文字単位 BLEU による翻訳自動評価

Etienne DENOUAL & Yves LEPAGE
エティエンヌ・ドヌアール & イヴ゛・ルパージュ

ATR 音声言語コミュニケーション研究所
619-0288「けいはんな学研都市」光台 2-2-2
{etienne.denoual,yves.lepage}@atr.jp

1 はじめに

翻訳自動評価手法 BLEU を提案した論文 (Papineni et al., 2001) では、その評価手法は言語に依存しないと強調しているが、実際に、その手法は単語単位であるため、英語以外に適用されていない。例えば、NIST[1]・TIDES [2]・IWSLT (Akiba et al., 2004) 等のキャンペーンでは、予め単語単位に分かち書きされたテキストに自動評価手法を適用している。このように、英日の自動翻訳評価キャンペーンは行われていない。

2 分かち書きの問題

通常、統計翻訳システムは、単語の単位であるレキシコンモデルに依存しており (Brown et al., 1993)、日本語のような言語に翻訳する場合には、分かち書きされたテキストを出力する。しかし、市販用翻訳システムの場合は、分かち書きのない出力も見られる。たとえば Systran[3] の日本語訳テキストでは単語間の空白はない。

もちろん、そのような異なるシステムの性能を計るために、標準分かち書きツールを適用し (例えば、茶筅 (Matsumoto et al., 1999))、翻訳自動評価手法の適用は可能であるが、そのようにして得られたスコアは分かち書きツールの誤り率によって片寄った結果になる[4]。

3 提案・実験データ

3.1 文字単位 BLEU

どの言語の電子テキストでも、文字そのものは直接扱えるため、BLEU の基本定義 (Papineni et al., 2001) を変換せずに、単語 n-gram 単位ではなく、文字 n-gram 単位で BLEU の定義を適用することを提案する。本論文では、普段使われている BLEU_{w4}、すなわち 4-gram を利用した BLEU の尺度を検討して、文字単位での同等の尺度を求める。文字単位 M-gram の尺度を BLEU_{cM} と呼ぶ。

3.2 データセット

本論文の実験では、4 つの自動翻訳システムにより翻訳された 510 文のデータを使用する。あわせて、2,040 翻訳候補になる。

それぞれの原言語の文について、13 文の参照訳が予め人手で用意されている。候補文と参照訳について簡単な統計情報を表 1 で示す。

表 1: データセットの属性の平均・標準偏差

	候補	参照訳
一文あたり文字数	30.65 ± 15.95	31.58 ± 18.02
一文あたり単語数	6.31 ± 3.26	7.08 ± 3.31
一単語あたり文字数	3.84 ± 2.10	3.80 ± 2.07

4 BLEU_{w4} と文字単位 BLEU の同等性

4.1 最尤相関値

BLEU_{w4} スコアの最尤相関を持つ BLEU_{cM} スコアの M を推定するため、すべての可能な M に対して、BLEU_{w4} と BLEU_{cM} スコアの線形相関 (Pearson's correlation) を計算した。

最尤相関値は値 17 で得られた。

4.2 最尤判断一致

BLEU スコアは 0~1 の値であるため、10 段階の成績に変換することができる。10 の狭い間隔で $[0, 1]$ をカバーするため、0 から 9 までの 10 段階の成績を選択した。BLEU のスコアは、そのような成績値に変換すると、離散的な評価値になり、同じ文の単語単位と文字単位の BLEU を比べる作業は 2 つの順位判定手法を比べる作業となる。このような 2 つの判定手法の同等性を計る手段の一つは kappa 係数である。従って、BLEU_{w4} と BLEU_{cM} の間の同等性を計るためその手段を使用した。

Kappa 係数の最大値は値 18 で得られた。

[1] http://www.nist.gov/speech/tests/mt/.
[2] http://www-nlpir.nist.gov/tides/.
[3] http://www.systranbox.com/systran/box
[4] 標準テキストで分かち書きツールの誤り率は 5%~10%であって、自動翻訳システム出力テキストでその誤り率を正確に推定するために評価キャンペーンが正確に必要になる。

