自 然 言 語 処 理 154−19
（２００３．３．７）

翻訳精度評価手法 BLEU の日英翻訳への適用

金山 博　荻野紫穂
日本アイ・ビー・エム株式会社　東京基礎研究所
242-8502 神奈川県大和市下鶴間 1623-14
kanayama@trl.ibm.com, shiho@jp.ibm.com

概 要

本稿では、翻訳システムの性能を自動的に評価する手法である BLEU を、日英翻訳システムに対して適切に用いるための方法について述べる。ここで提案する BLEU の拡張では、翻訳システムの開発段階ごとにおける訳質の変化を検知できるよう、従来は単語の n-gram を指標としていたところに、文字の n-gram を導入した。実験によって、日英翻訳においても他の言語ペアと同様に人手による評価と高い相関が得られたこと、翻訳システムの改良前後の訳出に対しても有意な差をもって評価値を出力できるようになったことを示す。

Application of BLEU to Evaluation of Japanese-English Translation

KANAYAMA Hiroshi, OGINO Shiho
Tokyo Research Laboratory, IBM Japan, Ltd.
1623-14 Shimotsuruma, Yamato-shi, Kanagawa-ken, 242-8502 Japan
kanayama@trl.ibm.com, shiho@jp.ibm.com

Abstract

This paper describes how to apply BLEU, an automatic evaluation method of translation systems, to Japanese-English translation. In order to detect slight enhancement of a translation system under development, we extended BLEU by introducing character-based n-gram in addition to word-based n-gram which is used in original BLEU. A series of experiment shows that BLEU can evaluate Japanese-English translation systems with high correlation against human evaluation as well as for other language pairs, and our method enabled BLEU to distinguish the performance of pre/post-enhancement of a translation system with significant difference.

1　はじめに

機械翻訳による訳出の良さを自動的に測定するための手法は、恣意姓の無い客観的な評価を得るためや、人手による評価のコストを削減するために重要な技術である。評価の結果は、複数の翻訳システムの優劣を判断する目的だけでなく、翻訳システムの開発段階における改善の度合いの測定や、辞書等の整備が遅れている分野の発見などに役立つ。

翻訳システムの性能を自動評価するには、システムによって翻訳した結果と、翻訳者が生成した正しい翻訳（以後、参照文と呼ぶ）とを比較し、それらが類似しているほど高い評価値を与えるという手法が一般的である。類似度の計算には DP マッチング

を用いたものが従来から考案されている [1]。

しかし、ある文を意味が通るように翻訳する方法は一通りではないため、たとえシステムの出力が実用上意味を把握できるものであっても、参照文と異なる訳し方をしているために、高い評価値が得られない場合がある。そこで、別の対訳コーパスから類似文を探索して、参照文の別訳として用いるという手法も試されている [2]。

BLEU (Bilingual Evaluation Understudy)[3] は、システムの出力と参照文の間における n-gram の一致度を用いる自動評価方法であり、フランス語・スペイン語・中国語・アラビア語の各言語を英語に翻訳するシステムの性能を、人手による評価と高い相関をもって求められることが確認されている [4]。また、辞書などの言語依存の知識を必要とせず、客観性が高く簡潔な手法である。

本稿では、BLEU の枠組みを用いて日本語→英語の翻訳システムの評価が適切にできるかどうかを検証する。さらに、一つの翻訳システムを辞書等の更新により改良していく場合に、その訳出が向上していく度合いを詳細に検知できるように、BLEU の拡張を試みる。

2 BLEU の評価方法

本節では、BLEU での評価の指標である **BLEU 値**の計算方法について解説する。ここでは日本語から英語への翻訳の性能評価を前提とする。

2.1 n-gram の正解率

まず、日本語の文と、それを翻訳者が英語に翻訳した結果の対を準備する。この英語訳を**参照文**と呼ぶ。それぞれの日本語の文に対して、異なる翻訳者によって作成された複数の参照文があることが望ましい。

また、対訳の翻訳元である日本語の文を、評価対象となる翻訳システムで翻訳した結果を**テスト文**と呼ぶ。以下に例を示す。

［例 1］

翻訳元

私は事務所の人に時計を直してもらった。

参照文 1

I had my watch repaired by an office worker.

