用いるハイブリッド翻訳システムはルールベース翻訳と統計翻訳を組み合わせる．ルールベース翻訳には東芝 Tauras[3] を用い，統計翻訳には Moses[4] を用いる．本調査で行う翻訳の手順を図 1 に示す．

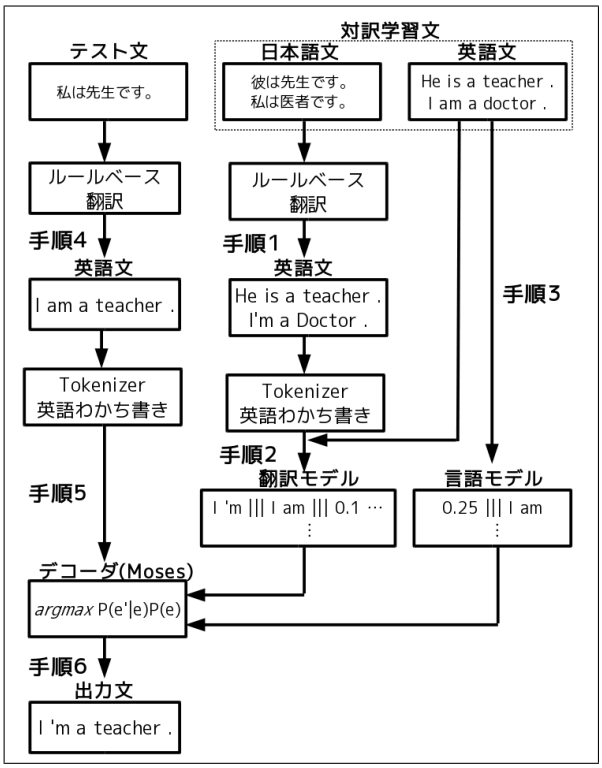

図 1　ハイブリッド翻訳の手順

手順 1　ルールベース翻訳を用いて，対訳学習文の日本語文を英語文に翻訳する．

手順 2　手順 1 で翻訳した英語文と対訳学習文の英語文を用いて翻訳モデルを作成する．

手順 3　対訳学習文の英語文を用いて言語モデルを作成する．

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 で翻訳した英語文をデコーダ (Moses) に入力する．

手順 6　手順 2 で作成した翻訳モデルと手順 3 で作成した言語モデルを用いて最終的な出力文の選出を行う．

3　実験データ

本調査に用いる実験データは，英語文に対しては，tokenizer.perl を用いて，英語文の分かち書きを行う．

3.1　学習データ

単文の学習データには，辞書の例文から抽出した単文 100,000 文対を用いる．重文複文の学習データには，辞書の例文から重文複文 100,000 文対を用いる [5].

表 1　単文の学習データの例

私 は 映画 を 見 に 行く 。
I go to see a movie .

表 2　重文複文の学習データの例

首 を 曲げ て 考え て おっ た 。
He was thinking with his head on one side .

3.2　テストデータ

単文の日英翻訳のテストデータには，辞書の例文から抽出した単文 10,000 文対を用いる．重文複文の日英翻訳のテストデータには，辞書の例文から重文複文 10,000 文対を用いる．

3.3　Development データ

単文の Development データには，辞書から抽出した単文 1,000 文対を用いる．重文複文の Development データには，辞書から抽出した重文複文 1,000 文対を用いる．

4　実験環境

4.1　フレーズテーブル

本調査における，フレーズテーブルの作成には，train-factored-phrase-model.perl を用いる．

4.2 *N*-gram モデルの学習

言語モデルには，*N*-gram モデルを用いる．*N*-gram モデルの学習には，“SRILM”[6] を用いる．本調査では，5-gram モデルを用いる．なお，スムージングには kndiscount を用いる．

4.3 デコーダのパラメータ

本調査の翻訳実験では，パラメータの最適化 [7] を行う．

5 評価方法

5.1 人手評価

本調査では，ハイブリッド翻訳とルールベース翻訳の対比較評価を行う．対比較評価はルールベース翻訳の翻訳結果とハイブリッド翻訳の翻訳結果からそれぞれランダムに 100 文ずつ抽出し，どちらが優れているかを判断する．固有名詞の未知語はローマ字変換し，それ以外の未知語は存在しないものとして評価を行う．判断基準を以下に示す．

