社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2009－ＦＩ－ 93 (1)
2009－ＮＬ－189 (1)
2009／1／22

大規模コーパスの効率的な Bleu 値計算方法の提案

岡田勇（創価大学） 宮澤信一郎（秀明大学）

要約 本研究は、機械翻訳システムの品質評価を扱う。コーパスの規模と性能との予測が可能となることは、研究上また実用上、大きな意義を持つ。そこで、日英対訳特許文書から生成されたコーパスを用いて、機械翻訳の性能とコーパスの規模との関係に関する見通しを得ることを目的とする。我々は、大規模なコーパス文書対の全ての組み合わせに対する Bleu 値を計算するための効率的な手法を提案し、その実行結果について報告する。その結果、本研究で定義したコーパスの性能の推移は、規模がある程度以上の場合、コーパス規模の対数に関する線形回帰で表せることを確認した。

Approach to Efficient Calculation of BLEU in Large Scale Corpora

Isamu Okada
Soka University

Shin-ichiro Miyazawa
Shumei University

Abstract. This paper reports on a performance estimate of Machine Translation System. To clarify the relationship between a size of corpora and its performance would be one of the most interest research topics. We aim to gain some prospects for the relationship by using a corpus generated by public patent claims in Japanese and English. This paper proposes an efficient calculating method of BLEU values for all combinatorial sentences in the large scale corpus. As a result of implementation, we indicate that the performance of the corpus can explain the logarithm of its size with a linear regression model under the condition of a certain size level.

1 はじめに

機械翻訳システムに大規模コーパスを用いる手法が提案されているが、その翻訳品質に対しては、多くの課題が残っている。しかし、現実にはコーパスを用いた機械翻訳の重要性が高まるにつれてある程度の品質をどのように保証するかは喫緊の課題となっている。特に、コーパスの規模と翻訳性能との関係について理解することは集めるべきコーパスの量に関する見通しを与えてくれるなど、重要な知見をもたらすだろう。

そのため本研究では、コーパスの規模とそれに基づく機械翻訳システムの品質評価について議論することにしたい。機械翻訳の品質には大きく人手評価と自動評価に分かれるが、人手評価には莫大なコストと時間がかかるため、自動評価でいかに正確な品質評価が可能であるかについて研究が進められてきている。さらに、隅田 (2005) は Bleu の提案によって機械翻訳研究には自動評価結果を示すことが当然視されている現状を指摘している。つまり、規模に対応した品質の大まかな見通しを与えるためには、自動評価でどのくらいの性能を測定できるかが重要となる。そこで、品質を自動評価する方法について議論する。

自動評価とは、あらかじめ正解文が設定されている場合では、正解文と翻訳文との類似性に基づいて品質を評価するのが一般的である。本研究では n-gram を用いた類似性指標を用いた自働評価手法について提案する。

ところで、類似性指標は文単位の評価を計測するものであって、それらを集約して翻訳システムの性能をどう評価するのであろうか。内元 (2003) や吉岡 (2005) では、交差検証法 (n-fold cross validation) が性能評価の方法として使用されている。これは、対象となるデータセットを n 分割し、そのうちの 1 分割を残りのデータセットから推定した時の誤差を全ての分割に対して集め、その自乗

和平均を取ることでデータセットの評価を行う手法である。本研究ではこれを参考に、新たな方法を提案する。

我々は、日英対訳特許文書から生成されたコーパスを用いる。内山 (2007) は、特許の機械翻訳が日英・英日の機械翻訳の対象として適切であると主張している。また、森下 (2008) は対訳特許文から専門用語辞書を半自動獲得する手法について新たな手法を提案している。このように日英対訳特許文は機械翻訳研究に良く用いられている。さらに、特許という対象の実用性からも、機械翻訳の性能とコーパスの規模との関係を議論することは有益である。

我々は、機械翻訳の性能とコーパスの規模との関係に関する見通しを持つために、全てのコーパス文書対の類似度を効率的に計算することにした。本研究では、大規模なコーパス文書対の全ての組み合わせに対する Bleu 値を計算するための効率的な手法を提案し、その実行結果について報告する。

2節では、本研究の翻訳品質の評価指標に関する議論を整理し、Bleu ならびにそれを構成する n-gram を基本としたコーパス性能の評価法について検討する。3節では、提案手法の詳細を説明する。4節では、研究対象と提案手法を実装した結果について報告する。5節では、結果について議論し、本研究をまとめる。