参照文 2

A person in the office repaired my watch.

テスト文 1

I had a man in the office repair a watch.

テスト文 2

I had the person of an office correct a clock.

BLEU 値の計算をするために、各テスト文の単語の n-gram のうち、一致するものが参照文のいずれかに含まれるものの割合を調べる。これを n-gram の正解率と呼ぶ。

例 1 のテスト文 1 では、出力された 11 語[1]のうち ‘I’, ‘had’, ‘a’, ‘in’, ‘the’, ‘office’, ‘watch’, ‘.’ の 8 語が参照文中に見つかるため、unigram の正解率は 8/11 となる[2]。また、‘I had’, ‘in the’, ‘the office’, ‘watch .’ の 4 つの bigram が参照文中に存在するため、bigram の正解率は 4/10 となる。trigram は ‘in the office’ のみなので 1/9 である。

一方、テスト文 2 では、unigram, bigram, trigram はそれぞれ 8/11, 2/10, 0/9 となり、テスト文 1 に比べて正解率が低く、テスト文 1 の方が良い翻訳であるという直感と一致する。

2.2 短い文に対する対処

n-gram の正解率を求める際に、テスト文の語数が分母となるため、参照文に出やすい一般的な単語だけからなる短い文を出力すれば、不当に高い正解率を得ることができてしまう。例えば、例 1 で “I had.” のみを出力すれば、unigram の正解率は 3/3 となるが、翻訳としては意味を成さない。この問題に対処するために、式 (1) の **brevity penalty** を定義し、テスト文が参照文より短い場合に、評価値に 1 未満の値を乗ずる。

$$\mathrm{BP} = \begin{cases} 1 & \text{if} \quad c > r \\ e^{(1-r/c)} & \text{if} \quad c \leq r \end{cases} \tag{1}$$

r, c はそれぞれ参照文・テスト文の総語数を表す。

[1] 句読点も 1 語と数える。

[2] テスト文 1 には ‘a’ が 2 回出現しているが、1 つの参照文で使われる ‘a’ は高々1 つなので、1 回分だけを正解とする。これは、“The the the the the.” のような無意味な出力の unigram の正解率が 1 となってしまうことを防ぐためである。

2.3 BLEU 値の計算

BLEU 値は、それぞれの n に対する n-gram の正解率 p_n の幾何平均に brevity penalty を乗じたものであり、式 (2) で計算される。幾何平均を用いるのは、p_n が n の増加に従って指数的に減少するためである。BLEU 値は 0 以上 1 以下の値をとる。

$$\text{BLEU} = \text{BP} \cdot \left(\prod_{n=1}^{N} p_n \right)^{\frac{1}{N}} \tag{2}$$

N の値が大きい（高次の n-gram までを考慮する）ほど、単語、ひいてはトピックの適正さよりも、文の流暢さを重視した指標になる。経験的に、他の言語対では $N = 4$ の時に人手による評価に近い評価値が得られることが報告されている [3]。

3 日英翻訳の BLEU による評価

経済白書平成 8 年版とその英訳 [5] のうち、日本語文の長さが 80〜90 文字である 23 文を用いて、3 つの機械翻訳システムの出力 S1, S2, S3 と、人手による翻訳 H1 の BLEU 値を測定した。H1 の翻訳は英語圏在住経験 2 年程度の非専門家によるものである。

前節の式 (2) の N を 2〜5 に変化させて実験を行った。例えば、「$N = 3$」は、unigram から trigram までの正解率を用いた指標である。また、「$N = 4$」について、‘-’ を常に単語の切れ目とみなして判定した場合[3]の BLEU 値も計測した。

