3B1-2

機械翻訳における文一致率による評価

A New Evaluation Method for Machine Translation using Sentence Correspondence Rate

村上　仁一[*1]　岡崎 響[*1]　石原 雅文[*1]
Jin'ichi Murakami　Hibiki Okazaki　Masafumi Ishihara

[*1]鳥取大学 工学部
Faculty of Engineering Tottori University

Currently, there are many evaluation methods for machine translation. These methods are divided automatic evaluation and human evaluation. Human evaluation has much costs. So, automatic evaluation methods are mainly used. However, there are many difference between human evaluation and automatic evaluation.

In this paper, we proposed a new automatic evaluation method. We used sentence correspondence rate between references and the outputs of translation systems. For the results of experiments, our method was effective and many new kind of aspects were obtained.

1. はじめに

現在，機械翻訳システムの性能評価において，複数の評価手法が提案されている．これらの翻訳結果の評価手法は，大きくわけて人手評価と自動評価がある．

人手評価は，一般的には翻訳された出力文を，文全体の意味や文法の正しさを考慮して人間が評価する．したがって信頼度は比較的高い．しかし，人が判断するため高コストである．また評価者によって着目点が異なるため，評価が同一にならないことが多い．一方自動評価にも多くの種類が提案されている．多くの評価方法は，人手で作成した入力文の正解文を必要とし，翻訳された出力文と正解文の単語の順序や出現頻度から評価する．自動評価は，人手で作成した入力文の正解文が必要だが，人手評価と比較するとコストは低い．しかし，人手評価と自動評価には大きな差が生じることが報告されている [1].

松本ら [1] は，人手評価と自動評価に差がある原因として，以下の原因を報告をしている．人手評価は，人間が文全体を着目して評価する．したがって，単語単位では均一に評価しない．一方，自動評価は，出力文と正解文の単語の語順や出現頻度を比較する．そのため，単語単位では均一に評価する．ところで，動詞は文を構成する重要な要素である．そのため，動詞が抜け落ちてると，意味不明な文章になる．しかし，自動評価において，動詞以外の大部分の単語が一致する場合，高い評価をする．よって人手評価と自動評価に大きな差が生じる．

そこで本研究では，単語ではなく文全体を評価する新たな自動評価法を提案する．具体的には出力文と正解文において，文を構成する単語が完全に一致した文数で評価を行う．単語が完全に一致した文を数えることで，文全体を考慮した評価が可能であると仮定する．最後に，提案手法と人手評価の相関を調査する．

2. 自動評価と人手評価が大きく異なる例

表 1 に自動評価と人手評価が大きく異なる例文を示す．本節では日英翻訳を想定する．表中の出力文は，入力文の機械翻訳の結果である．表中の正解文は，入力文の人手による翻訳である．

連絡先: 村上仁一，鳥取大学工学部，鳥取市湖山町南 4-101，0857-31-6788，0857-31-6787，murakami@ike.tottori-u.ac.jp

表 1: 自動評価と人手評価が大きく異なる例文

入力文	ホラー映画を見るのは楽しくありません。
出力文	This is interesting to see the horror movie .
正解文	This is not interesting to watch the horror movie .

この例では，人間の評価では，出力文の意味は入力文と逆になるため，低くなる．しかし，自動評価では，出力文と正解文において多くの単語が一致しているため，高くなる．

3. 提案手法 (文一致数)

本研究では，入力文の人手による翻訳文を正解文とし，”正解文と出力文の単語が完全一致する文数”（以下，文一致数）で，翻訳結果を評価する．以下に具体的な手順を示す．

手順１　入力文 10000 文を翻訳して出力文を得る．

手順２　入力文の正解文と，手順１の出力文を比較する．

手順３　文一致数を調査する．

4. 実験

提案した自動評価方法の信頼性を調査するために翻訳実験を行う．実験は，日英翻訳と英日翻訳の２種類を用いる．以下に実験の概要を示す．

4.1 コーパス

本研究では，辞書から抽出した実験には単文コーパスと重文複文コーパスの２種類を利用する [4]．日本文は，chasen を使って形態素解析を行う．表 2 に，利用したコーパスの文数を示す．また表 3 に使用した例文を示す．

