语言之物理特征计算

基于字符相似度的
机器翻译自动评价技术

杨沐昀
哈工大教育部-微软语言语音重点实验室
MOE-MS Joint Key Lab of NLP and Speech (HIT)

Foreword

How to do work simply!

For a given sentence, the easiest thing to do is……….

- Counting
- Counting words
- Can they be helpful?

BLEU method!

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:

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

Match Counting?

- Ranking the candidates
  - Simply comparing the candidate translation and the reference translations and counts the number of matches.
- 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*;

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.

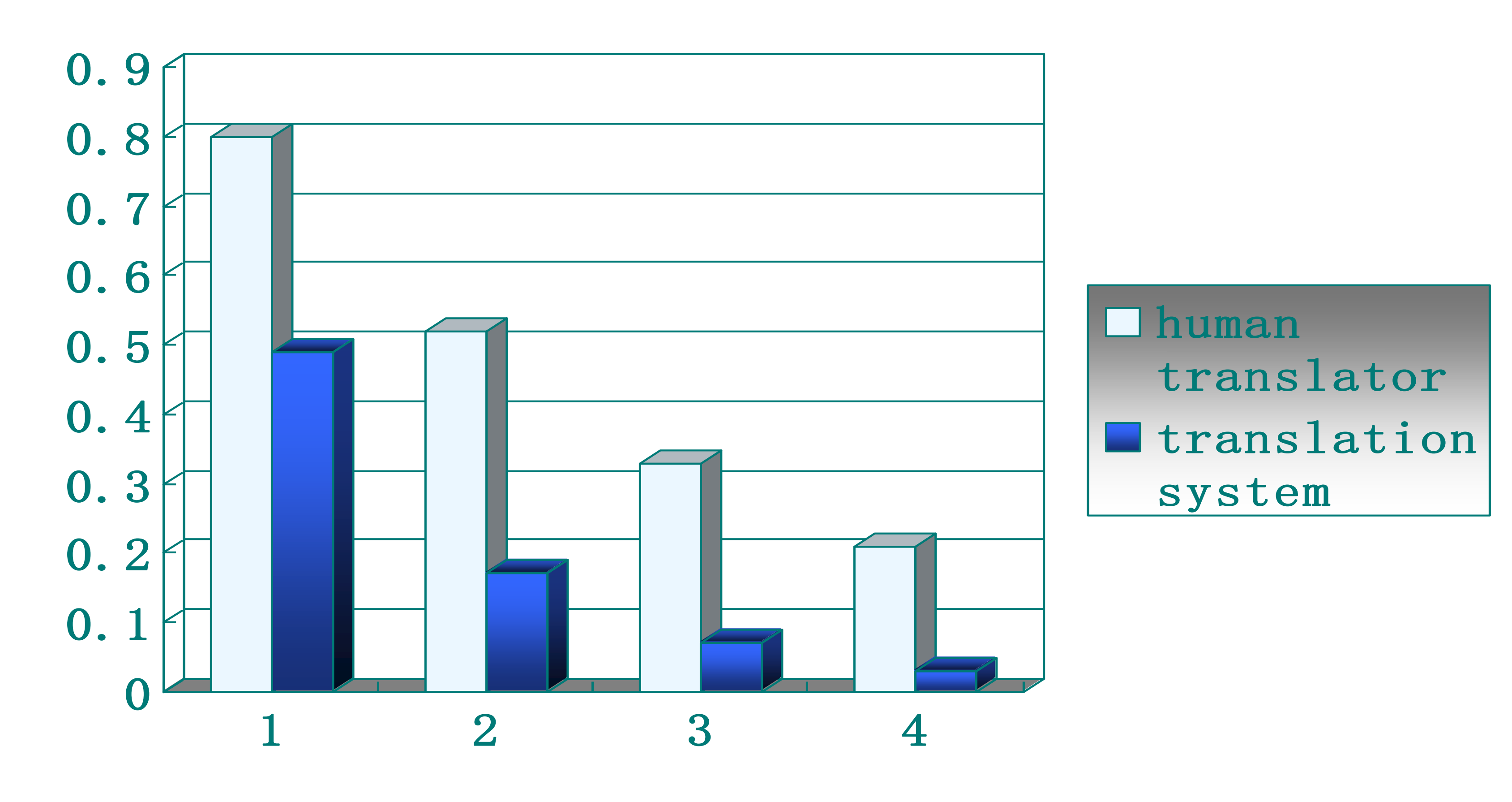
Are We Reasonable

- 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);