また、BLEU との比較のために、人手による評価を行った。これは、「問題ない翻訳 (5)」〜「意味が全く読みとれない (1)」の 5 段階の点数を、2 人で独立に付与したものの平均値である。結果を表 1 に示す。

これを、評価が最も低い S1 が 0、評価が最も高い H1 が 1 となるように線形に正規化すると、図 1 のようになった。また、この場合の各手法と人手評価との相関係数を表 2 に示す。

実験の結果、trigram までを用いた場合は極めて高い相関が得られたが、4-gram までを用いると S2 と S3 の値が逆転した。しかし、‘-’ のところで単語を分割すると、4-gram でも人手評価とほぼ一致するようになった。これは、今回の実験では Papineni らの実験 [3, 4] と比較して各 n-gram の正解率が低く、高次の n-gram の正解率を高めることで妥当な評価ができるようになったためと考えられる。高次の n-gram の正解率が低い場合、“i.e.” や “(5%)” のように 4-gram として数えられる訳出が参照文と一致するか否かが、BLEU 値に大きな影響を与えていた。

5-gram 以降は正解率がさらに減少するので、妥当な評価が難しくなる。以降の実験では、‘-’ で単語を分割し、4-gram までを用いることにする。

[3]従来の BLEU では、単語分割の際に ‘high-resolution’ を 1 語としていたところを、‘high’ ‘-’ ‘resolution’ の 3 単語とみなすように変更したもの。

	S1	S2	S3	H1
$N = 2$	0.200	0.230	0.247	0.246
$N = 3$	0.093	0.100	0.114	0.133
$N = 4$	0.056	0.060	0.056	0.079
$N = 4$ (‘-’ 分割)	0.059	0.063	0.074	0.087
$N = 5$	0.039	0.042	0.043	0.057
人手評価	1.9	2.1	2.8	3.6

表 1: 日英翻訳の BLEU 値と人手評価

$N = 2$	0.79
$N = 3$	1.00
$N = 4$	0.84
$N = 4$ (‘-’ 分割)	1.00
$N = 5$	0.92

表 2: 人手評価との相関係数

4 僅かな訳質の差の検出

4.1 類似した翻訳システムの BLEU 評価

前節の実験で、BLEU が日英翻訳の評価にも利用できることがわかった。では、翻訳システムを辞書増強などによって改良する際に、訳質が向上されているかを確認するために、自動評価を用いることができるだろうか。Papineni の実験 [3]、前節の実験とも、比較しているのは訳質に大きな差がある別個の翻訳システム（及び人間の翻訳）であるため、訳出に微妙な変更が加えられた時にその差を自動的に検出できるかどうかは明らかでない。

そこで、ある翻訳システムに対して、辞書増強作業の過程の 4 段階における訳出を評価する実験を行っ

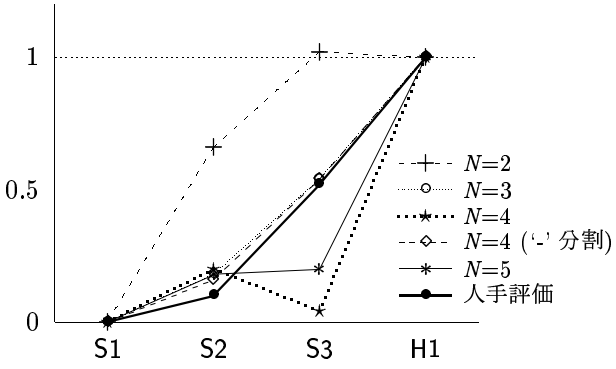

図 1: 日英翻訳の BLEU と人手による評価（正規化後）

	A-B	B-C	C-D
改良	33	41	23
改悪	9	11	1
その他	24	28	15

表 3: 人手評価による各システム間の改良/改悪の文数

た。各段階を、時期の早い順に A, B, C, D とする。

まず、各段階でどれだけの改良があったかを調べるために、各システムで日本語の文字数が 55 文字以下の 220 文を翻訳し、隣接する段階の翻訳結果を人手で比較した。2 つの訳出のうち、新しい側が訳出が良いと判断されたものが「改良」、新しい側が悪いものが「改悪」、何らかの差があるがどちらが良いとも判断できないものが「その他」である。この結果の表 3 から、A から B, C, D と各段階で辞書等を増強するにつれて、訳出が改良されたケースが多いことがわかる。