2 翻訳品質の評価法

翻訳品質をどう評価するのかについては多くの研究がある。このうち、自動評価手法としては、あらかじめ正解文が設定されている場合では、正解文と翻訳文との類似性に基づいて品質を評価するのが一般的である。ここでは翻訳品質の評価法として個々の文の類似性評価の方法についてと、それを全体として集約する方法について議論する。

2.1 類似性の評価

類似性を表す指標としては、Papineki(2002) の Bleu 値をはじめ多くが提案されており、それらの性能についても多くの研究がある。今村 (2004) は、機械翻訳の品質評価について複数の評価指標を比較しており、「完全訳を重視した場合、BLEU, WER の効果が高く、理解可能訳を重視した場合、WER, NIST の効果が高」いとの知見を得ている。岸田 (2006) は、Bleu 以外に、WAMU という指標を定義し、それらを用いて言語横断検索の性能を回帰モデルで予測する方法を提案している。江原 (2007) では、 Bleu の欠点として、単語の文字面のもの一致に基づく点と構文的な自然性が認識できない点を上げ、新たな評価指標の開発を試みているが、基本的には n-gram の概念を用いている。このように Bleu には多くの議論がなされているが、それに代わる有効な指標が、現状では提案されていない。また本研究目的に照らし、厳密な正確性ではなく、限られた時間と資源制約で概算値を求めるほうが望ましいことから n-gram 一致率を基にする単純な指標である Bleu 値を、翻訳品質の評価指標として取り上げる。

Papineki(2002) による Bleu 値の定義は次の通りである。

長さ c の候補文 C に対する長さ r の正解文 R の n-gram Bleu 値は

$$BP = \exp(1 - \max\{1, r/c\})$$

なる BP を用いて、

$$Bleu(R, C) = BP \exp(\sum_{i=1}^{n} log(p_i)/n)$$

で与えられる。ここで p_n とは n-gram 正解率（一致率）のことである。

英文の品質評価に際して Bleu 値を用いる場合、金山 (2003) によると、人手評価との比較において、英単語を単位とする 4-gram で計算することが望ましいことが示されている。日英対訳特許文において、4-gram による Bleu を用いるとすると、定義から、 $Bleu > 0$ であるための必要十分条件が、4-gram で一致する単語列を有することであるから、文の 4-gram の出現パターンを検討することが本質的に重要である。

次に、Bleu の計算を効率的に行うための方法について述べる。Bleu の定義から、4-gram までの

句の一致率を数え上げることが重要である。そのため、文を事前に 4-gram に分割しソーティングすることで効率的に Bleu を計算できる方法を提案する。

2.2 コーパスの評価

コーパスの規模を変えたときの翻訳性能の変化を調べるために、コーパスの評価をいかに行うか議論したい。交差検証法では、コーパス全体をテスト文と参照文に分割し、参照文からテスト文がどのくらい再現できるかを評価するものである。一方、コーパスによる機械翻訳とは、コーパスの中から最も翻訳文に近い候補文をいかに抽出できるかにかかっている。このような優れた翻訳システムが構築できたとすると、抽出された候補文は、目的となるテスト文との類似性が最も高いものであるとみなして差し支えない。つまり、全てのコーパスを構成する文の中で最も高い類似性を持つ文の類似度がコーパスの性能を表すものとなる。そのため、通常はコーパスの規模が大きくなれるほど、そのコーパスの性能は向上する。一方で、コーパスを大きくするにはコストがかかるので、コストと性能のトレードオフが存在し、そのバランスについて議論できることが重要となる。

このような考え方に基づき本論文では、規模 s のコーパス C の翻訳性能 $R(s)$ を次のように定義することとする。

まず、コーパスの対象とすることができる文のサイズに比べて s が十分小さい場合は、どの文を対象とするかでいくつもの $R(C)$ を計算できるため、規模 s のコーパスを複数作成して計算した $R(C)$ の平均を性能とする。そこで、以下ではある規模 s のコーパスを構成したものとして、その性能を定義する。