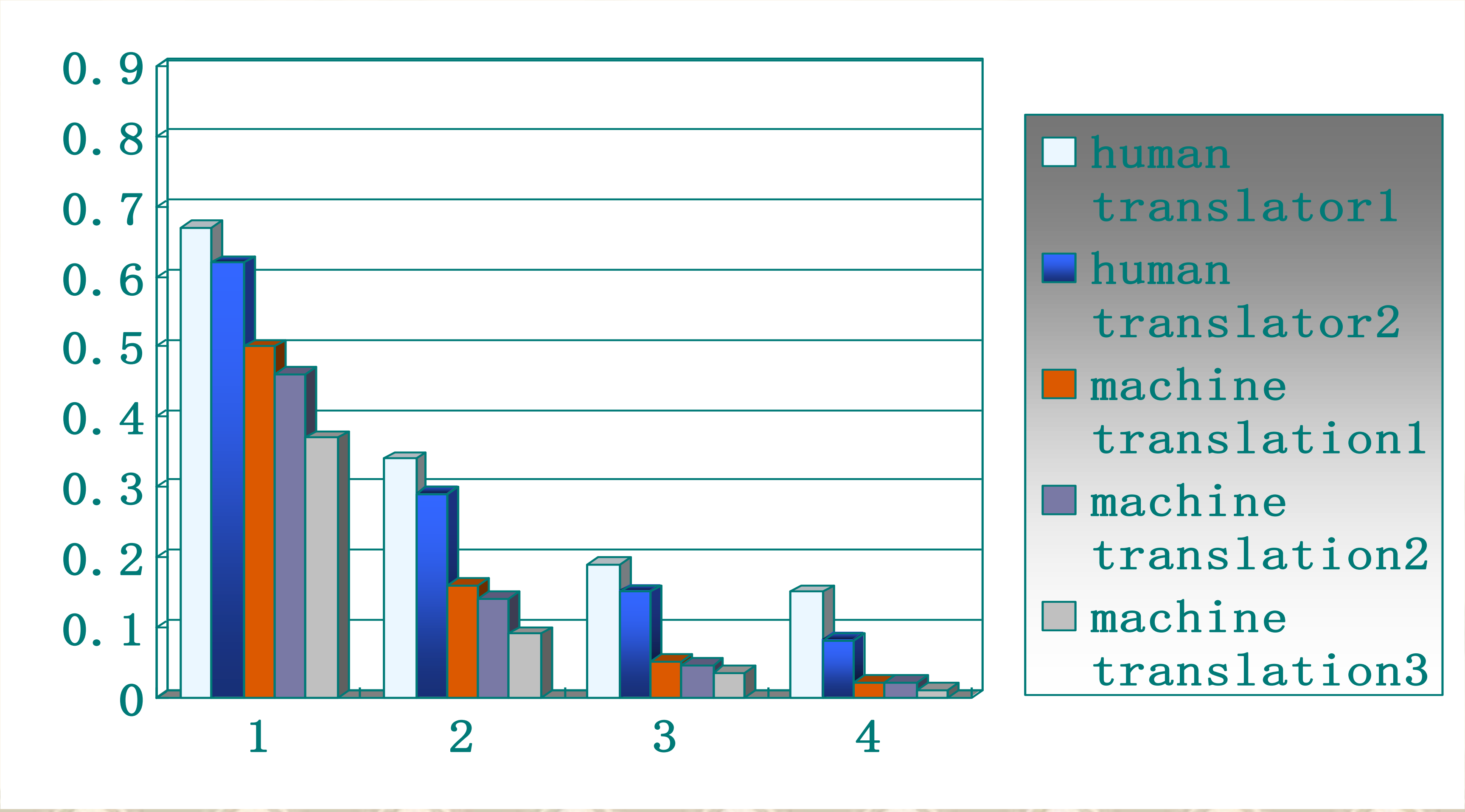
Modified n-gram Precision on Translation Text

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

Compare Human Translator and a Translation System



Compare Multiple Human Translators and MT Systems



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 in average!

Problem: Sentence Length

- Recall Issue

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

- The modified unigram precision is 2/2, and the modified bigram precision is 1/1!

Recall is not an Easy Issue

- 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

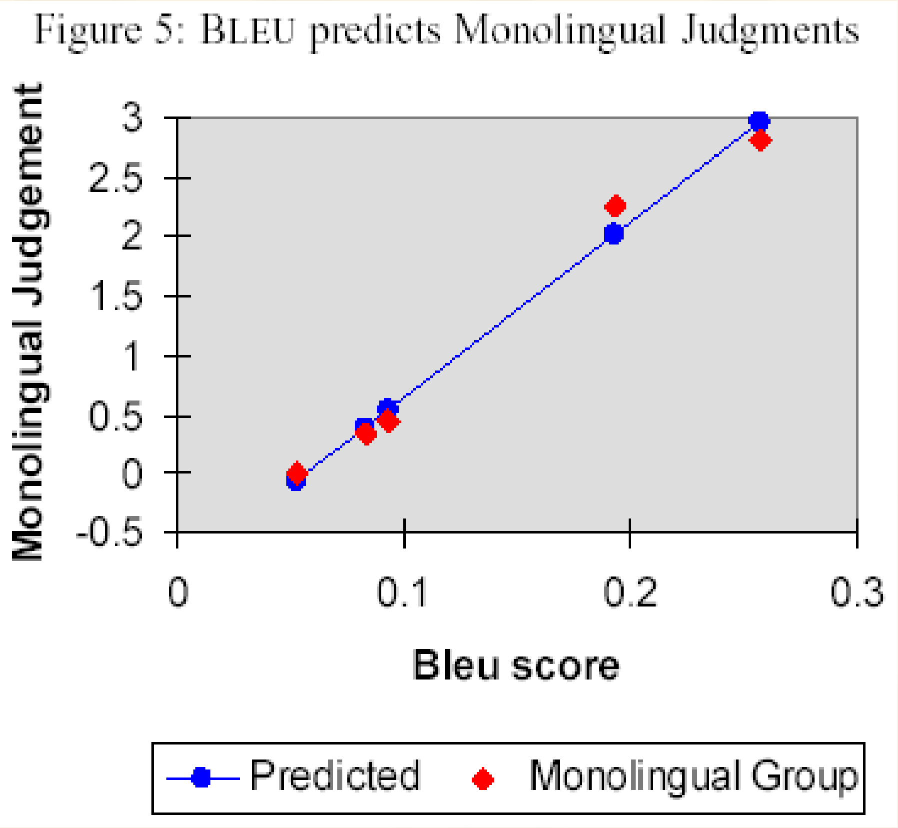
unigram:	bigram:	trigram:
	$Count_{clip}(the, book) = 1$	$Count_{clip}(the, book, is) = 1$
	$Count_{clip}(book, is) = 1$	$Count_{clip}(book, is, on) = 1$
	$Count_{clip}(is, on) = 1$	$Count_{clip}(is, on, the) = 1$
	$Count_{clip}(on, the) = 1$	$Count_{clip}(on, the, desk) = 1$
	$Count_{clip}(the, desk)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

