Statistical Machine Translation

Marcello Federico
FBK-irst Trento, Italy
Galileo Galilei PhD School – University of Pisa

Pisa, 7-19 May 2008

Part IV: MT Evaluation

- Human vs. machine evaluation
- Human evaluation metrics
- Automatic metrics
- Issues with automatic metrics
- Evaluation Campaigns
- Correlation Human and Automatic Scores
- Outlook

Evaluating MT Performance

How do we evaluate the output of a MT system?

- **Human MT evaluation**:
  - criteria: adequacy and fluency
  - pros: very accurate, high quality
  - cons: expensive and slow
- **Automatic MT evaluation**:
  - criteria: "similarity" to professional human translation
  - pros: inexpensive and quick
  - cons: quality is "slightly" lower than human check

Evaluation bottleneck: MT developers need to monitor the effect of daily changes to their systems in order to weed out bad ideas from good ideas!

Human Evaluation of MT

Let us introduce the Human Assessment Procedure used at LDC in the 2001 Chinese-English track MT evaluation under the DARPA TIDES program.

- A team of English native judges provide multiple assessments of adequacy and fluency of sampled segments of translations of news stories.
- Adequacy assessment: judges compare each segment to a gold standard selected by a bilingual linguist among several human translations.
- Fluency assessment: wrt grammar of Standard Written English and requires no comparison.
- Judges evaluate fluency and adequacy of each translations at once.
- Judges are timed & encouraged to work quickly ($<$ 30"/sentence) and comfortably.
- Assessors are strongly encouraged to provide their intuitive reaction.

LDC Human Evaluation of MT: Fluency

A fluent sentence is one that is well-formed grammatically, contains correct spellings, adheres to common use of terms, titles and names, is intuitively acceptable and can be sensibly interpreted by a native speaker of English.

Possible scores:

1. Incomprehensible
2. Disfluent English
3. Non-native English
4. Good English
5. Flawless English

LDC Human Evaluation of MT: Adequacy

The judge is presented with the gold-standard translation and should evaluate how much of the meaning expressed in the gold-standard translation is also expressed in the output translation.

Possible scores:

1. None
2. Little
3. Much
4. Most
5. All

Requirements for Automatic Metrics

- Low Cost (wrt Human Evaluation)
- Objective (unbiased)
- Informative (for System Developers)
- Efficient

Automatic Evaluation of MT

Automatic scoring methods typically compare the output against multiple high-quality human translations, called references:

- **Word alignment methods**
  - WER: ratio of smallest edit distance and output length
  - SER: 0 if WER is 0, and 1 otherwise
- **N-gram matching methods**
  - BLEU: compute weighted sum of counts of the matching n-grams
  - NIST: modification of BLEU
- **Task completion methods**
  - CLIR: compare IR performance with automatic and manual translations
  - IE: check information extraction performance
  - others

Automatic Evaluation of MT: WER

- Output: `it is a guide to action which ensures that the military always obeys the commands of the party`
- 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`

We can see that the lowest edit distance is with Reference 1.

Automatic Evaluation of MT: WER

Best alignment between Output and Reference 1:

T:	it	is	a	guide	to	action	*which	ensures	that	the	military
R1:	it	is	a	guide	to	action	*that	ensures	that	the	military

T:	*always	*obeys	*the	-	commands	*of	*the	*party
R1:	*will	*forever	*heed	party	commands	-	-	-

The edit distance sums up to: 4 substitutions + 1 deletion + 3 insertions = 8
Hence, the Word Error Rate is $WER = \frac{8}{18} = 0.44$

- WER cannot take into account word re-orderings, e.g. look at the different positions of word `party`.
- WER compares the output with only one reference.

Automatic Evaluation of MT: BLEU

- Rational: the closer MT is to human translation, the better.
- Idea: check matches of words and phrases between
  - one hypothesis (the translation produced by MT) and
  - a set of references (professional human translations)
- Criterion: the more the matches, the better the hypothesis
- Proposed by IBM [Papineni et al., 2001] (name from IBM's company color)
- A numerical measure of closeness between texts
- Needs good quality references to cover linguistic variety
- Not perfect: small changes in the text may determine big changes in the meaning

Important: only the target language is taken into account!

BLEU score: Two Components

- **Modified N-gram Precision**:
  percentage of N-grams in the MT output that occur in references (co-occurrence)
  – matches of shorter N-grams (N=1,2) capture adequacy
  – matches of longer N-grams (N=3,4,...) capture fluency

- **Sentence Brevity Penalty** (rewards Recall):
  penalizes short MT outputs

- **BLEU score is the product** of:
  – the geometric mean of the single n-gram precisions
  – the brevity penalty

BLEU: Modified N-gram Precision

$$PRECISION_{BLEU} = exp\left\{\sum_{n=1}^{N}\frac{1}{N}log(p_n)\right\} \tag{1}$$

where

$$p_n = \frac{\sum_{hypo \in TestSet}\sum_{Ngram \in hypo} Count_{matched}(Ngram)}{\sum_{hypo \in TestSet}\sum_{Ngram \in hypo} Count(Ngram)}$$

$$N = 4$$

Matches at each sentence, score on the entire test set.

BLEU Modified N-gram Precision: an Example

Hypo: `it is a guide to action which ensures that the military always obeys the commands of the party`

Ref1: `it is a guide to action that ensures that the military will forever heed party commands`

Ref2: `it is the guiding principle which guarantees the military forces always being under the command of the party`

Ref3: `it is the practical guide for the army always to heed the directions of the party`

BLEU 1-grams precision: 17/18

Hypo: it is a guide to action which ensures that the military always obeys the commands of the party

Ref1: it is a guide to action that ensures that the military will forever heed party commands

Ref2: it is the guiding principle which guarantees the military forces always being under the command of the party

Ref3: it is the practical guide for the army always to heed the directions of the party

BLEU 2-grams precision: 10/17

Hypo: it is a guide to action which ensures that the military always obeys the commands of the party

Ref1: it is a guide to action that ensures that the military will forever heed party commands

Ref2: it is the guiding principle which guarantees the military forces always being under the command of the party

Ref3: it is the practical guide for the army always to heed the directions of the party

BLEU 3-grams precision: 07/16

Hypo: it is a guide to action which ensures that the military always obeys the commands of the party

Ref1: it is a guide to action that ensures that the military will forever heed party commands

Ref2: it is the guiding principle which guarantees the military forces always being under the command of the party

Ref3: it is the practical guide for the army always to heed the directions of the party

BLEU 4-grams precision: 04/15

Hypo: it is a guide to action which ensures that the military always obeys the commands of the party

Ref1: it is a guide to action that ensures that the military will forever heed party commands

Ref2: it is the guiding principle which guarantees the military forces always being under the command of the party

Ref3: it is the practical guide for the army always to heed the directions of the party

BLEU: Brevity Penalty

$$BP_{BLEU} = \begin{cases} 1 & \text{if } LenHypo > LenRef \\ exp\left(1 - \frac{LenRef}{LenHypo}\right) & \text{if } LenHypo <= LenRef \end{cases} \quad (2)$$

- **Brevity Penalty is calculated over the entire set** (not for each sentence)
- LenHypois the total length of hypothesis
- LenRef is the effective reference length, that is total length of references with closest length to each hypothesis translation (depends on hypothesis!)

BLEU Score Computation

$$BLEU_{score} = BP_{BLEU} * PRECISION_{BLEU} \quad (3)$$

- BLEU ranges from 0 to 1, while BLEU% from 0 to 100
- **The more references, the higher the score**
- **Pros**
  – high correlation with human assigned scores
  – ranking equivalent to human ranking
- **Cons**
  – no co-occurrence of 4-grams (e.g. 4-grams) $\Rightarrow$ score is 0.0
  – longer N-grams dominates shorter N-grams

BLEU limitations: example

Ref:	a b c d e f g h i j k l m n o p q r s
Hyp 1:	a b c d f e g i h j l k m o n p r q s
Hyp 2:	a b c d e f g x x x x x x x x x x x x

	Hyp 1	Hyp 2
1-gram	1.0000	0.3684
2-gram	0.1666	0.3333
3-gram	0.1176	0.2941
4-gram	0.0625	0.2500
BLEU Score	0.1871	0.3083

The NIST score

Proposed by NIST (National Institute of Standard and Technology) in 2002

Rational

- reduce effect of longer N-grams: use arithmetic mean over N-grams counts instead of geometric mean of co-occurrences over N
- weight more heavily the more informative N-grams
- reduce impact of BP: BLEU is very sensitive to variations in translation length

$$NIST_{score} = BP_{NIST} * PRECISION_{NIST} \tag{4}$$

NIST score: Precision

$$PRECISION_{NIST} = \sum_{n=1}^{N} \left\{ \frac{\sum_{all\_w_1 \dots w_n\_that\_co-occur} Info(w_1 \dots w_n)}{\sum_{all\_w_1 \dots w_n\_in\_hypo}(1)} \right\} \quad (5)$$

where

$$Info(w_1 \dots w_n) = -\log_2 \left(\frac{Count(w_1 \dots w_n)}{Count(w_1 \dots w_{n-1})} \right)$$

$$N = 5$$

- $Info$ weights more the words that are difficult to predict
- $Count$ is computed over the full set of references
- Precision range: no theoretical limit, practically $[0..20]$

NIST score: Brevity Penalty

$$BP_{NIST} = exp \left\{ \beta * log^2 \left[min \left(\frac{LenHypo}{LenRef}, 1 \right) \right] \right\} \quad (6)$$

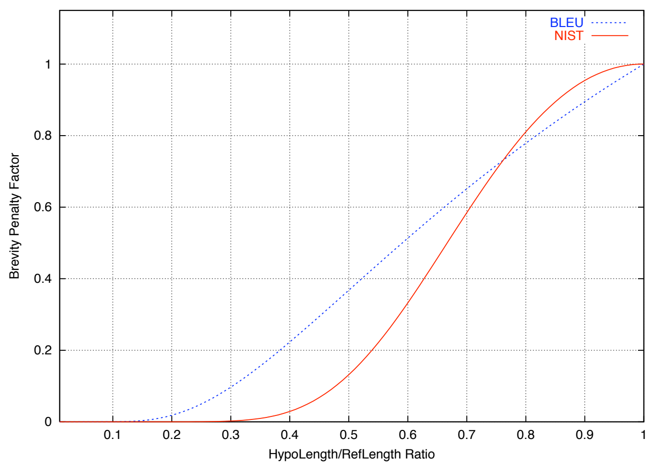
- $LenHypo =$ total length of hypothesis
- $LenRef =$ average length of all references (does not dipend on hypothesis!)
- $\beta = -4.22$, chosen so that $BP = 0.5$ when $LenHypo = 2/3 * LenRef$

Example revised with NIST N-gram Precision

Ref:	a b c d e f g h i j k l m n o p q r s
Hyp 1:	a b c d f e g i h j l k m o n p r q s
Hyp 2:	a b c d e f g x x x x x x x x x x x x

	Hyp 1	Hyp 2
BLEU Score	0.1871	0.3083
NIST Score	4.2479	1.5650

BLEU vs. NIST Brevity Penalty



- BLEU penalizes more than NIST hypotheses slightly shorter than references
- NIST penalizes much more than BLEU very short hypotheses

BLEU or NIST?

- Both scores have shown high correlation with human scores
  - BLEU correlates better with fluency
  - NIST correlates better with adequacy

BLEU or NIST? A Case Study

In CSTAR 2003 Evaluation (Chinese to English) three labs took part:

- CMU (Pittsburgh, USA)
- IRST (Trento, Italy)
- NLPR (Beijing, China)

Results:

	BLEU	NIST
Chi2Eng CMU	0.2733	5.6830
Chi2Eng IRST	0.3884	8.1383
Chi2Eng NLPR	0.5542	3.4013

- BLUE and NIST show the same behaviour for CMU and IRST, but ...
- for NLPR: the highest BLUE and the lowest NIST! Why??

Case Study

Manual inspection outcome:

- NLPR: shorter and more accurate sentences, several empty sentences (== few but precise)
- CMU and IRST: longer and less accurate sentences, no empty sentences (== verbose but imprecise)
- Intrinsecally different approaches used by the NLPR and CMU, IRST
- NLRP: cascade of Example-based MT and Rule-based MT
- CMU, IRST: Statistical MT

Case Study

Effect of NLPR's shorter sentences on the scores through the BP

BLEU

	LenHypo	LenRef	LenHypo/LenRef	BP	Precision	Final Score
CMU	3346	3307	>1	1	0.2733	0.2733
IRST	4047	3549	>1	1	0.3884	0.3884
NLPR	1967	3109	0.63	0.56	0.9896	0.5542

NIST

	LenHypo	LenRef	LenHypo/LenRef	BP	Precision	Final Score
CMU	3346	3421	0.98	0.999	5.6835	5.6830
IRST	4047	”	>1	1	8.1383	8.1383
NLPR	1967	”	0.58	0.29	11.7286	3.4013

- BP reduces NLPR BLUE score to almost 1/2!
- BP reduces NLPR NIST score to less than 1/3!

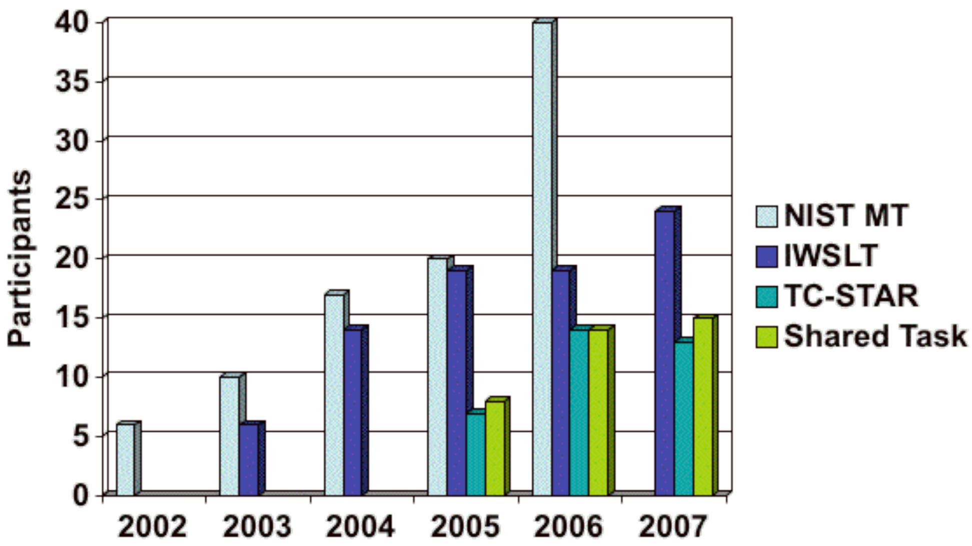
Why Evaluations

- Evaluations started in the ASR community around the 80s
  – controlled experimental setting (LRs, tools)
  – evaluation infrastructure (external organization)
  – goal is to measure progress and compare methods
  – evaluations followed by a workshop

- Open MT evaluations started in 2002 (NIST MT WS)
  – large LRs for statistical MT
  – introduction of automatic scores and subjective evaluations

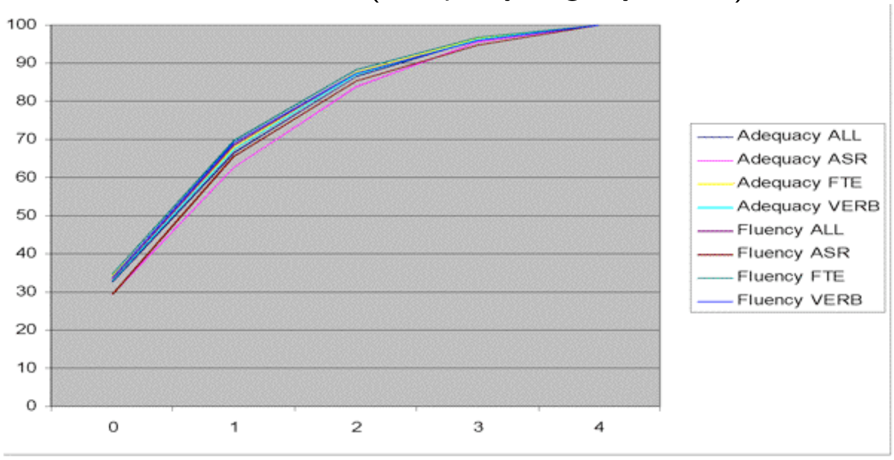
- Today: many open evaluations in many sectors of HLT

Evaluation Campaigns on MT

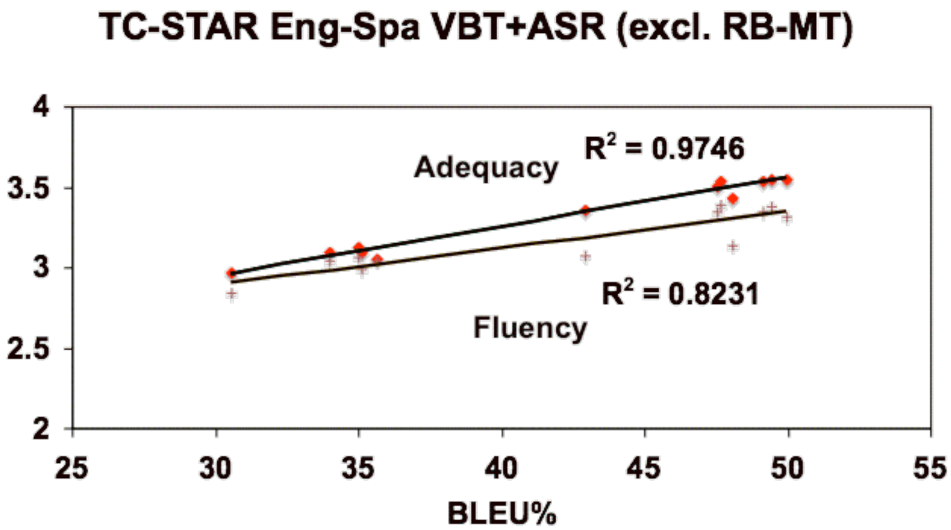


Consistency of Graders

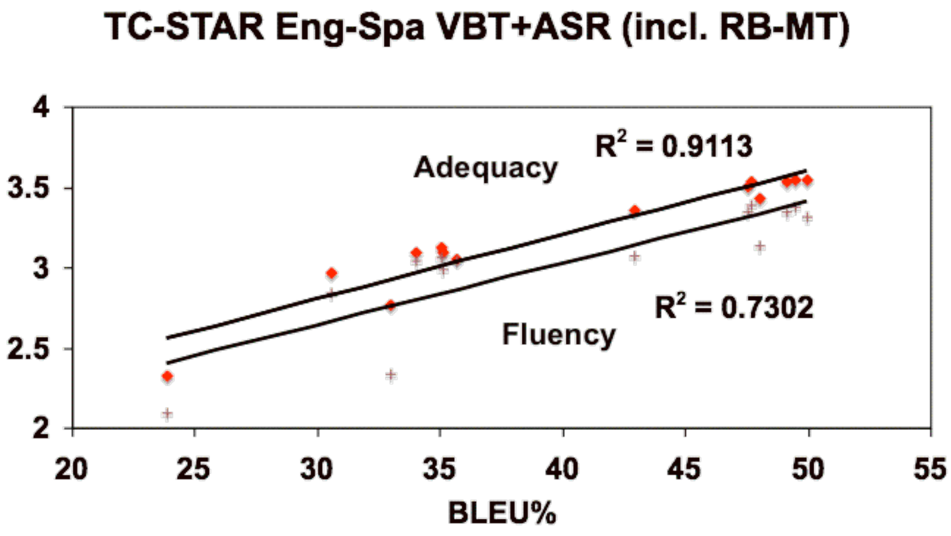
In TC-STAR 2006 Eval, each sentence was evaluated by two graders (tot. 125)
Intra-grader Fluency differences:
– 33% sentences with score $\Delta = 0$
– 65% sentences with score $\Delta \leq 1$ (adequacy slightly worse)



Correlation Subjective-Automatic Score



Correlation Subjective-Automatic Score (2)



Experience with Evaluation

- **Automatic scores** are:
  - Very useful in development cycle of statistical MT systems
  - Useful when comparing different statistical MT systems
  - Useless to compare systems of different nature
- **Subjective scores** are:
  - Very useful to assess general level of performance
  - Useful when comparing systems of different nature
  - Slightly more informative than automatic scores when comparing statistical systems

Outlook: Automatic Scores

- MT research needs **new automatics scores**:
  – Informative: to profile system behavior
  – Discriminative: to tell if and where improvements are
  – Effective: to be computed quickly and often
- We need **more deep insight** into system behavior:
  – More complex and informative benchmarks (used many times)
  – Encourage development of open tools for MT output profiling

Outlook: Human Evaluation

Subjective evaluation should be more efficient:

- Use trained and expert graders only
- