4. 自動評価尺度 BLEU

内山将夫@NICT
mutiyama@nict.go.jp

自動評価尺度BLEU

Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. (2002) BLEU: a method for Automatic Evaluation of Machine Translation. ACL.

前提：

MT訳とプロの翻訳者による(複数の)翻訳が似ていれば似ているほど，そのMT訳は良いだろう.

自動評価に必要なもの

- (複数の)良質の翻訳
  あらかじめ用意しておく
- 似ている度合を測定する尺度 → BLEU

注：BLEUの妥当性の日本語についての現状

- これからBLEUについて,
- そのMTの評価尺度としての妥当性を検討したものは,
- 中国語–英語，アラビア語–英語が主であり,
- 中日や英日で，しっかりと検証した研究はないようである.
- これについて，我々は,
- NTCIR-7における日英特許翻訳タスクを通じて
- 調査する予定であるが,
- 今のところは，そのような調査結果はないようであるので,
- 英語を例文として利用する.

ngram の重なりによる類似度の例

1番目の良いMT訳の方が，参照訳と共通するngramが多い.

MT訳

1. $_1$It is a guide to action $_2$which $_3$ensures that the military $_4$always obyes the $_5$commands $_6$of the party.
2. It is to ensure the troops forever hearing the activity guidebook that party direct

参照訳

1. $_1$It is a guide to action that $_3$ensures that the military will forever heed Party $_5$commands.
2. It is the guiding principle $_2$which guarantees the military forces $_4$always being under the command $_6$of the party.
3. It is the practical guide for the army $_4$always to heed the directions $_6$of the party.

一般に

1. 多くの ngram を参照訳と共有するMT訳の方が
2. そうでないものよりも良い訳と言えるのではないか

欠点

1. ngram の共有は字面しかみていないので，同義語でも異なるとみなされる．また，活用を考慮していない
2. 語順があまり評価に反映されない

ngramの重なり具合の測り方の悪い例

$$\text{ngram精度} = \frac{\text{参照訳中にあるngram数}}{\text{MT訳中のngram数}}$$

不都合な例

MT: the the the the the the the

Ref1: The cat is on the mat.

Ref2: There is a cat on the mat.

MT訳はtheのみからなり，theはRef1とRef2の双方に出現しているため，上記定義だと

$$\text{1gram精度} = \frac{7}{7}$$

となる．これはおかしい．

修正されたngram精度

$$P_n = \frac{\Sigma \text{ngram ある参照訳での ngram の共有数の最大値}}{\text{MT 訳中の ngram 数}}$$

MT: the the the the the the the

Ref1: The cat is on the mat.

Ref2: There is a cat on the mat.

$P_1 = \frac{2}{7}$
$P_2 = 0$

最初の例での計算

MT1: $P_1 = \frac{17}{18} = 0.94$, $P_2 = \frac{10}{17} = 0.59$

MT2: $P_1 = \frac{8}{14} = 0.57$, $P_2 = \frac{1}{13} = 0.08$

MT訳

1. $_1$It is a guide to action $_2$which $_3$ensures that the military $_4$always obyes the $_5$commands $_6$of the party.
2. It is to ensure the troops forever hearing the activity guidebook that party direct

参照訳

1. $_1$It is a guide to action that $_3$ensures that the military will forever heed Party $_5$commands.
2. It is the guiding principle $_2$which guarantees the military forces $_4$always being under the command $_6$of the party.
3. It is the practical guide for the army $_4$always to heed the directions $_6$of the party.

複数文を翻訳したときの ngram精度

$$P_n = \frac{\Sigma_{\text{MT訳}} \Sigma_{\text{MT訳のngram}} \text{修正された共有ngram数}}{\Sigma_{\text{MT訳}} \Sigma_{\text{MT訳のngram}} \text{ngramの数}}$$

- 分母としては，全てのMT訳における全てのngramの数
- 分子は，各MT訳について，修正したngramの共有数を求めて，それを全MT訳について足したもの

修正ngram精度の統合

$$\sum_{n=1}^{N} \frac{1}{N} \log P_n$$

$P_n =$ 修正ngram精度
$N =$ ngramの最大長 (英語では4が多い)

いくつもの ngram精度を組合せることにより，数値が安定することを意図している.