Hybrid ○ ハイブリッド翻訳の結果がルールベース翻訳の結果より優れている場合

RBMT ○ ルールベース翻訳の結果がハイブリッド翻訳の結果より優れている場合

差なし ハイブリッド翻訳の結果とルールベース翻訳の結果の文質が変わらない場合

同一出力 ハイブリッド翻訳の結果とルールベース翻訳の結果が同じ場合

5.2 自動評価

本調査では，出力文の評価として BLEU[8]，NIST[9]，METEOR[10]，TER[11]，WER を用いる．

6 翻訳実験

6.1 ルールベース翻訳

ルールベース翻訳は，入力文に日本語のテスト文を用いて翻訳を行う．

6.2 ハイブリッド翻訳

ハイブリッド翻訳の統計翻訳に用いる，学習データ，Development データ，テストデータは全て日本語文をルールベース翻訳で英語文に翻訳したデータを用いる．

7 実験結果

7.1 人手評価結果

表 3 に人手評価の結果を示す．

表 3 ハイブリッド翻訳とルールベース翻訳の自動評価結果

	Hybrid ○	RBMT ○	差なし	同一出力
単文	14	34	42	10
重文複文	7	57	35	1

表 3 から，人手評価において，ルールベース翻訳はハイブリッド翻訳より翻訳精度が高い．

7.1.1 ハイブリッド翻訳が優れていると判断した例

表 4，表 5 中の “RBMT” はルールベース翻訳の出力を示し，“Hybrid” はハイブリッド翻訳の出力を示す．

表 4 において，ハイブリッド翻訳の出力は，ルールベース翻訳より主語が適切であるので，ハイブリッド翻訳が優れていると判断した．

表 4 単文においてハイブリッド翻訳が優れていると判断した例

入力文	毎年 ダービー へ 行く 。
参照文	I go to the Derby every year .
RBMT	It goes to the Derby every year .
Hybrid	I ’m going to the Derby every year .

表 5 において，ハイブリッド翻訳の出力は入力文の意味に沿うが，ルールベース翻訳の出力では，入力文の意味に沿わないため，ハイブリッド翻訳が優れていると判断した．

表 5 重文複文においてハイブリッド翻訳が優れていると判断した例

入力文	波 が 岩 に 当たっ て 砕ける 。
参照文	The waves break on the rocks .
RBMT	A wave gives it a shot at a rock .
Hybrid	The waves break on the rock .

7.1.2 ルールベース翻訳が優れていると判断した例

表 6，表 7 中の “RBMT” はルールベース翻訳の出力を示し，“Hybrid” はハイブリッド翻訳の出力を示す．

表 6 において，ルールベース翻訳の出力の “remade” は “作り替えた” の意味に沿うが，ハイブリッド翻訳の出力の “made” では “作り替えた” の意味に沿わないため，ルールベース翻訳が優れていると判断した．

表 6 単文においてルールベース翻訳が優れていると判断した例

入力文	父 は 犬 小屋 を 大きく 作り替え た 。
参照文	Father rebuilt the doghouse to make it bigger .
RBMT	The father remade the doghouse greatly .
Hybrid	My father made the doghouse .

表 7 において，ルールベース翻訳の出力では入力文の意味に沿うが，ハイブリッド翻訳の出力は，入力文の “なる” に対応する語句がなく，ハイブリッド翻訳出力の “throw” に対応する語句が入力文にないため，ルールベース翻訳が優れていると判断した．

表 7 重文複文においてルールベース翻訳が優れていると判断した例

入力文	その 山 は 秋 に なる と ハイカー たち で にぎわう 。
参照文	The hill is thronged with groups of hikers every fall.
RBMT	If the mountain becomes autumn, it will be crowded with hikers.
Hybrid	If the peak fall , it is crowded with hikers throw.

7.2 自動評価結果

表 8，表 9 に自動評価における結果を示す．表中の “Moses” は統計翻訳のスコアを示し，“RBMT” はルールベース翻訳のスコアを示し，“Hybrid” はハイブリッド翻訳のスコアを示す．

表 8 単文における翻訳精度の評価

	BLEU	NIST	METEOR	TER	WER
Moses	0.1420	4.9470	0.3886	0.7068	0.7342
RBMT	0.1329	4.8524	0.4007	0.7240	0.7471
Hybrid	0.1774	5.4575	0.4353	0.6706	0.6947

表 9 重文複文における翻訳精度の評価

	BLEU	NIST	METEOR	TER	WER
Moses	0.1191	4.4831	0.3542	0.7784	0.8135
RBMT	0.0942	4.0271	0.3635	0.8518	0.8838
Hybrid	0.1481	4.9317	0.3921	0.7388	0.7698