表 2: 使用したコーパスの文数

学習データ	100,000 文
ディベロップメントデータ	1,000 文
入力文（テスト文）	10,000 文

表 3: 使用したコーパスの例文

単文	
入力文 1	銀行 は ちょうど 駅 の 向かい に ある 。
正解文 1	The bank is just across from the station .
入力文 2	火 は 台所 から 出 た 。
正解文 2	The fire started in the kitchen .
入力文 3	学校 は 4月 から 始まり ます 。
正解文 3	School starts in April .
重文複文	
入力文 1	彼 は 言う ばかり で 行動 が 伴わ ない 。
正解文 1	He is all talk and no action .
入力文 2	あれ が 彼 の 住ん で い た 家 で ある 。
正解文 2	That is the house in which he lived .
入力文 3	彼 は 泳げる と 言っ た 。
正解文 3	He said he could swim .

4.2 翻訳システム

本研究で使用する翻訳システムを，以下に示す．

1. ルールベース翻訳

   ルールベース翻訳とは，人手によって構成された変換規則を元に翻訳を行うシステムである．現地点では，日英・英日翻訳において，統計翻訳より翻訳精度が高いことが知られている．本研究では，ルールベース翻訳に東芝の Tauras[8] と富士通の Atlas[3] を使用する．

2. 句に基づく統計翻訳

   統計翻訳とは，対訳コーパスより翻訳規則を生成し，翻訳を行うシステムである．入力文が与えられた時，デコーダで翻訳モデルと言語モデルの確率を組み合わせ，生成確率が最大となる文を求めて翻訳を行う翻訳システムである．統計翻訳には，単語に基づく統計翻訳，句に基づく統計翻訳，階層型統計翻訳などがある．現在，句に基づく統計翻訳は，対訳句を用いて翻訳を行う方法で，翻訳方式のなかで最も良く利用されている．本実験では，句に基づく統計翻訳として moses[5] を用いる．また，パラメータチューニングを行う．

3. 階層型統計翻訳

   階層型統計翻訳 [9] は，木構造を用いて翻訳を行う．具体的には SCFG に基づいて翻訳を行う．そのため，階層型統計翻訳は句に基づく統計翻訳よりも，人手の評価において翻訳精度が高くなることが知られている [10]．本実験では階層型統計翻訳として moses[5] を用いる．

4. ハイブリッド翻訳

   ハイブリッド翻訳 [11] とは，前処理としてルールベース翻訳を，後処理として統計翻訳を用いる．ハイブリッド翻訳は，自動評価において最も翻訳精度が高くなることが知られている．以下に手順を示す．

   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．

   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．

   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．翻訳モデルと言語モデルは，手順２と手順３で作成されたものを使用する．

   本研究では，後処理としての統計翻訳として，句に基づく統計翻訳と，階層型統計翻訳の２種類を利用する．

本研究では計８種類の翻訳システム利用した．使用した翻訳システムを表 4 に略記と共に示す．

表 4: 本研究で使用する翻訳システム

	翻訳システム	略記
1	ルールベース翻訳 (Tauras)	RBMT(t)
2	ルールベース翻訳 (Atlas)	RBMT(a)
3	句に基づく統計翻訳 (Moses)	PSMT
4	階層型統計翻訳 (Moses)	HSMT
5	前処理:ルールベース翻訳 +後処理:句に基づく統計翻訳 (Tauras+Moses)	RBMT(t)+PSMT
6	前処理:ルールベース翻訳 +後処理:句に基づく統計翻訳 (Atlas+Moses)	RBMT(a)+PSMT
7	前処理:ルールベース翻訳 +後処理:階層型統計翻訳 (Tauras+Moses)	RBMT(t)+HSMT
8	前処理:ルールベース翻訳 +後処理:階層型統計翻訳 (Atlas+Moses)	RBMT(a)+HSMT

4.3 評価方法

4.3.1 人手評価

人手評価は，各翻訳システムの 10,000 文の出力文から，ランダムに 100 文を抽出する．そして，１文ずつ 1(悪い翻訳)〜5(良い翻訳) でランクづけし，平均値を算出する (adequacy)．評価者は 1 名である．

4.3.2 自動評価法

本論文では比較のため，提案手法の他に自動評価法として BLEU，NIST，METEOR，RIBES[7] を用いる．

5. 実験結果

5.1 出力例

表 5 に文一致した例と文一致しなかった例を示す．

表 5: 出力例文

文一致した例	
入力文	犬 が 私 の 手 を なめ た 。
正解文	The dog licked my hand .
出力文	The dog licked my hand .
文一致しなかった例	
入力文	酒 は 米 から 作ら れる 。
正解文	Sake is made from rice .
出力文	The Sake is made from the United States .

