第二讲：补充材料2

开放性答案的自动评价策略

基于字符相似度的机器翻译自动评价技术

杨沐昀

语言技术研究中心

计算学部

Motivation

- Why automatic evaluation for MT?
  - Manual evaluation is expensive, inconsistent and time consuming.
  - MT development need instant feedback on his efforts
    - Whether my algorithm, my model, new weight help?
  - Large scale, objective evaluation is of substantial significance for any research.

How ?

- Do we need to study how people recognize good translation?
  - Word, phrase, sentence structure and pattern?
  - A long history of translation argues what is good translation!
- In most cases, “whether better” matters more than “how better”!
- Can we accomplish this by a simple way?

Observations!

- The closer a (machine) translation is to a professional human translation, the better it is!
  - A corpus of good quality human reference translations
  - A numerical translation closeness metric!!!

Examples

Example 1: *Which may be better, intuitively?*

Candidate 1: It is a guide to action which ensures that the military always obeys the commands of the party

Candidate 2: It is to insure the troops forever hearing the activity guidebook that party direct

Reference 1: It is a guide to action that ensures that the military will forever heed party commands

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

Reference 3: It is the practical guide for the army always to heed the directions of the party

Counting Word-match?

- Ranking the candidates
  - Simply comparing the candidate translation and the reference translations and counts the number of matched-word.
- Assumption 1: simple counting method (by unigram word)
  - Counting the number of candidate translation words which occur in any reference translation and then divides by the total number of words in the candidate translation

Exhausted Counting

Example 2：

- Candidate: *the the the the the the the*
- Reference1: *the cat is on the mat.*
- Reference2: *there is a cat on the mat.*

  - Simple standard unigram count is 7/7;
  - Each word should be modified as exhausted after the match identified;
  - Thus, the modified unigram precision is *2/7*;

$i.e.\ Count_{clip}(n\text{-}gram)=2$

Modified Bigram Precision

Example 1:

Candidate 1: It is a guide to action which ensures that the military always obeys the commands of the party

Candidate 2: It is to insure the troops forever hearing the activity guidebook that party direct

Reference 1: It is a guide to action that ensures that the military will forever heed party commands

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

Reference 3: It is the practical guide for the army to heed the directions of the party

- candidate 1 achieves a modified bi-gram precision of 10/17
- whereas the candidate 2 achieves a modified precision of 1/13.

