Statistical Significance Tests for Machine Translation Evaluation

松本 宏

長岡技術科学大学

2015 年 6 月 3 日

Paper

文献

- Statistical Significance Tests for Machine Translation Evaluation.
  - KOEHN, Philipp.
  - In: EMNLP. 2004. p. 388-395.

Agenda

1. Paper
2. Abstract
3. Introduction
4. BLEU
5. 予備実験
6. Statistical Significance
7. Bootstrap Resampling

Abstract

翻訳機の性能比較において、それぞれが出力する結果の差異がそれぞれのシステムの差を表しているのだろうか？
そこで、推計的有意性を提案する bootstrap resampling 手法つかって計算し、比較を行う

Introduction

(2004 年) 統計的機械翻訳と自動翻訳評価が登場し機械翻訳が盛んになっていた
しかし、(BLEU) スコアの向上が統計的有意性があるかが不明そこで、boostrap resampling 手法を提案する

BLEU

BLEU は翻訳評価に頻繁に利用される評価手法
式は以下

$$BLEU = BP \cdot \exp(\sum_{n=1}^{N} \log p_n) \tag{1}$$

$$BP = \min(1, e^{1-\frac{r}{c}}) \tag{2}$$

r : 参照訳の単語長さ

c : 翻訳の単語長さ

n-gram のプリシジョン p_n をサイズ N(=4) まで測る
人で評価との相関があるとの報告もある

BLEU の特徴

2 つの特徴

- 高次 ngram と BP (*brevity penalty*) への依存
- *BP* の強い影響

高次 ngram と BP への依存

System	1-gram	4-gram	%BLEU
Spanish	62.6%	14.7%	28.9%
Portuguese	60.9%	13.4%	27.4%
Danish	60.8%	13.3%	26.9%
Greek	59.4%	12.1%	25.3%
German	58.3%	9.8%	22.6%
Finnish	56.1%	7.8%	20.2%

Figure : 翻訳比較:ngram precision と BLEU 値

- unigram では 60%前後に対して、4-gram では 15%以下
- Spanish 対 Finnish では 2 倍差

BP の強い影響

BP の役割はシステムが断片翻訳を行わせないためのもの

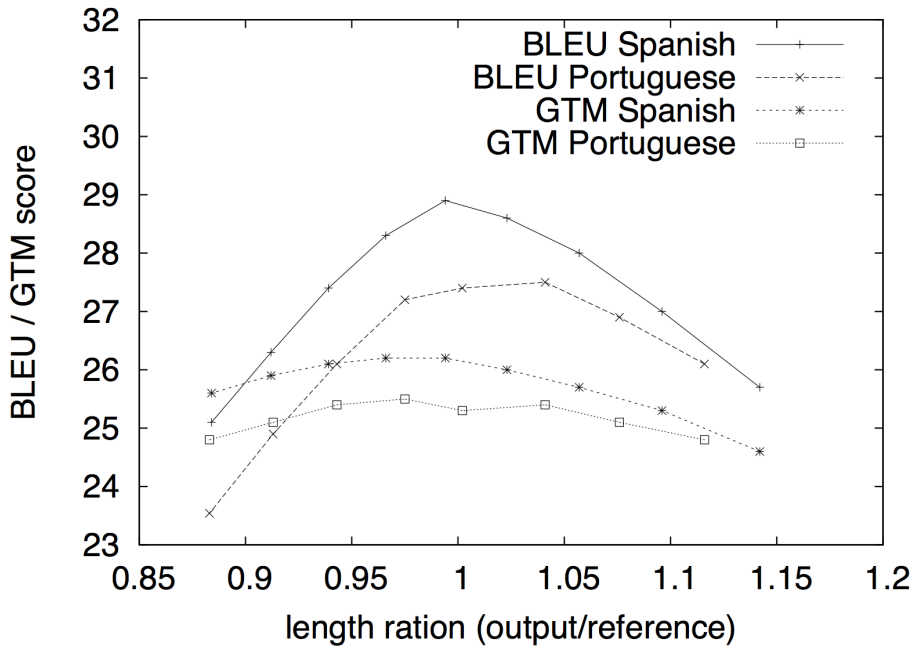

Figure : 出力長さによる BLEU と GTM での影響

- BLEU スコアでは出力が短くなる程 (グラフ左) 下降しているのが分かる
- GTM (n-gram Precision/Recall 評価) ではそうでもない