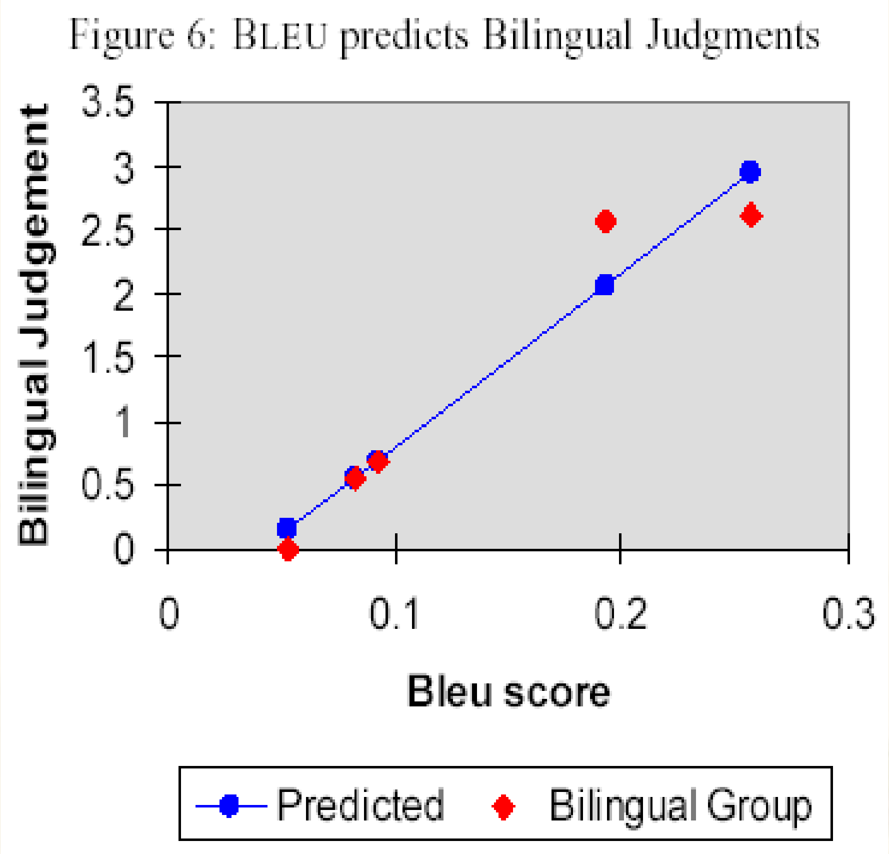
$$\left.\begin{array}{l} c = 6 \\ r = 6 \end{array}\right\} = e^{1-\frac{r}{c}} = e^0 = 1 = BP$$

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

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

BLEU Evaluation--Consistency





Adopted by NIST for TIDES Project

- Corpus used to evaluation of 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 (%) s	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	-

Outline

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

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!

课下深入学习材料

BLEU后的机器翻译自动评价改进
----从专家学习到机器学习

杨沐昀
哈工大教育部-微软语言语音重点实验室
MOE-MS Joint Key Lab of NLP and Speech (HIT)

Where Could BLEU Be Wrong

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

- Parameter?
- Equation?
- Algorithm?
- ….
- ….
- How to do a Systemic Study?

Outline

How to do research by exhaustive study

- Defects in BLEU
  - How to exploit n-gram potentiality?
- NIST Metric: A Weighted Ngram
  - Exhaust the model (formalism)
- Skip Ngram: Get Different Kind of Ngram
  - Exhaust the feature space
- LTE: The Fundamental Unit for Ngram
  - Exhaust the unit space

Defects in BLEU

- Could BLEU Be Wrong?
  - Challenge it by our common sense
  - How about MT and Human Translation
    - which is better?
    - Evaluate human translation?
- Compare Human *vs* MT Translation quality
  - IWSTL2004：506句，每句16个人工标准译文
  - NIST_mt04 :1788句，每句4个人工标准译文
  - 在标准译文中任意选出一个句子作为候选译文，其余15个作为参考译文，使用BLEU工具进行自动评价
  - Guess the Max and Min score!

Defects in BLEU

- MT Beats Human?