4.3 最尤類似行動

BLEU は修正 n-gram 精度の幾何平均で定義されている。従って、ある文にある n-gram が見つかるためには、2 つの $(n-1)$-gram が見つからなければならないので[5]、以下の特徴を満たす。

> すべての N、すべての候補、すべての参照訳集合について、
>
> $$\mathrm{BLEU}_{wN} \leq \mathrm{BLEU}_{w(N-1)}$$

以上の特徴を使用して、主に $\mathrm{BLEU}_{cM} \leq \mathrm{BLEU}_{w(4-1)} = \mathrm{BLEU}_{w3}$ という条件を満たす M 値を求めた。そのような M は BLEU_{w4} の 4 と同等であると考えられる。

90% のしきい値を設定すると、少なくとも 90% の場合でスコアが対応する BLEU_{w3} スコア以下になる M の最小値は 18 であった。

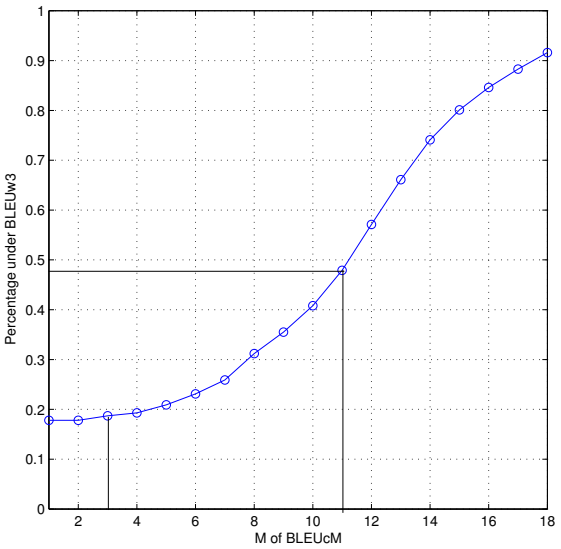

図 1: 1~30 で BLEU_{w3} スコア以下の BLEU_{cM} スコアの率

5 $\mathbf{BLEU}_{c18} \simeq \mathbf{BLEU}_{w4}$

以上の結果に基づいて、BLEU_{w4} と最尤相関・最尤判断一致・最尤類似行動を持つ M の値はそれぞれに 17・18・18 であったため、4 単語の単語単位 BLEU_{w4} に相当する M 文字単位 BLEU は $M = 18$ の値であろうと結論した。

採用した 4 つのシステムにあたりそれぞれの BLEU_{w4} と BLEU_{c18} スコアを表 2 に示す。

[5]例えば、4-gram の abcd が見つかるためには、abc と bcd の 2 つの 3-gram が見つからなければならない。

文字単位で計算すると単語単位で計算するより平均 0.047 減少した。N 単位以下の文は N-gram BLEU の尺度で計算するとやむを得ずスコアが 0 となるため、観測した現象を説明できる。採用したデータの中には、18 文字以下の文の数 (350) は 4 単語以下の文の数 (302) より多いので、0 のスコアを当る文は BLEU_{c18} で多くなり、システムのスコアは BLEU_{c18} で BLEU_{w4} より低くなった。

システム 2 とシステム 3 の順序は変わったが、主な序列は変わっていない。実際、システム 2 とシステム 3 のスコアの差異は統計的に優位とは考えられない。(Zhang et al., 2004) では、2%の信頼区間を報告しているため、システム 2 とシステム 3 をお互いに位づけるのは困難であろう。

6 換算式

6.1 尺度の粒度

データセットによる BLEU_{w4} と BLEU_{c18} スコア分布を図 2 で示す。この図を見ると、BLEU_{w4} スコアが特定の領域にしか存在しないことが明らかになった。BLEU_{c18} の場合はそのような現象は観察できない。BLEU_{w4} スコア分布で現れた周期的な集積はフーリエ変換で明白になる。BLEU_{w4} スコアのデータで最初のフォルマントは 20 であったので、周律は $1/20 = 0.05$ であると結論付けられる。これに反して、BLEU_{c18} フーリエ変換データを人手で検査すると相対極大値も見つからず、周律は見られない。