システムとコーパスの関係

この研究の目的は異なるシステムでの性能比較を行うこと
だが、ここでは同システムで異なるコーパスでの比較を行う
目的言語を英語とし、原言語を変える
例:

- ドイツ語 $\rightarrow$ 英語
- スペイン語 $\rightarrow$ 英語

システム構成

- システム: 統計的機械翻訳 [2]
- コーパス: Europarl[1]
  - スペイン語
  - ポルトガル語
  - ギリシャ語
  - ドイツ語
  - フィンランド語
  - デンマーク語

目的

スペイン-英語翻訳 (西英翻訳) & ポルトガル-英語翻訳 (葡英翻訳) に着目
西英翻訳, 葡英翻訳での参照翻訳との一致度合いの比較を行う
テスト・セットには 30,000 文含まれている
このテスト・セットを 2 通りの手法で分割する

- 1 ブロック 300 文の 100 ブロック, 順番に分割
- 100 文ずつ加算していく, 増やす文はランダム (*Broad Sampling*)

300 文/ブロックの場合

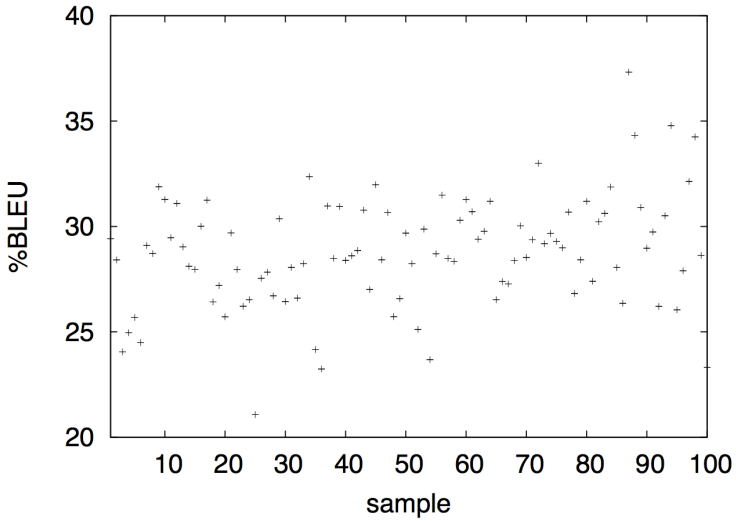

Figure : 各点 300 文の BLEU スコア分布

- BLEU 値はブロックごとに異なる: 21%から 37%の間に分散
- 要因は様々: 未定義語, 文長, 統語度合いなどなど

Broad Sampling

翻訳難度に影響を及ぼす要因

- 書き方スタイルや内容のトピックなどは局所的には類似、大域では大きく異る
- そのような要因が他に影響を及ぼしている

→ よって、テスト・セットの異なる部位から集めたサブ・テストセットを構築

100 文ずつ増加

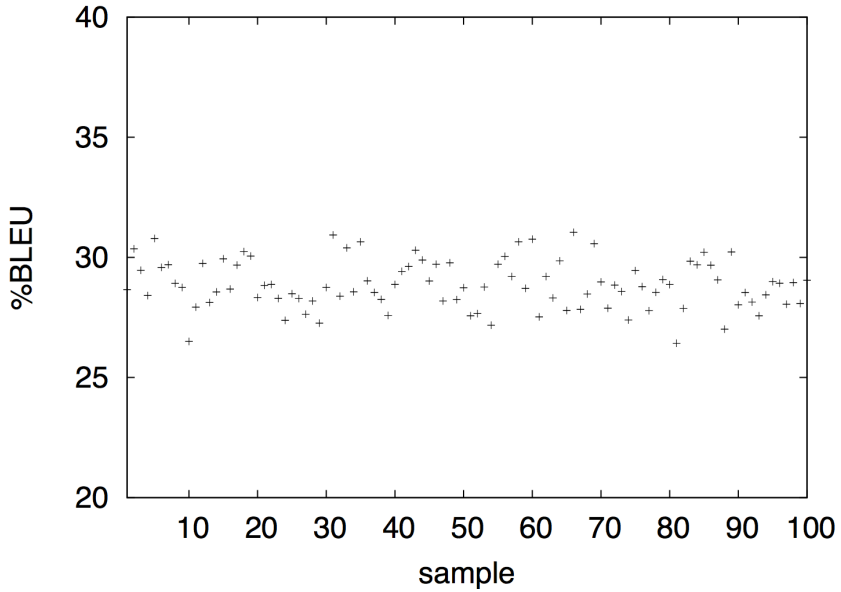

Figure : 出力長さによる BLEU と GTM での影響

- 27% から 31%の間に収まる

システムの比較

スペイン語-英語翻訳 と デンマーク語-英語の比較を行う

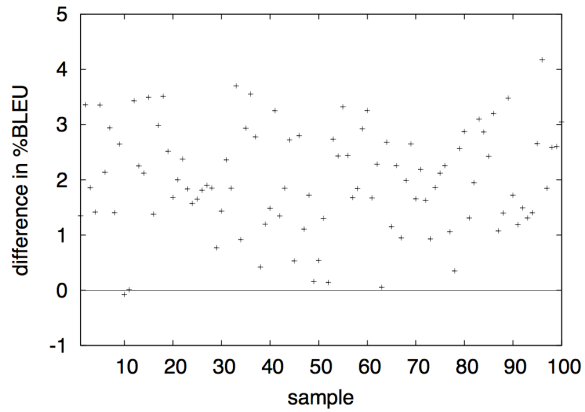

Figure : スペイン語-英語 v.s. デンマーク語-英語

- スペイン語のほうが 4%いい
- デンマーク語一度だけ outperform