交差検証法では、コーパスの一部をテスト文としていたが、本研究では、コーパスを構成する文ではない文集合からテスト文を構成するものとし、その集合を T とする。なぜなら一般的に $Bleu(x,y) \simeq Bleu(y,x)$ であるので、交差検証法では対称性が生じやすく、テストデータをコーパスから独立したほうが望ましいと考えられるからである。T の全ての要素 t に対して、C の全ての要素との Bleu 値を計算し、そのうち最大となるものを c_t とする。すなわち、

$$c_t = arg\max_{c \in C} Bleu(t,c)$$

である。ある優れた翻訳システムを使用すれば、t に対するコーパス候補文として c_t を選択することが出来るはずであるから、これをコーパスの性能の基礎とする。そこで、$Bleu(t,c_t)$ を全てのテスト文に対して行った平均をコーパス C のテスト文集合 T に対する性能 $R(C|T)$ と定義する。すなわち、

$$R(C|T) = ave(Bleu(t,c_t)) \quad \forall t \in T$$

である。これを用いて規模 s のコーパス性能 $R(s)$ を計算する。

この計算は、規模が異なるごとに行う必要があり、計算量爆発の原因となりうるため、それを回避する手段を提案する。その本質は、テスト文と最大規模のコーパス文とのすべての組み合わせに対する Bleu を最初に計算しておいてから、任意の規模のコーパスの性能を計算するというものである。次節において詳細に説明する。

3 提案手法とアルゴリズム

ここでは、Bleu 値を効率的に計算する方法と、コーパスの性能を計算する方法を提案し、そのアルゴリズムを説明する。

3.1 Bleu の効率的な計算方法

あるテスト文 x に対する候補文 y における Bleu 値を $Bleu(x,y)$ と定義し、この効率的な計算方法を説明する。そのため前処理として、対象文ごとに、全ての 4-gram の句と最後の 3,2,1 単語からなる句を、ソートしてファイルに保存する。例えば、文 x として "The system works in the front of the neck freely." 文 y として "The system runs in the front of the head." とする場合を考えよう。x に対

して表1のようなファイル Fx を、y に対して表2のようなファイル Fy を生成する。

表1：文 x から生成されるファイル Fx

freely
front of the neck
in the front of
neck freely
of the neck freely
system works in the
the front of the
the neck freely
the system works in
works in the front

表2：文 y から生成されるファイル Fy

front of the head
head
in the front of
of the head
runs in the front
system runs in the
the front of the
the head
the system runs in

この前処理により、Bleu 値を計算するための作業領域として、対象文の量の約4倍のメモリ領域を用いることになる。

次に、Fx のすべての行 lx に対して以下の動作を行う。lx の第1単語と一致する行が Fy にある場合、その行 ly が lx と第何単語まで一致するか調べ、その値 k に対し、$G_{Fx}(k)$ に1を加える。ただし ly が複数あっても追加する値は最大のもの一つのみとする。こうして出来た $G_{Fx}(k)$ から n-gram 一致率を算出することができ、

$$p_k = \frac{\sum_{k \leq i} G_{Fx}(i)}{\#ly - k + 1}$$

となる。ただし、ly の個数を $\#ly$ とする。以上から Bleu を計算することが出来る。上記の例では $Bleu(x, y) = 0.459$ となる。この方式で計算すると計算量は、単語数のオーダーとなり効率的である。

3.2 コーパス性能の効率的な計算方法

任意の規模のコーパスの性能評価を効率的に行うため、テスト文 T と最大コーパスを構成する場合のコーパス文 C に関する全ての $Bleu(T, C)$ を計算し、Bleu 値表を作成する。これを用いて任意の規模 s のコーパスの性能 $R(s)$ を以下の方法で計算する。

最大コーパスを分割して、規模 s のコーパス C_i を作成する。Bleu 値表のうち C_i を構成する文の行を取り出し、テスト文ごとに最大 Bleu 値を算出し、その平均を $R(C_i|T)$ とする。これを全てのコーパス C_i に対して行った平均を $R(s)$ とする。この方法によると計算量は (テスト文の規模×最大コーパスの規模) のオーダーとなるので効率的といえる。