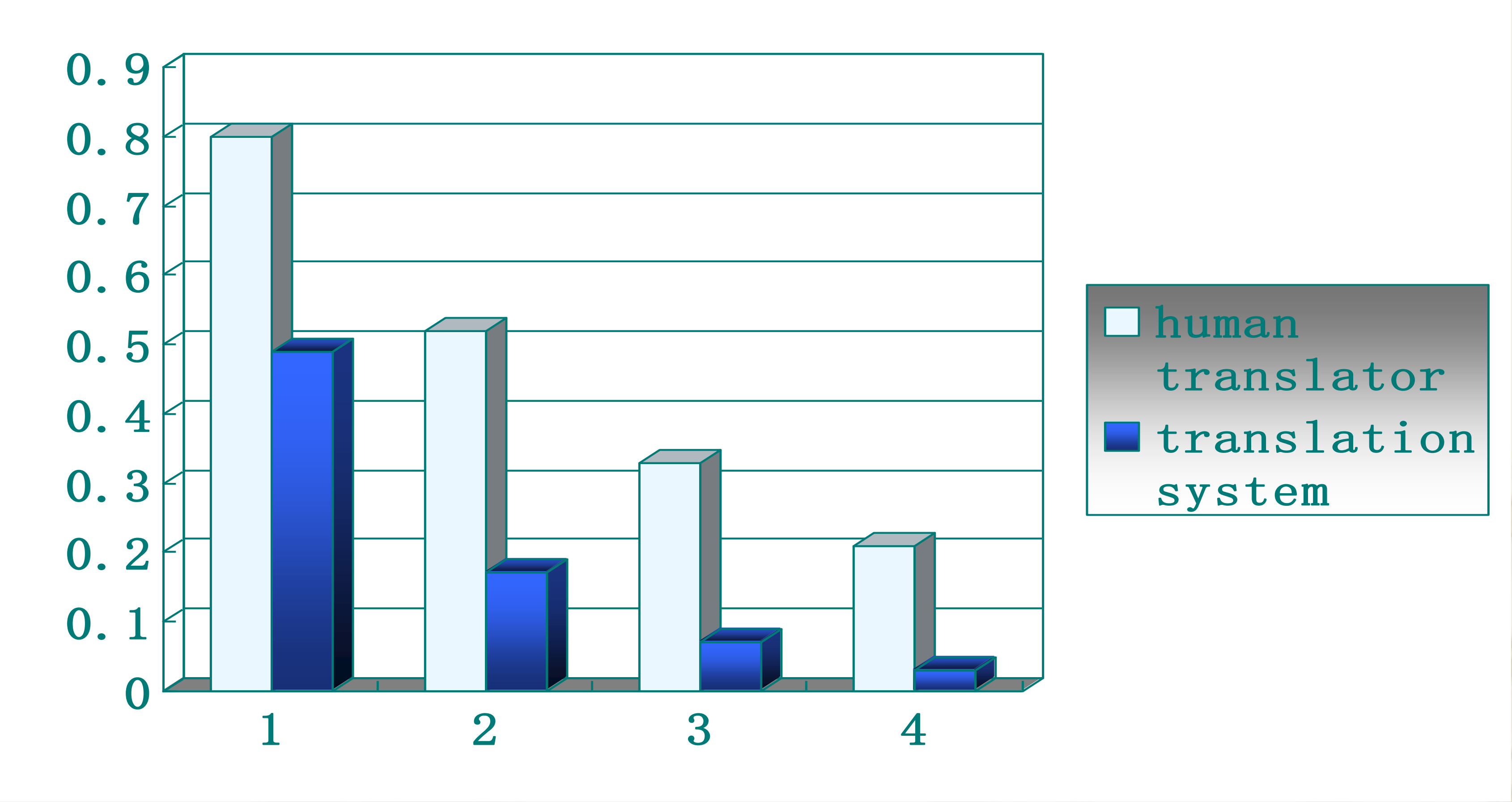
Are We Reasonable

- This sort of modified n-gram precision scoring captures two aspects of translation quality
  - Unigram tends to satisfy adequacy (忠实度)
  - The longer n-gram matches account for fluency (流利度);