次のような解釈ができる。単語単位尺度である BLEU_{w4} では、内在精度があると考えられる。その尺度のスコアは 0 から 1 まで公平に分布していなくて、0.05 の倍数の値の付近にあたるようである。逆に、文字単位尺度である BLEU_{c18} では、$[0, 1]$ の間隔でよりなだらかに分布するようである。

結論として、BLEU_{w4} を BLEU_{c18} に交代すると、単語単位の尺度の内在精度を消滅させることができた。

6.1.1 $\mathbf{BLEU}_{w4} \rightarrow \mathbf{BLEU}_{c18}$ 換算

BLEU_{w4} の精度を利用して単語単位と文字単位の尺度の間に文に対して次の換算を提案できる。同じ BLEU_{w4} 精度区間にあたる BLEU_{c18} スコアを検討してそのスコアの平均・標準偏差を計算した。結果は図 3 に示す。平均は 1 に近くなると対応する BLEU_{w4} スコアにも近くなる。従って、スコアが高いほど相関率が高い。標準偏差が平均 0.078 であって、BLEU_{w4} での x のスコアは BLEU_{c18} で $x \pm 0.078$ に均等であろうと結論付けることができる。

6.1.2 $\mathbf{BLEU}_{c18} \rightarrow \mathbf{BLEU}_{w4}$ 換算

BLEU_{c18} スコアには周期性は見られないので、逆の計算をするために、BLEU_{c18} 精度をあてにすることができず、$[0, 1]$ の値間隔を 10 の間隔に分けて、それぞれの BLEU_{w4} スコアの間隔の標準偏差を計算した。

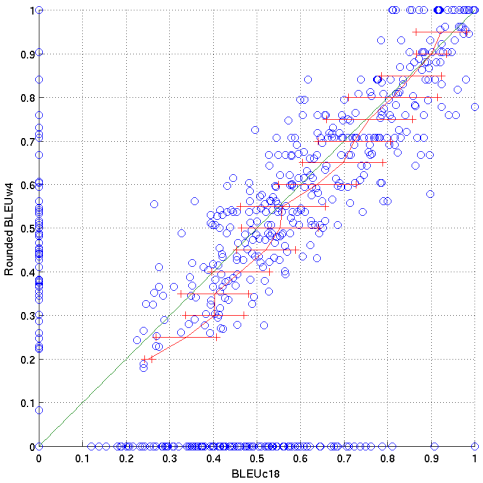

図 3: BLEU_{w4}(縦座標) 対 BLEU_{c18}(横座標)— 精度のステップあたりの BLEU_{c18} スコアの平均・標準偏差

結果は表 3 に示す。非常に低い値や高い値には標準偏差値は低くて、単語単位 BLEU を文字単位 BLEU と置き換えるには低い値や高い値のあたりが安全であろうと思われる。実際に、実験の結果によると 85% の文の文字単位 BLEU スコアは 0.2 以下であるか 0.6 以上であって[6]、その二つの区間で BLEU_{w4} の標準偏差は 0.125 である。

7 おわりに

本論文では、文字単位 M-gram と単語単位 N-gram で BLEU の適用を試みた。普段に使用されている 4-gram 単語の値について検討した結果、M-gram 文字単位で 18 が対応していることを示した。その対応をより詳細に検討して、実用的な換算式を推定した: $\mathrm{BLEU}_{c18} \simeq \mathrm{BLEU}_{w4} \pm 0.078$.

本研究は、日本語のような分かち書きのない言葉へ翻訳する翻訳システムの自動評価キャンペーンのため、文字単位 BLEU の適用を容易にする。

謝辞

本研究は、情報通信研究機構の研究委託「大規模コーパスベース音声対話翻訳技術の研究開発」により実施したものである。

[6] これは文のスコアである。システムに対してのスコアはそのような文のスコアの平均であって、普通は、システムの全体スコアは $[0.30, 0.60]$ にあたる。しかし、実験によると、システムのスコアは二つの異なる母集団 (低い文スコアと高い文スコア) の平均である。

参考文献

Yasuhiro Akiba, Marcello Federico, Noriko Kando, Hiromi Nakaiwa, Michael Paul, and Jun'ichi Tsujii. 2004. Overview of the IWSLT04 evaluation campaign. In *Proc. of the International Workshop on Spoken Language Translation*, pages 1–12, Kyoto, Japan.