表 8, 表 9 より，全ての自動評価において，ハイブリッド翻訳はルールベース翻訳よりスコアが高い．

8 考察

7 章の実験結果より，人手評価において，ルールベース翻訳はハイブリッド翻訳より翻訳精度が高いという結果であった．しかし，自動評価において，ハイブリッド翻訳はルールベース翻訳より高いスコアであった．例を表 10 に示す．なお，表 10 中の “RBMT” はルールベース翻訳の出力を示し，“Hybrid” はハイブリッド翻訳の出力を示す．また，“RBMT BLEU”，“Hybrid BLEU” はルールベース翻訳とハイブリッド翻訳の 1 文に対するそれぞれの BLEU スコアを示す．

表 10 自動評価と人手評価の評価が逆転した例

入力文	その 機械 の 構造 に は 欠陥 が ある 。
参照文	There is a fault in the machine ’s construction .
RBMT	The structure of the machine has a defect .
Hybrid	The structure of the is a fault in the machine .
RBMT BLEU	0.0000
Hybrid BLEU	0.4799

表 10 において，人手評価では，ルールベース翻訳は意味がわかる文であるが，ハイブリッド翻訳は意味がわからない文である．一方，自動評価では，ハイブリッド翻訳がルールベース翻訳よりも BLEU スコアが良い．

その原因としては，ハイブリッド翻訳において，動詞の位置が不適切であるため，文構造が崩れ，結果として，翻訳精度の低下を引き起こすのではないかと考えている．

9 学習データが増加した場合の翻訳実験

7 章の実験結果より人手評価において，ハイブリッド翻訳はルールベース翻訳より翻訳品質が悪いという結果であった．そこで，学習データを増加させた場合について，翻訳実験を行う．

9.1 実験データ

実験データに対する処理は 6 章の実験と同様である．

学習データ

単文と重文複文の学習データを統合した 281,707 文対を用いる．

テストデータ

3.2 節と同様のテストデータを用いる．

Development データ

3.3 節と同様の development データを用いる．

9.2 実験環境

4 章と同様の実験環境で，実験を行う．

10 学習データが増加した場合の実験結果

10.1 人手評価結果

表 11 に人手評価における結果を示す．

表 11 ハイブリッドとルールベース翻訳の自動評価結果

	Hybrid ○	RBMT ○	差なし	同一出力
単文	16	30	46	8
重文複文	12	22	65	1

表 11 から，人手評価において，学習データを増加した場合でも，ルールベース翻訳はハイブリッド翻訳より翻訳精度が高い．

10.1.1 ハイブリッド翻訳が優れていると判断した例

表 12，表 13 中の “RBMT” はルールベース翻訳の出力を示し，“Hybrid” はハイブリッド翻訳の出力を示す．

表 12 において，ハイブリッド翻訳の出力は入力文の意味に沿うが，ルールベース翻訳の出力では，“のどに骨が立つ” という意味になるため，不適切である．よって，ハイブリッド翻訳が優れていると判断した．

表 12 単文においてハイブリッド翻訳が優れていると判断した例

入力文	のど に 魚 の 骨 を 立て て しまっ た 。
参照文	I had a fish bone stuck in my throat .
RBMT	The bone of a fish has been stood to the throat .
Hybrid	A fish bone got stuck in my throat .

表 13 において，ルールベース翻訳出力の “rather from” では “むしろ” という意味にはならならず，入力文の意味とは異なる．よって，ハイブリッド翻訳が優れていると判断した．

表 13 重文複文においてハイブリッド翻訳が優れていると判断した例

入力文	その 言葉 は 慣習 的 に 男 より むしろ 女 について 言う 場合 に 使わ れ て いる 。
参照文	The word is customarily used of women rather than men.
RBMT	The language is used when saying about a woman rather from a man as usual .
Hybrid	The word is used in about women rather than a man as usual .

10.1.2 ルールベース翻訳が優れていると判断した例

表 14，表 15 中の “RBMT” はルールベース翻訳の出力を示し，“Hybrid” はハイブリッド翻訳の出力を示す．

表 14 において，ルールベース翻訳の出力は入力文の意味に沿うが，ハイブリッド翻訳の出力では，動詞が 2 つあるため文として不適切である．よって，ルールベース翻訳が優れていると判断した．

表 14 単文においてルールベース翻訳が優れていると判断した例

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

表 15 において，ハイブリッド翻訳の出力は，入力文の “広くなる” に対応する語句がないため，ルールベース翻訳が優れていると判断した．

表 15 重文複文においてルールベース翻訳が優れていると判断した例

入力文	道幅 が 2 メートル 伸び て 広く なる 。
参照文	The road is to be widened by two metres .
RBMT	Road width is extended 2 meters and becomes large .
Hybrid	The width extended two meters .