5.2 人手評価との比較および他の自動評価との比較

本実験では，日英翻訳と英日翻訳，単文データと重文複文データ，全 8 種類の翻訳システム，合計 32 種類の実験を行った．実験結果を表 7 にまとめる．また，人手評価との相関係数と，提案手法との相関係数も同時に示す．この表から読み取れることを以下に示す．

1. 提案手法の有効性

   提案手法を含めた 5 種類の自動評価と，人手評価との相関を見ると，日英翻訳の単文の実験において，METEOR と RIBES は提案手法より高い．しかし，他の全ての実験において，提案手法が最も優れている．したがって，提案手法の有効性が示された．また，日英翻訳と英日翻訳を比較すると，英日翻訳において有効性が高い．

2. 提案手法と他の自動評価方法との相関

   提案手法と他の自動評価方法との相関をみると，BLUE との相関が比較的高い．

3. 人手評価の傾向

   日英翻訳と英日翻訳，単文データと重文複文データの 4 種類の表から人手評価において以下の傾向がある．

   - RBMT は最も高い．
   - PSMT が最も低い．
   - 自動評価方法の精度は，以下の順番になる．
     $RBMT > RBMT + PSMT = RBMT + HSMT > HSMT > PSMT$

4. 英日翻訳の単文の自動評価方法の問題

   日英翻訳の重文複文の実験において，全ての自動評価方法において人手評価との相関が負になった．ただし，提案手法は-0.04 と他の自動評価方法と比較して最も低い値になった．

6. 考察

6.1 提案手法の有効性

今回の実験において提案手法の有効性が示された．しかし，実験の評価データが少ない．また人手評価も 1 名である．そして，英日翻訳の単文の実験では，人手評価と負の相関係数になった．したがって実験における信頼性がまだ低いと考えている．

今後，より多くのシステムで多くの人手評価を行い，提案手法の信頼性の向上を図りたい．

6.2 提案手法の問題点

以下に提案手法において問題になった文を示す．

表 6: 提案手法の問題点

入力文	彼の妹はとてもかわいい。
出力文	His younger sister is so cute .
正解文	His younger sister is so pretty .

この例では cute が pretty になっているため，文は一致しなかった．このような例を解決する方法として以下の方法を考えている．

1. 複数の正解文 (Multi-Reference)

   今回の実験では，正解文は 1 文である．しかし通常，翻訳において正解文は複数ある．複数の正解文を利用することにより，この問題を緩和できると思われる．ただし，複数正解文を作成することは，コストがかかる．

2. 複数の出力文の候補文 (N-best)

   今回の出力文は 1 文であった．そこで出力文を複数にすることにより，この問題を緩和できると考えている．

7. おわりに

従来の自動評価は，文全体に対して評価を行わず，文に存在する部分的な単語に対して評価を行う．結果，動詞のような重要な単語が抜け落ちた翻訳結果でも，大部分の単語が正解文と一致していれば，高いスコアを算出してしまう問題がある．そこで本研究では，正解文と出力文の単語が完全に一致する文数を，新たな自動評価法として提案した．結果として，提案手法は人手評価と高い相関が得られた．そして，他の自動評価方法より高かった．これは，提案手法の有効性を示している．しかし，実験の量が少ないため，信頼性がまだ低いと考えている．今後は，より多くの実験をおこないたい．また複数の正解文を利用した場合や，複数の出力文を利用した場合の提案方法を調査し，人手評価と比較を行いたい．

参考文献

[1] 松本 拓也, ” 機械翻訳における自動評価と人手評価の考察”, 平成 23 年度卒業論文, 2011.

[2] AAMT/Japio 特許翻訳研究会, ” 自動評価手法がもたらした歓喜と失望，そして，希望”, 2012.

[3] 日英翻訳ソフト ATLAS, http://software.fujitsu.com/jp/atlas/

[4] 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，pp.119-130, Mar. 2012.

[5] Philipp Koehn, Marcello Federico, Brooke Cowan, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, Evan Herbst, “Moses, Open Source Toolkit for Statistical Machine Translation”, Proceedings of the ACL 2007 Demo and Poster Sessions, pages 177–180, June 2007.

[6] Franz Josef Och, “Minimum Error Rate Training in Statistical Machine Translation”, In Proceeding of the 41st Annual Meeting of the Association for Computational Linguistics, pp.160–167, 2003.