次に、各段階の訳出を、経済白書平成 8 年版のうちの 2870 文の対訳を用いて、BLEU で評価した。また、対訳コーパスを 50 分割したものの各部分に対する BLEU 値を測定し、それぞれの部分における A と B, B と C, C と D での BLEU 値の差の平均値・標準偏差と求め、それらを用いて差分の t 値を算出した。その結果を表 4 に示す。

A と B、及び C と D の差分の t 値はそれぞれ 0.27, 1.99 であり、自由度 49, 信頼係数 95% の t 値である 2.01 を超えていない。従って、A と B の間、及び C と D の間では評価値に差がないという仮説が棄却できず、BLEU ではこれらを有意な差として判定できなかったことになる。

		A	B	C	D
BLEU 値		0.0530	0.0531	0.0592	0.0602
差分	平均値	$6.9 \cdot 10^{-5}$	$6.9 \cdot 10^{-3}$	$9.2 \cdot 10^{-4}$	
	標準偏差	$1.8 \cdot 10^{-3}$	$5.2 \cdot 10^{-3}$	$3.2 \cdot 10^{-3}$	
	t 値	0.27	9.26	1.99	

表 4: 各段階の BLEU 評価

		正解率増	変化なし	正解率減
A-B	改良	**14**	11	8
	改悪	3	1	**5**
B-C	改良	**18**	5	18
	改悪	2	4	**5**
C-D	改良	**11**	6	6
	改悪	0	0	**1**

表 5: 文ごとの人手評価と BLEU 評価の比較

4.2 文ごとの n-gram 正解率

上記の実験で、それぞれのペア間で改良・改悪と人手評価で判断された文のうち、1 文の n-gram の正解率の増減を調べたものが表 5 である。太字の数字が、人手で判断した改良・改悪を BLEU で検知できたものであるが、改良を検出した割合は 44%, 改悪を検出した割合は 52%にとどまっている。

その原因は、以下の例 2 のように、辞書の変更によって訳出が改善されても、参照文の単語と一致するものが増えない場合が多いことである。

［例 **2**］

翻訳元 一方、これには重大な要因が含まれる。

参照文 By contrast, this includes an important factor.

C On the other hand, the serious factor is contained by this. (unigram : 5/13)

D On the other hand, the serious factor is included in this. (unigram : 4/13)

D は C に比べて「含む」の訳が改善されているものの、included が参照文の includes と一致しないため、unigram の正解率は上昇しないばかりか、下がっている。この問題を軽減するためには、参照文と単語が一致していない場合でも、表記が近づいた場合に評価を高めるような機構が求められる。

4.3 文字 n-gram の導入

テスト文と参照文の間で、単語の一部が一致した場合にも得点を与えられるように、単語 n-gram に加えて文字 n-gram を導入する。ここでの文字 n-gram とは、文を単語に分割した上で、各単語の中に含まれる n 文字の連続である。例として、“factor is included” という句の中には、5-gram は 6 つ (‘facto’, ‘actor’, ‘inclu’, ‘nclud’, ‘clude’, ‘luded’), 6-gram は 4 つ (‘factor’, ‘includ’, ‘nclude’, ‘cluded’) 含まれる。

この指標を用いると、例 2 の文字 n-gram の正解率は、C では 5-gram が 2/11, 6-gram が 1/7 で、D では 5-gram が 5/10, 6-gram が 3/6 となり、D の方に高い値を与えることができる。