BLEU-4	MaxHuman	MinHuman	Total MTsys	MT > Human
IWSLT2004	0.9704	0.2866	20	13
NIST_mt04	0.5064	0.2824	17	2

Defects in BLEU

- To Evaluate Human Translation
  - 数据： 152篇翻译，阅卷点正式评分
    - 某英语水平考试英汉翻译试题：12分
    - 1段英文、3个句子
    - 1个标准译文+3个手工译文
  - Word for Chinese, segmentation or not?

各分数段的样本分布

分数	1	2	3	4	5	6	7	8	9	10	11	12
样本数	9	8	7	6	13	14	16	19	25	16	12	7

Defects in BLEU

- Performance on Human Translations
  - 4个参考译文时与人工评价的相关性较好；
  - 按字计算时与人工评价的相关性较好；
  - 无论是按字匹配，按词匹配，按词性匹配，还是按词与词性同时匹配的， BLEU的性能<0.7

参考译文数	按字	按词匹配	按词性	词+词性
1	0.573	0.539	0.560	0.548
4	0.684	0.624	0.673	0.620

Defects in BLEU

- Analysis
  - Treat word equally
  - Treat n-gram equally
  - Geometric Mean
  - Ngram is not for structure
  - ………
- Solutions?

NIST Metric: A Weighted Ngram

- Background
  - July 2001 TIDES PI meeting in Philadelphia, IBM described BLEU
  - Compare MT output with expert translations in word N-grams
  - Elegant in its simplicity, and strong correlation with human judgment.
  - DARPA commissioned NIST to develop an MT evaluation facility based on the IBM work.

NIST Metric: A Weighted Ngram

NIST's Examination: Bad performance for HT

The Corpus	The Systems	Adequacy (%)	Fluency (%)	Informativeness (%)
1994 French Corpus	5 MT Systems	95.7	99.7	91.4
1994 Japanese Corpus	4 MT Systems	97.8	85.6	98.3
1994 Spanish Corpus	4 MT Systems	97.5	97.2	94.3
2001 Chinese Corpus	6 Commercial MT Systems	95.2	97.1	-
	7 Professional Translators	70.5	16.6	-

NIST Metric: A Weighted Ngram

- Motivation: weight more heavily those N-grams that are more informative

$$Info(w_1 \ldots w_n) = \log_2 \left(\frac{\text{the \# of occurrences of } w_1 \ldots w_{n-1}}{\text{the \# of occurrences of } w_1 \ldots w_n} \right)$$

- How informative n-gram is favored by this?
  - similar to IDF;

NIST Metric: A Weighted Ngram

$$Score = \sum_{n=1}^{N} \left\{ \left[\sum_{\substack{\text{all } w_1 \ldots w_n \\ \text{that co-occur}}} Info(w_1 \ldots w_n) \Bigg/ \sum_{\substack{\text{all } w_1 \ldots w_n \\ \text{in sys output}}} (1) \right] \right\} \cdot \exp\left\{ \beta \log^2 \left[\min\left(\frac{L_{sys}}{\bar{L}_{ref}}, 1 \right) \right] \right\}$$

- β is chosen to make the brevity penalty factor = 0.5 when the # of words in the system output is 2/3rds of the average # of words in the reference translation

NIST Metric: A Weighted Ngram

- Candidate 1: one book is on the desk

- Reference 1: there is a book on the desk
- Reference 2: the book is on the table