長さに関しての修正ngram精度P_nの性質

参照訳よりも長いMT訳の場合

共有ngram数は，参照訳にあるngram数を越えないので，参照訳よりも長いMT訳は，P_nが小さくなる．

参照訳よりも短いMT訳の場合

短い訳のngram精度は，高くなる．これは困った．
$P_1 = 2/2,\ P_2 = 1/1$

MT訳: of the

参照訳1: It is a guide to action that ensures that the military will forever heed Party commands.

参照訳2: It is the guiding principle which guarantees the military forces always being under the command of the party.

参照訳3: It is the practical guide for the army always to heed the directions of the party.

短かすぎるMT訳へのペナルティ

MT訳の長さとコーパス中の文長(単語数)の比較

$$c = \sum_{MT\text{訳}} MT\text{訳の長さ}$$

$r = \sum_{\text{参照訳集合}}$ 参照訳中で，対応するMT訳に最近の長さ

コーパス全体で，長さを計算することにより，一文一文の長さの違いには，あまり影響されないようにする．

BP (brevity penalty)

$$\mathrm{BP} = \begin{cases} 1 & \text{if } c \geq r \\ \exp(1 - r/c) & \text{if } c < r \end{cases}$$

- MT訳 $\geq$ 参照訳 のときには，BP $= 1$ (なにもしない)
- MT訳 $<$ 参照訳なら BP < 1 としてペナルティとする

自動評価尺度 BLEU

$$\mathrm{BLEU} = \mathrm{BP} \times \exp(\sum_{n=1}^{N} \frac{1}{N} \log P_n)$$

- 短いMT訳へのペナルティ $\times$
- ngram精度の幾何平均

BLEUにより始めて可能になること

- システムの安価で素早い比較

BLEUにより，開発→評価→開発 ... というサイクルが素早く回りはじめた．最近の統計的機械翻訳進展の原動力の一つである．
現状のMTの研究における，標準的な評価尺度である．

人手による評価とBLEUの比較

人手による評価

- 単言語グループ10人 (英語を母語とするもの)
- 2言語グループ10人 (中国語が母語)
- 各人は，中英翻訳システムが翻訳した500文の英文のうちで 50文を評価
- 各文は，5つのシステム (S1,S2,S3,H1,H2) が英語に翻訳．ただし，H1 とH2は人手による翻訳．H1は日英ともに母語ではない．H2は英語が母語.
- 評価の方法は，1(非常に悪い)〜5(非常に良い) の点を付ける.
- 単言語グループは，出力された英語の読み易さなどのみをチェック
- 2言語グループは，入力中国語と出力英語とを比較してチェック
- 各人が，各文毎に，各システムの評価をするので，システムを比較するときには，同一人の同一文におけるシステム間の評価値の差を求めて，その差が0かどうかにより，検定をする.

評価値の差の検定 (paired t-test)

- ある人u，ある文i，あるシステムsについて，評価値$r(u,i,s)$がある．
- 同じ人と文について，別のシステムs'について，評価値$r(u,i,s')$がある．
- これより，s と s' の評価値の差は $d(u,i,s,s') = r(u,i,s) - r(u,i,s')$である．
- これを全ての人と文について平均すると$m(s,s') = \Sigma_{u,i}\, d(u,i,s,s')$
- 分散は$v(s,s') = \frac{1}{n}\Sigma_{u,i}(d(u,i,s,s') - m(s,s'))^2$である．ただし，$n = \Sigma_{u,i}\, 1$.
- もし，sとs'で評価値に差がなければ，$m(s,s') = 0$なので，
- これが実際に 0 かどうかを調べるために，