文字 n-gram q_n に基づいた BLEU 値 BLEU_c を式 (3) のように定める。そして、式 (2),(3) の指標を組み合わせて、式 (4) の**拡張 BLEU 値**を定義する。

$$\mathrm{BLEU}_c = \mathrm{BP}_c \cdot \frac{1}{M-K+1} \sum_{n=K}^{M} q_n \qquad (3)$$

$$\mathrm{BLEU}' = (1-t)\mathrm{BLEU} \cdot t\mathrm{BLEU}_c \qquad (4)$$

単語 n-gram は、n が通常 1 文内の単語数より小さいことを前提としていたのに対し、文字 n-gram の場合は、n が一単語内の文字数よりも大きくなる場合がある。このため、文字 n-gram は単語 n-gram のように n に従って指数的に減少しないので、式 3 では相加平均を用いている。

図 2 は、文字 n-gram の正解率を A, B, C, D について測定したものである。全ての n に対して正解率の低い順に A, B, C, D となっている。$n \leq 4$ では、無関係な文字列と偶然一致している場合が多くなる。また、$n \geq 10$ のものは頻度が小さく、単語に偏りがみられる。従って、式 (3) で $K=5$, $M=9$ として実験を行う。簡単のため、式 (4) の t は 1/2 とする。

文ごとの人手評価による改良・改悪と拡張 BLEU 値の増減との関係は表 6 のようになった。この場合、改良の 54%と改悪の 57%が検出できており、従来の BLEU 値を用いた場合の結果（表 5）と比べてより多くの訳出の変更を適切に評価できていることになる。

また、表 7 は 4.1 節と同様の実験の結果である。A と B, B と C, C と D の差分の t 値が全て 2.01 を上回っているため、4 段階の訳出が順に改良されていることが確信度 95%以上の有意な差をもって示された。

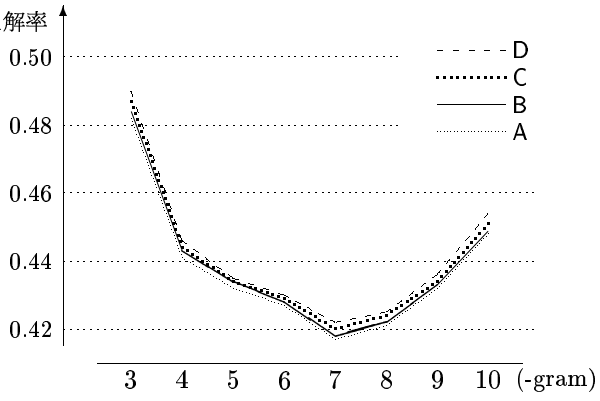

図 2: 文字 n-gram の正解率

		正解率増	変化なし	正解率減
A-B	改良	**19**	5	9
	改悪	4	0	**5**
B-C	改良	**20**	4	17
	改悪	1	4	**6**
C-D	改良	**13**	4	6
	改悪	0	0	**1**

表 6: 文ごとの人手評価と拡張 BLEU 評価の比較

5 考察

5.1 文字 n-gram 導入の妥当性

文字 n-gram を導入することにより、前節の例 2 の includes / included のように、テスト文と参照文の間で単語としては一致しない場合でも、長い文字列が一致する場合に評価値を上げることができた。この他にも同様の例として次のようなものがあった。

1. industry / industrial
2. disposable / disposability
3. evaluating / valuation
4. comparatively / relatively

		A	B	C	D
拡張 BLEU		0.2350	0.2358	0.2448	0.2457
差分	平均値	$5.2 \cdot 10^{-4}$	$1.1 \cdot 10^{-2}$	$9.4 \cdot 10^{-4}$	
	標準偏差	$1.5 \cdot 10^{-3}$	$1.1 \cdot 10^{-2}$	$2.6 \cdot 10^{-3}$	
	t 値	2.46	6.86	2.53	