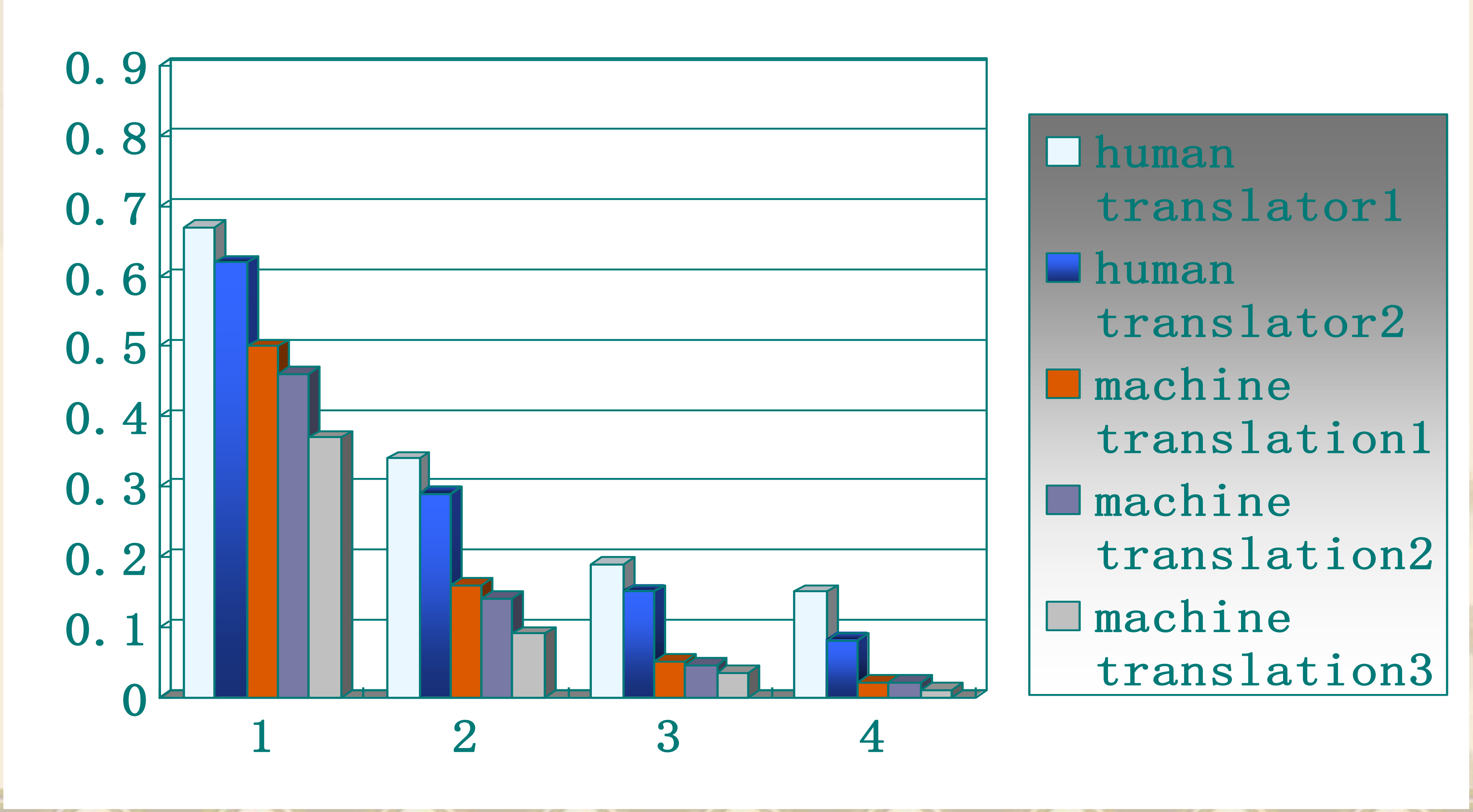
Modified n-gram Precision

$$Pn = \frac{\sum_{C \in \{candidates\}} \sum_{n-gram \in C} Count_{clip}(n-gram)}{\sum_{C \in \{candidates\}} \sum_{n-gram \in C} Count(n-gram)}$$

Compare One Human Translator and One Translation System



Compare More Human Translators and MT Systems



Q1: How to Deal with $P_1 \ldots P_n$

- Shall we choose a best P_i or combine them?
- How to combine: average?
- Note the modified n-gram precision decays roughly exponentially with n:
  - Unigram > Bi-gram > trigram
- How to take account of this?
  - Smooth the sharp difference!

Q2: Sentence Length Issue

- Considering Recall

Candidate1:of the

Reference1: It is a guide to action that ensures that the military will forever heed party commands

Reference2: It is the guiding principle which guarantees the military forces always being under the command of the party

Reference3: It is the practical guide for the army to heed the directions of the party

- Bad recall but: the modified unigram precision is 2/2, and the modified bigram precision is 1/1!

Q2: Sentence Length Issue

- Recall varies in reverse ratio to precision
  - Candidate1: I always invariably perpetually do.
  - Candidate2: I always do.

  - Reference1: I always do.
  - Reference2: I invariably do.
  - Reference3: I perpetually do.
- Note: The recall rate of candidate1 is better than candidate2, but the translation quality is poorer

Solution from Mathematics

- Precision may balance long sentences;
- We may penalize the short ones with a brevity penalty;
- Average logarithm against arithmetic average and geometric mean?
  - Log is a good smoothing function!

BLEU Metric

$$BLEU = BP \bullet \exp\left(\sum_{1}^{N} w_n \log p_n\right)$$

$$BP = \begin{cases} 1 & if\ c > r \\ e^{(1-r/c)} & if\ c \le r \end{cases}$$

$$N = 4, w_n = 1/N$$

BLEU: An Example

- Candidate 1: the book is on the desk

- Ref1: there is a book on the desk
- Ref2: the book is on the table

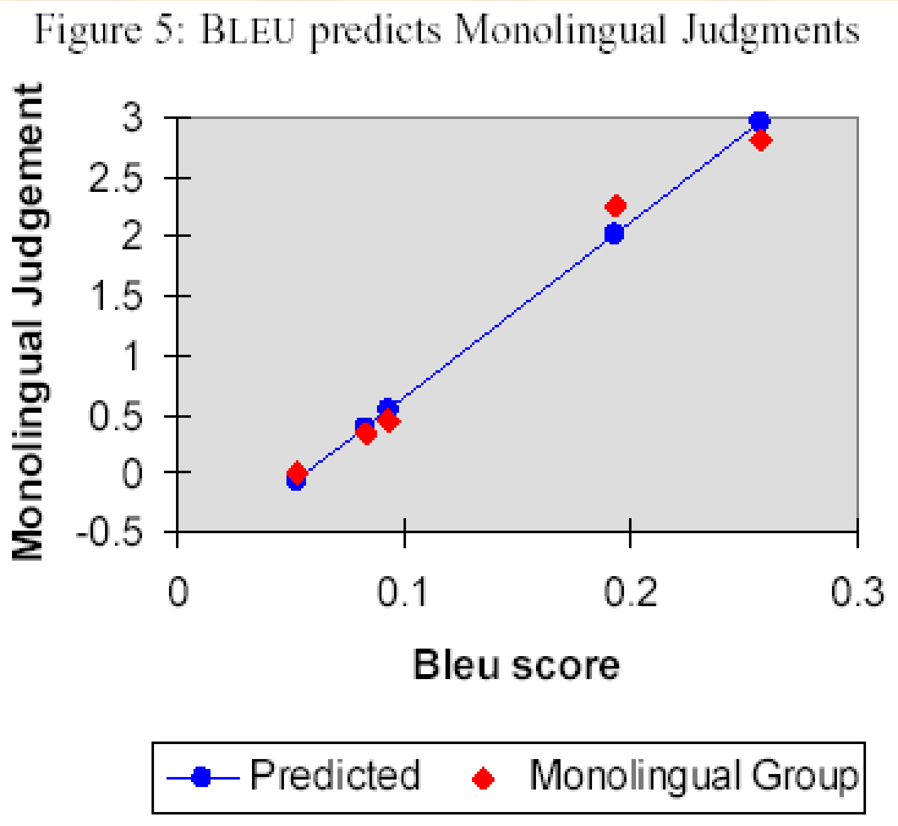
unigram:	bigram:	trigram:
	$Count_{clip}(the, book) = 1$	$Count_{clip}(the, book, is) = 1$
	$Count_{clip}(book, is) = 1$	$Count_{clip}(book, is, on) = 1$
	$Count_{clip}(is, on) = 1$	$Count_{clip}(is, on, the) = 1$
	$Count_{clip}(on, the) = 1$	$Count_{clip}(on, the, desk) = 1$
	$Count_{clip}(the, desk) = 1$	
$\sum_{unigram \in C} Count(unigram) = 6$	$\sum_{bigram \in C} Count(bigram) = 5$	$\sum_{trigram \in C} Count(trigram) = 4$
$p_1 = 1$	$p_2 = 1$	$p_3 = 1$