Peter E. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics, Special Issue on Using Large Corpora: II*, 19(2):263–311, March.

Y. Matsumoto, A. Kitauchi, T. Yamashita, Y. Hirano, H. Matsuda, and M. Hasahara. 1999. Japanese morphological analysis system ChaSen version 2.0. Technical report NAIST-IS-TR99009, Nara Institute of Technology.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. Research report RC22176, IBM, September.

Ying Zhang, Stefan Vogel, and Alex Waibel. 2004. Interpreting BLEU/NIST scores: how much improvement do we need to have a better system? In *Proceedings of LREC 2004*, volume V, pages 2051–2054, Lisbonne, May.

表 2: 四つのシステムの BLEU スコア。

	システム 1		システム 2		システム 3		システム 4
overall BLEU_{w4} スコア	0.349	$>$	0.312	$\sim$	0.305	$>$	0.232
overall BLEU_{c18} スコア	0.292	$>$	0.267	$\sim$	0.279	$>$	0.183
difference	0.057		0.045		0.036		0.049

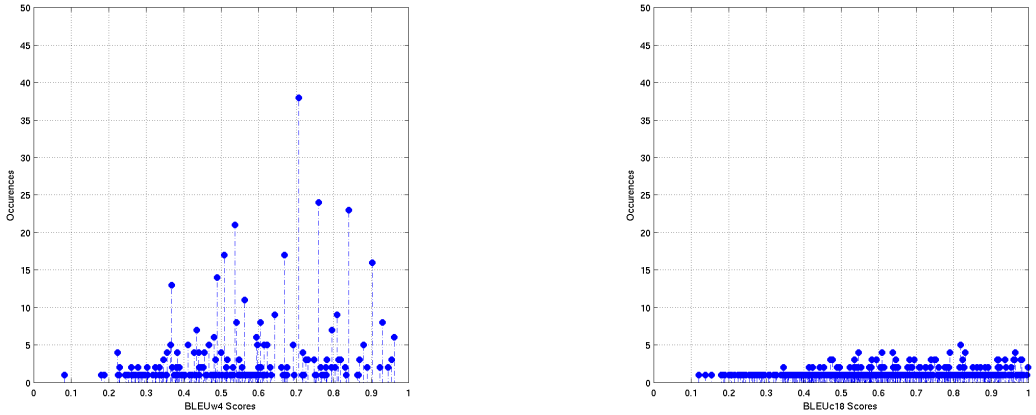

図 2: データセットで BLEU_{w4} と BLEU_{c18} スコア分布。

表 3: BLEU_{c18}・BLEU_{w4} 間の換算同等

BLEU_{c18} での間隔	]0 , 0.20[	[0.20 , 0.30[	[0.30 , 0.40[	[0.40 , 0.50[	[0.50 , 0.60]	[0.60 , 0.70]	[0.70 , 0.80]	[0.80 , 0.90]	[0.90 , 1]
BLEU_{c18} 値率	nr	2.36%	4.55%	10.46%	15.68%	12.98%	12.31%	15.01%	26.64%
BLEU_{w4} 値率	nr	2.70%	5.56%	8.26%	14.50%	10.79%	13.49%	7.96%	36.59%
標準偏差	± 0	± 0.102	± 0.052	± 0.113	± 0.073	± 0.058	± 0.060	± 0.062	±0.033
BLEU_{w4} での対応間隔	[0 , 0]	]0.01 , 0.4[	[0.25 , 0.45[	[0.29 , 0.61[	[0.43 , 0.67]	[0.54 , 0.76]	[0.64 , 0.86[	[0.74 , 0.96[	[0.87 , 1]