$$t = \frac{m(s,s')}{\sqrt{\frac{v(s,s')}{n-1}}}$$

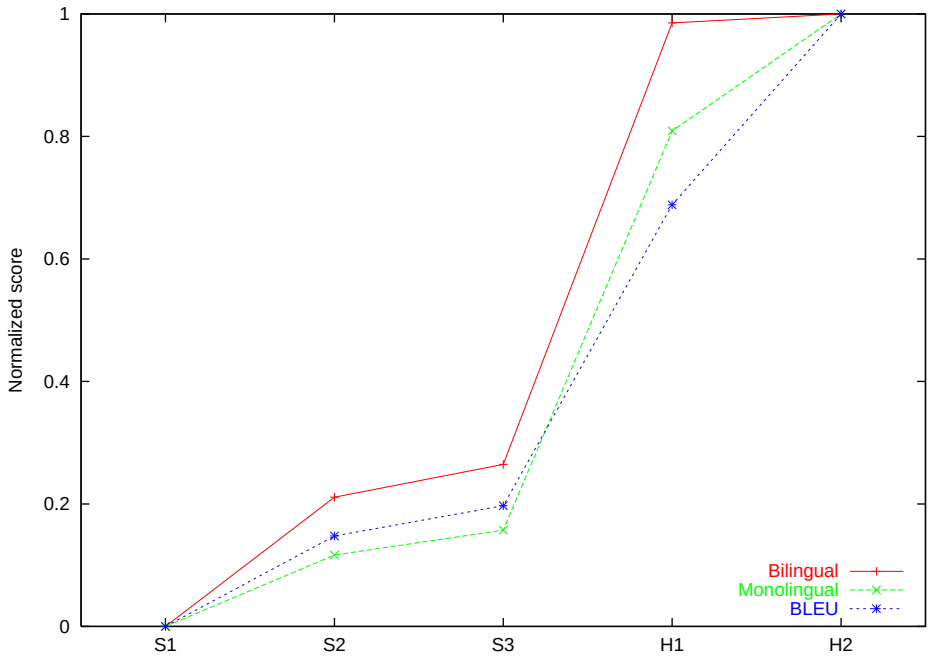
  を計算する．
- t がある値よりも大きければ，統計的に有意差がある．

BLEUおよび人手による評価値

システム	BLEU	単言語	2言語
S1	0.0527	0	0
S2	0.0829	0.326	0.551
S3	0.0930	0.44	0.691
H1	0.1934	2.265	2.574
H2	0.2571	2.8	2.612

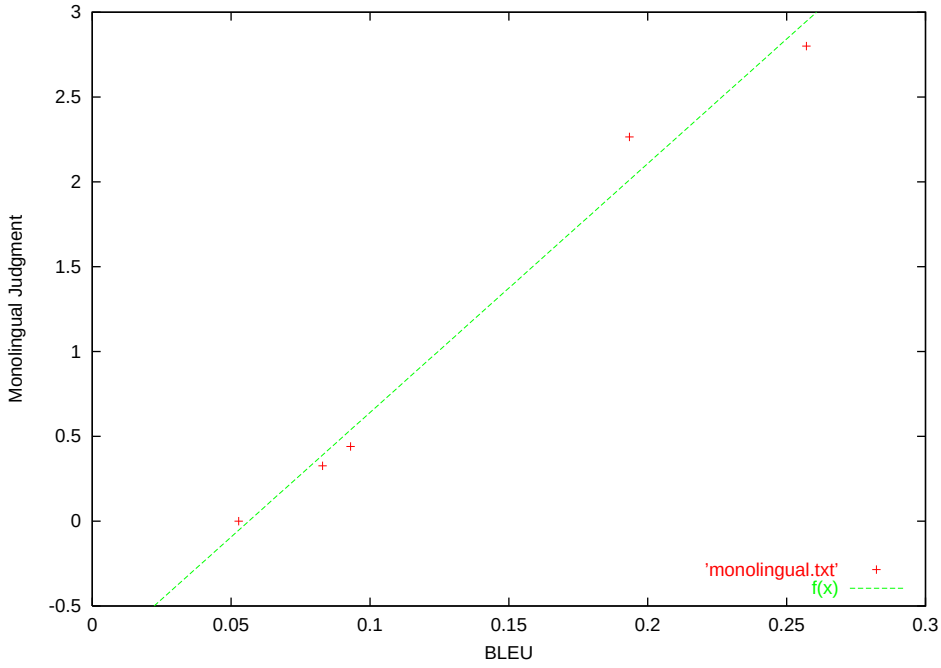
単言語グループと2言語グループにおける数値は，S1の数値を0とし，S2 は，$m(S1, S2)$ を評価値とし，$S3 = S2 + m(S3, S2)$というように，評価値の差に基づいて点数を付けた.