unigram:	bigram:	trigram:
$Info(w_i) = \log_2 \frac{\# \; words \; of \; all \; refs}{\# \; occuring \; in \; all \; refs}$	$Info(w_1 \ldots w_n) = \log_2 \frac{the \; \# \; occurences \; of \; w_1 \ldots w_{n-1} \; in \; all \; refs}{the \; \# \; occurences \; of \; w_1 \ldots w_n \; in \; all \; refs}$	
$Info(one) = null$	$Info(one, book) = null$	$Info(one, book, is) = null$
$Info(book) = \log_2 (13/2)$	$Info(book, is) = \log_2 (2/1)$	$Info(book, is, on) = \log_2 (1/1)$
$Info(is) = \log_2 (13/2)$	$Info(is, on) = \log_2 (2/1)$	$Info(is, on, the) = \log_2 (1/1)$
$Info(on) = \log_2 (13/2)$	$Info(on, the) = \log_2 (2/2)$	$Info(on, the, desk) = \log_2 (2/1)$
$Info(the) = \log_2 (13/3)$	$Info(the, desk) = \log_2 (3/1)$	
$Info(desk) = \log_2 (13/1)$		
$\sum_{\substack{all \; unigram in \\ candidate}} (1) = 6$	$\sum_{\substack{all \; bigram in \\ candidate}} (1) = 5$	$\sum_{\substack{all \; trigram in \\ candidate}} (1) = 4$
$\sum Info(w_i) / \sum_{\substack{all \; uni in \\ candidate}} (1) \approx 2.32$	$\sum Info(w_i w_{i+1}) / \sum_{\substack{all \; bigram in \\ candidate}} (1)$ ≈ 0.717	$\sum Info(w_i w_{i+1} w_{i+2}) / \sum_{\substack{all \; trigram in \\ candidate}} (1$ $= 0.25$

NIST Metric: A Weighted Ngram

- Candidate 1: one book is on the desk

- Reference 1: there is a book on the desk
- Reference 2: the book is on the table

brevity penalty:

$$L_{sys} = 6 \qquad \bar{L}_{ref} = 6.5 \qquad \beta = 4.217$$

$$BP = \exp\left\{ \beta \log^2 \left[\min\left(\frac{L_{sys}}{\bar{L}_{ref}}, 1 \right) \right] \right\} = 0.973$$

Score:

$$N = 3$$

$$scorce = \sum_{n=1}^{N} \left\{ \sum Info(w_1 \ldots w_n) \Big/ \sum_{\substack{\text{all } w_1 \ldots w_n \\ \text{in candidate}}} (1) \right\} \bullet BP$$

$$= (2.32 + 0.717 + 0.25) * 0.973$$

$$\approx 3.199$$

NIST Metric: A Weighted Ngram

- Primary Differences with BLEU
  - Geometric mean versus arithmetic sum
  - Uniform weight versus information weight
  - Selective use of n-gram
    - N=4 for BLEU
    - N=5 for NIST

Skip Ngram: Get Different Kind of Ngram

- **Skip N-gram:** any pair of same length n-gram in sentence order, allowing for arbitrary gaps
  - SNR uses the n-gram pair with arbitrary gap **&** Rouge uses the one gram pair with arbitrary gap

E.g. Skip-Ngrams (N <= 2) in "Australia reopens embassy in Manila" :

"Australia, reopens", "Australia, embassy", "Australia, in", "Australia, manila", "reopens, embassy", "reopens, in", "reopens, manila", "embassy, in", "embassy, manila", "in, manila"

"Australia reopens, embassy in", "Australia reopens, in manila", "reopens embassy, in manila"

SNR: Skip *N*-gram Regression

- Skip N-gram Match
  - Partial match: if only one of the two members is matched
  - Full match: if both members are matched
  - Ordered full match: if both members are matched in their sentence order
- Skip N-gram is calculated in recall
  - Partial match recall: M_p
  - Full match recall: M_f
  - Ordered full match recall: M_o

SNR: Skip *N*-gram Regression

- SN score: straight mean of three skip N-gram recall
  - $SN = (M_p + M_f + M_o)/3$
- SNRegression score
  - Assign weights to M_p, M_f and M_o
  - Resolve the weights by SVM
  - Trained by five-fold cross validation on the data provided by NIST Metrics-MATR 2008 evaluation

Source of Data	LDC2008E43
Genre	Newswire
Number of documents	25
Total number of segments	249
Source Language	Arabic
Number of system translations	8
Number of reference translations	4
Human assessment scores	Score 1-7

SNR: Skip *N*-gram Regression

- SNR series are outstanding in all metrics
- SN series exceed all baselines except METEOR
- Comparing with the Rouge, the most similar metric of this work, all novel metrics performs better
- Regression method improves the performance significantly
- Stem is always helpful in this experiment
- Longer gram is not helpful in some cases.