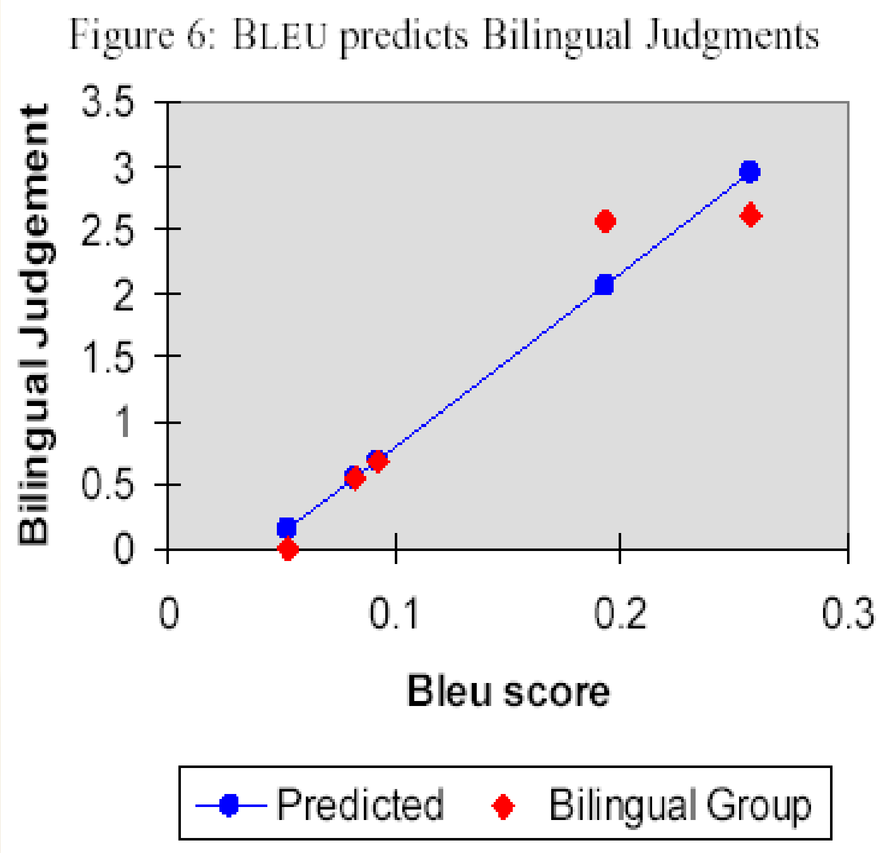
$$\left.\begin{array}{l} c = 6 \\ r = 6 \end{array}\right\} = e^{1-\frac{r}{c}} = e^{0} = 1 = BP$$

$$BLEU = BP \bullet \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

$$= \exp\left[\frac{1}{3}(\log 1 + \log 1 + \log 1)\right] = 1$$

Evaluation BLEU: Consistency





Pearson Correlation Coefficient: 0.99

Adopted by NIST for TIDES Project

- Corpus used to evaluate N-gram Scoring

Corpus	Source language	#of documents	#of human translations	#MT systems
DARPA 1994 French-English	French	100	2	5
DARPA 1994 Japanese-English	Japanese	100	2	4
DARPA 1994 Spanish-English	Spanish	100	2	4
DARPA 2001 Chinese-English	Chinese	80	11	6

Correlation between BLEU Score and Human Assessment

Corpus	Systems	Adequacy (%)	Fluency (%)	Informatics (%)
DARPA 1994 French-English	5 MT systems	95.7	99.7	91.4
DARPA 1994 Japanese-English	4 MT systems	97.8	85.6	98.3
DARPA 1994 Spanish-English	4 MT systems	97.5	97.2	94.3
DARPA 2001 Chinese-English	6 Commercial systems	95.2	97.1	-

Pearson Correlation Coefficient

Outline

- Summary
  - How to processing language by simple method;
  - How to frame your intuition into good formula;
  - Simple->reliable->beautiful

References

- The website for NIST MT Evalution: http://www.nist.gov/speech/tests/mt/index.htm
- *BlEU: a method for automatic evaluation of machine translation,* Kishore Papieni, Salim Roukos, Todd Ward, Wei-Jing Zhu, ACL 2002.

Thanks!