[7] 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 年次大会発表論文集, pp.1111–1114, 2011.

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

[9] Chiang David, “A Hierarchical Phrase-Based Model for Statistical Machine Translation.”, Proceedings of ACL-2005, pp.263-270, 2005.

[10] 久保田裕介, 村上仁一, 徳久雅人, “階層型統計翻訳の調査”, 言語処理学会第 18 回年次大会, pp.259-262, 2012.

[11] 村上仁一, 徳久雅人, “ルールベース翻訳と統計翻訳を結合した特許翻訳”, AAMT/Japio 特許翻訳研究会　第 1 回特許情報シンポジウム, pp.46-53, Dec. 2010.

表 7: 実験結果

単文　日英翻訳						
	提案手法	人手評価	BLEU	NIST	METEOR	RIBES
PSMT	178	2.42	0.13	4.82	0.45	0.71
HSMT	187	2.65	0.14	4.91	0.46	0.72
RBMT(t)	157	4.15	0.13	4.78	0.47	0.73
RBMT(a)	375	4.01	0.15	5.02	0.49	0.73
RBMT(t)+PSMT	323	3.44	0.17	5.36	0.5	0.75
RBMT(a)+PSMT	523	3.72	0.19	5.66	0.51	0.75
RBMT(t)+HSMT	304	3.5	0.17	5.41	0.5	0.75
RBMT(a)+HSMT	503	3.71	0.19	5.62	0.51	0.75
人手評価との相関係数	0.44		0.31	0.31	0.59	0.52
提案手法との相関係数		0.44	0.92	0.89	0.91	0.83

単文　英日翻訳						
	提案手法	人手評価	BLEU	NIST	METEOR	RIBES
PSMT	159	2.83	0.18	4.63	0.45	0.77
HSMT	143	2.77	0.18	4.57	0.45	0.77
RBMT(t)	83	3.99	0.14	4.01	0.4	0.76
RBMT(a)	490	4.08	0.14	3.94	0.38	0.75
RBMT(t)+PSMT	243	3.63	0.23	5.29	0.51	0.8
RBMT(a)+PSMT	436	3.68	0.24	5.49	0.52	0.8
RBMT(t)+HSMT	223	3.68	0.23	5.31	0.51	0.8
RBMT(a)+HSMT	434	3.72	0.24	5.53	0.52	0.81
人手評価との相関係数	0.48		-0.08	-0.08	-0.13	0.12
提案手法との相関係数		0.48	0.27	0.28	0.19	0.27

重文複文　日英翻訳						
	提案手法	人手評価	BLEU	NIST	METEOR	RIBES
PSMT	44	2.58	0.12	4.47	0.42	0.67
HSMT	56	2.54	0.12	4.55	0.42	0.68
RBMT(t)	20	3.98	0.09	3.95	0.4	0.67
RBMT(a)	118	3.75	0.09	4.02	0.41	0.65
RBMT(a)+PSMT	122	2.95	0.14	4.84	0.44	0.68
RBMT(t)+PSMT	89	3.14	0.14	4.84	0.45	0.7
RBMT(a)+HSMT	107	2.98	0.14	4.79	0.44	0.69
RBMT(t)+HSMT	85	3.12	0.14	4.87	0.44	0.71
人手評価との相関係数	-0.04		-0.63	-0.66	-0.46	-0.36
提案手法との相関係数		-0.04	0.42	0.43	0.53	0.11

重文複文　英日翻訳						
	提案手法	人手評価	BLEU	NIST	METEOR	RIBES
PSMT	23	2.17	0.14	4.17	0.39	0.71
HSMT	29	2.46	0.14	4.29	0.4	0.72
RBMT(t)	5	3.92	0.12	3.75	0.37	0.71
RBMT(a)	92	3.87	0.11	3.72	0.36	0.7
RBMT(a)+PSMT	98	3.41	0.19	5	0.46	0.76
RBMT(t)+PSMT	50	3.55	0.19	4.87	0.45	0.76
RBMT(a)+HSMT	72	3.35	0.19	4.9	0.45	0.76
RBMT(t)+HSMT	37	3.56	0.19	4.88	0.45	0.76
人手評価との相関係数	0.32		-0.02	-0.04	0.09	0.27
提案手法との相関係数		0.32	0.32	0.34	0.34	0.36