10.2 自動評価結果

表 16，表 17 に自動評価における結果を示す．表中の “Moses” は統計翻訳のスコアを示し，“RBMT” はルールベース翻訳のスコアを示し，“Hybrid” はハイブリッド翻訳のスコアを示す．

表 16 単文における翻訳精度の評価

	BLEU	NIST	METEOR	TER	WER
Moses	0.2054	5.8628	0.4523	0.6956	0.7195
RBMT	0.1329	4.8524	0.4007	0.7240	0.7471
Hybrid	0.2186	5.8951	0.4647	0.6447	0.6693

表 17 重文複文における翻訳精度の評価

	BLEU	NIST	METEOR	TER	WER
Moses	0.1729	5.3296	0.4108	0.8204	0.8544
RBMT	0.0942	4.0271	0.3635	0.8518	0.8838
Hybrid	0.1902	5.4848	0.4281	0.7037	0.7362

表 16，表 17 より，全ての自動評価において，ハイブリッド翻訳はルールベース翻訳よりスコアが高い．

11 考察

実験結果より，学習量を増加した実験において，人手評価では，ルールベース翻訳がハイブリッド翻訳より，良い翻訳であった．この結果から，現状では人手評価に関して，ルールベース翻訳が最も良いと考えられる．また，学習量を増加させても，人手評価において，ルールベース翻訳はハイブリッド翻訳より翻訳精度が高いという結果であった．

今回の実験に関しては，学習量の増加に対する変化を見るという視点では，追加する学習データの量が少ないと考えられる．よって，より明確な変化を見るためには，さらに学習量を増加させて再実験する必要があると考えている．

12 おわりに

本調査では，ルールベース翻訳とハイブリッド翻訳の人手評価を行った．その結果，人手評価では，ルールベース翻訳はハイブリッド翻訳より翻訳精度が高く，自動評価では，ハイブリッド翻訳はルールベース翻訳より自動評価のスコアが高いという結果であった．

自動評価では，ハイブリッド翻訳とルールベース翻訳のスコアの差が大きいにもかかわらず，人手評価ではルールベース翻訳が翻訳精度が高いという結果であった．その原因としては，動詞の位置が不適切であるため，文構造が崩れたことが原因であると考えられる．また，この評価の不一致は，現在の自動評価の問題点を示していると考えている．

13 謝辞

この調査を行う際に，様々なご支援，ご教授を賜わりました，東芝研究開発センター知識メディアラボラトリーの翻訳チームの皆さまに深く御礼を申し上げます．

参考文献

[1] L.Dugast, J.Senellart, and P.Koehn, “Statistical postediting on SYSTRAN’s rule-based translation system” , in Second Workshop on SMT, 2007, pages.179-182

[2] 福田智大, 村上仁一, 徳久雅人, 池原悟, “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, pp.672-675, 2010.

[3] 東芝ルールベース翻訳システム “Tauras” http://www.mt-archive.info/Nagao-1989-Amano.pdf

[4] Philipp Koehn et al., “Moses: Open Source Toolkit for Statistical Machine Translation” , Association for Computational Linguistic, pp.177-180, (2007).

[5] 西山七絵, 村上仁一, 徳久雅人, 池原悟, “単文文型パターン辞書の構築” , 言語処理学会第 11 回年次大会, pp.372-375, 2005.

[6] SRILM The SRI Language Model Toolkit http://www.speech.sri.com/projects/srilm

[7] Franz Josef Och “Minimum Error Rate Training in Statistical Machine Translation” , Association for Computational Linguistics, pp.160-167, (2003).

[8] Kishore Papineni, Salim Rukos, Todd Ward, and Wei-Jing Zhu , “BLEU: a Method for Automatic Evaluation of Machine Translation” , Association for Computational Linguistics, pp.311-318, (2002).

[9] George Doddington, “Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics”, Human Language Technology, (2002).

[10] Satanjeev Banerjee and Alon Lavi, “METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments” , Association for Computational Linguistics, pp.65-72, (2005).

[11] Matthew Snover and Bonnie Dorr et al., “A Study of Translation Edit Rate with Targeted Human Annotation”, The Association for Machine Translation in the Americas, pp.223-231, (2006)