BLEU および人手による評価値



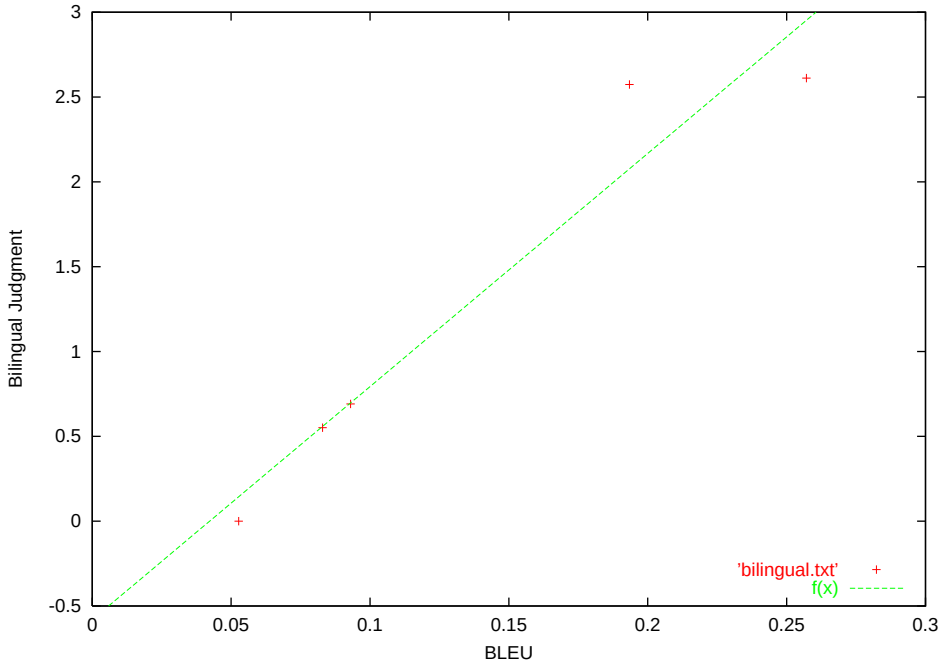
- 縦軸が評価値で，横軸がシステムである
- 0-1 に正規化したスコアを利用している
- BLEU と単言語および 2 言語の評価は似ている
- BLEU は単言語グループの評価に似ている
- S3 と H1 のような大きな差だけでなく，
- H1 と H2，S2 と S3 のような小さい差も検出可能である

BLEU と単言語グループのスコアの比較



BLEU は，単言語グループのスコアと相関が高い

BLEU と単言語グループのスコアの比較



BLEU は，2 言語グループのスコアとも相関が高い．

まとめ

- BLEU は MT 訳と参照訳との類似性を表す尺度である
- BLEU と人手評価との相関は高い

BLEU の欠点

- 意味は同じでも字面が違う ngram とマッチしない
- コーパス全体での値は求められるが文毎の値は求められない

  → どの文が上手く翻訳でき，どの文が翻訳できなかったかがわからない

しかし，現在では，ほぼ全ての研究で利用されている．