このように、小さなテスト・セットでテストする場合は代表サンプルを決めるべきである

Statistical Significance

あるサンプルによる翻訳結果よりシステムの性能評価を行っている
統計的有意性は真の訳質がテストセットの計測値の信頼区間にどこに在るか推定するもの
100 文の翻訳結果を 2 値 (正しい/正しくない) で評価し, 30 文正しいとされた時, これは 30%の性能といえるか？若干の誤差が生じていると考えられるそこで、どのくらいの誤差かを調べる

Bootstrap Resampling

考えは大規模なテストセットから、小さなブロックを抽出し BLEU を計測
これを繰り返し、最後に Top2.5%, Bottom2.5%を切り捨てると信頼区間
[a,b] に在る 95%の BLEU が得られる.

Bootstrap Resampling

仮定

大規模なテストセットから n テストセット文を n 文のテストセットから抽出するのは無限資源から抽出するのと同じである

Bootstrap Resampling Result

300 文/セット を 100 セットに対して行った
1000 回 BLEU を測定し、Top 25 と Bottom 25 を削除した
区間は 26 番目によいから 975 番目までの BLEU 値となる

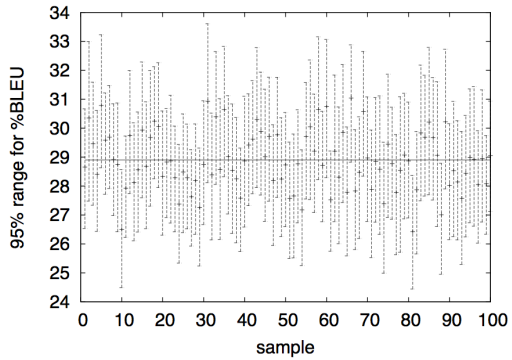

Figure : Bootstrap Resampling Result

30,000 文のテストセットで３文以外は信頼区間に在ることがわかった
95%の信頼レベルで実際は 97%正しかった

References I

P. Koehn.

Europarl: A multilingual corpus for evaluation of machine translation.
In *Unpublished*, 2002.

P. Koehn, F. J. Och, and D. Marcu.

Statistical phrase-based translation.
In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54. Association for Computational Linguistics, 2003.