4 使用データと計算結果

ここでは使用データについて説明し、提案手法を実装した結果について紹介する。

4.1 使用データ

日英対訳特許文として、(財) 日本特許情報機構から提供された公開特許公報要約と PAJ 対訳データの 2003 年公開全件 348,061 件の中から、英文対訳で特許概要を説明している部分 (タグ <SDOAB LA="E">) を対象とする。この部分は実質的に特許の概要を説明している部分である。ここは「解決すべき問題 (PROBLEM TO BE SOLVED:)」という部分と「解決 (SOLUTION:)」という部分に分かれる。前者はほぼ To 不定詞で始まり問題を名詞句の形で表現している。後者で具体的な概要が文として記述されている。そこで全データのうち、後者の部分が1文以上あるもの 348,058 件のうち、327,680 件を対象にした。この部分の2文目以降は、1文目と同じ専門用語や名詞句の使用が

散見されることから、1文目に対する Bleu 値が高くなりやすくコーパスの性能に影響を与えやすい。しかし、これらの文は同時に翻訳されるのが普通であるから、全ての文書対に対する Bleu 値の計算を想定する場合は、コーパスに含めるべきではなく、今回の対象には適さないため、英文説明の1文目のみを抽出し対象データとする。

表3：テスト文とコーパス文の基本統計量

	テスト文	コーパス文
規模	1000	100000
1文あたりの平均単語数	54.10	52.78
総単語種数	4744	62948

このデータに対し $32768 = 2^{15}$ 文ごとにグループ0から9を作り、各グループごとに0から順に、文に番号を付与する。各グループに対して、文番号0－99の文をテスト文とし、文番号10000－19999の文をコーパスを構成する文の候補とする。表3にデータの基本統計量を示す。

4.2 計算結果

前節で提案した方法を Linux 上にて perl で実装した結果、最も処理時間のかかる、全組み合わせに対する Bleu の事前計算であっても、1日程度のオーダーで計算可能であることが分かった。また、

コーパスの規模別の性能は表4と図1にまとめる。また、規模別の n-gram の一致率の平均の推移を表4と図2に、性能と n-gram 平均一致率との相関を表5にまとめた。ここで相関とは、n-gram の一致数と性能という二つの変量に関する全部で10種類ある規模ごとに計算した相関係数のことである。

表4：規模別のコーパス性能と n-gram の平均一致率

規模	性能	1-gr.	2-gr.	3-gr.	4-gr.
100	0.00066	0.0185	0.0018	0.0004	0.0000
200	0.00117	0.0207	0.0024	0.0007	0.0001
500	0.00233	0.0238	0.0036	0.0013	0.0002
1000	0.00374	0.0263	0.0047	0.0020	0.0004
2000	0.00579	0.0287	0.0060	0.0029	0.0007
5000	0.00979	0.0313	0.0079	0.0043	0.0013
10000	0.01370	0.0325	0.0095	0.0057	0.0020
20000	0.01823	0.0340	0.0114	0.0073	0.0028
50000	0.02488	0.0356	0.0143	0.0098	0.0042
100000	0.02999	0.0369	0.0164	0.0115	0.0055

表5：性能と n-gram との相関係数

n-gram	相関係数	F 値
1-gram	0.912	0.0002
2-gram	0.993	0.0000
3-gram	0.999	0.0000
4-gram	0.994	0.0000

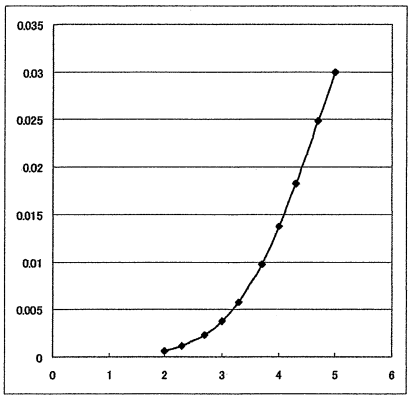

図1：コーパスの規模(横軸)と性能(縦軸)の関係

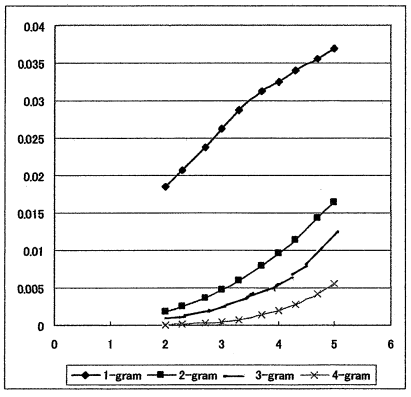

図2：コーパスの規模(横軸)と
平均 n-gram 一致率(縦軸)の関係

5 議論とまとめ

本研究では、大規模コーパスの翻訳性能を少ない記憶領域と計算時間で測定する方法を提案し実装した。その結果、計算すべき 1000 × 10 万件のうち、Bleu が正となるデータも 32,628 件と少なく、重複計算を防ぐことで 1 日程度の計算時間で測定できた。

表 4 ならびに図 1 から、本研究で定義したコーパスの性能の推移は規模が 5000 以上であると、コーパス規模の対数に関する線形回帰で表すことが出来ると予測できる。回帰直線と決定係数を求めると

$$R = -0.0486 + 0.0157 \log s$$

となり、決定係数は 0.997 の高い水準となっている。全ての規模に対する回帰直線の決定係数も 0.92 と高い。この回帰直線がどのくらいの規模に関して有効なのかは、今後の議論であるが、日英対訳特許における近似的なコーパスの性能予測としては有効であろう。さらに、性能予測曲線は規模の増加に伴いやがて収束することが推測できるので、全体として S 字曲線を描くと思われる。この点も今後検討したい。

また、表 5 からコーパスの性能と n-gram との相関は概ね高く、3-gram では 99.9% となっている。つまり、Bleu で計測した大まかな性能は 3-gram 一致率で計測する場合と近似することを意味する。

本研究は日英対訳特許文書という限定されており、ただちに一般的な性能評価に結びつかない。表 3 から明らかなように、テスト文の総単語種類数は 5000 弱であるが、その 100 倍の規模のコーパス文のそれは 6 万を超えていることから、対象が専門用語の多い特殊性を有することが示唆される。これをどのように拡張して、一般性を持たせるかは今後の課題となろう。

さらに、コーパスの性能として 4-gram までの一致率を使った Bleu を用いたが、他の指標についても検討する必要があるかもしれない。例えば、特許文は特定の専門用語が散見され、その存在によって性能を大幅に低下することがありうる。例えば "I watch an orange under the table." という文と "I watch an apple under the table." という文は 4-gram で一致する部分がないので $Bleu = 0$ であるが、文の構造は完全に再現され、たった一つの名詞の置き換えで正解文と一致できる良質な翻訳であると判定しても良い。日英対訳特許でもこれと同種の問題が起きる可能性は高く、これに対処できるような指標は検討すべきであろう。

参考文献

[今村 2004] 今村賢治, 隅田英一郎, 松本裕治, 機械翻訳自動評価指標の比較, 第 10 回言語処理学会年次大会, 452-455.

[江原 2007] 江原暉将, 新しい機械翻訳自動評価基準 NMG の提案, Japio 2007 YEARBOOK, (財) 日本特許情報機構, 238-265.

[金山, 2003] 金山博, 荻野紫穂, 翻訳精度評価手法 BLEU の日英翻訳への適用, IPSJ SIG Notes, 2003(23), (2003-NL-154-19), 131-136.

[岸田 2006] 岸田和明, 三田図書館・情報学会, ２００６年度研究大会.

[森下 2008] 森下洋平, 宇津呂武仁, 山本幹雄, 対訳特許文書からの専門用語対訳辞書半自動獲得におけるフレーズテーブルと既存対訳辞書の併用, 情報処理学会研究報告, 2008-NL-187, 91-98.

[Papineki 2002] Papineki, K., et.al., Bleu : a Method for Automatic Evaluation of Machine Translation, Proc. ACL 2002, 311-318.

[隅田 2005] 隅田英一郎, 佐々木裕, 山本誠一, 機械翻訳システム評価法の最前線, 情報処理, 46(5), 552-557.

[内元 2003] 内元清貴, 関根聡, 村田真樹, 井佐原均, 用例に基づく手法と機械学習モデルの組み合せによる訳語選択, 自然言語処理, 10(3), 87-114.

[内山 2007] 内山将夫, 山本幹雄, 藤井敦, 宇津呂武仁, 特許情報を対象とした機械翻訳:共通基盤による評価タスクを目指して, IPSJ SIG Notes, 2007(76), (2007-Nl-180-23), 133-138.

[吉岡 2005] 吉岡篤志, 徳久雅人, 村上仁一, 池原悟, 名詞句パターン辞書を用いた日英機械翻訳の試作 - bi-gram による訳出選択の場合, 平成 17 年度電気・情報関連学会中国支部連合大会論文集, 305-306.