表 7: 各段階の拡張 BLEU 評価

1〜3 のように互いに派生語であるものほか、4 のように語尾が共通する類義語が得られている。全く関連のない語のために正解率が上がってしまうケースは稀であった。

文字 n-gram を使わずに、活用形を原形に正規化してから比較するという方法も考えられるが、上記の例のような派生語を辞書で網羅するのは困難である。また、評価の際に辞書等のリソースを用いると、評価結果がリソースに大きく左右され、恣意性が高まってしまう。リソースの作成にコストが掛かる、評価手法が言語依存になるなどの点からも好ましくない。

なお、4.3 節の実験で、人手評価で A-B 間で改良と判定されたもののうち、文字 n-gram の指標を用いても正解率が増えていない 19 文を観察すると、その全てが、参照文との比較により改良を検知するのは困難なものであった[4]。

従って、テスト文と参照文の類似性を言語知識に依らずに表層のみから判断するには、文字 n-gram を利用することは妥当であるといえる。

5.2 BLEU 値と拡張 BLEU 値の使い分け

文字 n-gram を用いた拡張 BLEU 値は、辞書の変更によるシステムの改良を検知できることがわかったが、他システムとの比較にも用いることができるかを試してみたい。表 8 は、3 節で評価したものと同じ 4 つのシステムの拡張 BLEU 値を計測したものである。

拡張 BLEU 値と人手評価との相関係数は 0.52 となり、通常の BLEU 値を使った場合と比べて遙かに低い値となった。原因は、人手の翻訳である H1 の文字 n-gram の正解率が低いことである。機械翻訳と比較して、人手の翻訳は専門用語や抽象語などを正確に訳せない場合が多い。また、長い単語の繰り返しを避けるために、意図的に長い単語を訳出しないことがある。文字 n-gram の正解率は低いものの、構文解析の誤りが少なく、文の構成が適切なため、単語 n-gram の値は高くなっていた。従って、拡張 BLEU 値は機械翻訳相互の比較、特に同一システムの変更を検知する際に用い、人手の翻訳を含む全体的な評価の時には従来の BLEU 値を用いるのがよい。

[4]引用符の位置が正しくなったり、参照文で使われていない前置詞が変更されているなど。

	S1	S2	S3	H1
拡張 BLEU 値	0.283	0.286	0.307	0.293
人手評価	1.93	2.11	2.78	3.57

表 8: 日英翻訳の拡張 BLEU 値と人手評価

6 まとめ

BLEU の手法は、日英翻訳の評価に対しても、人手による評価と高い相関を示し、他の言語対と同様に適用できることが確認された。

また、単語 n-gram に加えて文字 n-gram の値を用いることにより、辞書の変更による訳出の向上を検知できる確率を高めることができた。この方法は、翻訳システムの辞書整備の際に、訳出の向上の度合いを確認しながら作業をするために有用である。

今後は、これらの指標と訳出の変化との定性的な関連を見出して、複数の観点から訳質を評価し、開発作業に寄与できる情報を引き出せるようにしたい。

参考文献

[1] Keh-Yih Su, Ming-Wen Wu, and Jing-Shin Chang. A new quantitative quality measure for machine translation systems. In *Proc. of 14th COLING*, pages 433–439, 1992.

[2] Keiji Yasuda, Fumiaki Sugaya, Toshiyuki Takezawa, Seiichi Yamamoto, and Masuzo Yanagida. An automatic evaluation method of translation quality using translation answer candidates queries from a parallel corpus. In *Proc. of MT Summit VIII*, pages 373–378, 2001.

[3] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proc. of the 40th ACL*, pages 311–318, 2002.

[4] Kishore Papineni, Salim Roukos, Todd Ward, John Henderson, and Florence Reeder. Corpus-based comprehensive and diagnostic MT evaluation: Initial arabic, chinese, french, and spanish results. In *Proc. of HLT2002*, pages 124–127, 2002.

[5] H. Isahara and M. Haruno. Japanese-English aligned bilingual corpora. In Jean. Veronis, editor, *Parallel Text Processing: Alignment and use of translation corpora. (Text, Speech and Language Technology, Vol. 13)*, pages 313–334. Kluwer Academic, 2000.