Criterion	Correlation
METEOR	0.705
ROUGE-4	0.654
ROUGE-9	0.663
ROUGE*	0.655
BLEU	0.609
GTM	0.543
SN-1	0.686
SN-4	0.665
SN-1-Stem	0.689
SN-4- Stem	0.670
SNR-1	0.716
SNR-4	0.741
SNR-1- Stem	0.720
SNR-4- Stem	**0.745**

Pearson on segment level

(The SNR scores are the average of 5-fold validation)

Skip Ngram: Get Different Kind of Ngram

- A SVM Regression Based Skip-Ngram Approach to MT Evaluation
  - Longer skipped gram pair.
  - Multiple statistics of skip-Ngram by full, partial and ordered matching
  - Regression to human assessments using multiple statistics as features

LTE: The Fundamental Unit for Ngram

- Method: BLEU counted in letter
- Original BLEU
  - Candidate: As you wish no problem
  - Reference: No problem as you like
- LTE
  - Candidate: A s y o u w i s h n o p r o b l e m
  - Reference: N o p r o b l e m a s y o u l i k e.

LTE: The Fundamental Unit for Ngram

Experiment on training data

LET	1	2	3	4	5	6	7	8
Pear	.264	.548	.656	.682	.687	**.687**	.684	.678
Spea	.484	.609	.674	.700	.711	**.716**	.717	.715

- Why 6-gram is the best?
  - The average word length is 6.

LTE: The Fundamental Unit for Ngram

Compare with other metrics

Correlation	BLEU-4	NIST-5	Meteor0.6	Rouge*
Pearson	0.605	0.735	0.774	0.685
Spearman	0.608	0.686	0.724	0.683

- LTE
  - Pearson: 0.687
  - Spearman: 0.716

LTE: The Fundamental Unit for Ngram

- Advantages
  - Letter match is another resolution to partial word-match;
  - Captures more about translation adequacy
  - Less sensitive to errors of short words; which may be less important;
- Defects
  - Inherited for BLEU framework

常见的研究现状：相关因素庞杂

- Many successful metrics available
    - BLEU/NIST/GTM/Rouge/Meteor……
- How far can we go by exhausting the plain words of the language?
    - Combining metrics to avoid bias
- Linguistic knowledge helpful?
    - POS is the key knowledge at word level

建模策略：常见类型

- Machine Learning Approach
  - Classification: ?
  - Ranking: V
    - SVM Ranking
    - Sensitive but not for robust, diagnostic……
  - Regression?
    - better, but for another metric
    - Robust and generalization can be achieved via multiple features

Case Sutdy：SVM-Ranking for MTE

- Features Selected
  - BLEU1, BLEU2, LetterBLEU
  - Meteor
  - ROUGE-9, recall, case insensitive
- Features Removed
  - BLEU4, NIST5
  - Rouge in other conditions
  - GTM, WER……

Case Sutdy：SVM-Ranking for MTE

- Linguistic info still beneficial?
- How shall we use POS?
  - Each pos is small in quantity in one sentence.
  - Cluster them
- POS Selected
  - Verb: MD,VB, VBD,VBG,VBN,VBP,VBZ
  - Noun: NN,NNS,NNP,NNPS,FW
  - Other: LS, SYM, punctuation
  - Not used: adj, adv, prep……

Case Sutdy：SVM-Ranking for MTE

- In NIST MART Metrics08
  - 10 Top1 out of 144 indexes;
  - Top 2 in weighted score over multiple reference tracks;
  - Top 8 in weighted score over all tracks;
- In WMT 2011
  - 3 champions out of 4, with the last ranked top2.

Case Sutdy：SVM-Ranking for MTE

- Why POS improves performance
  - POS carries word syntax information;
  - POS is a good cluster of words;
  - Smoothing is always helpful in statistics;

这里给出的是BLEU出现后的10年间，改进工作的一个脉络；Evaluation现在的焦点是Quality Estimation，大有可为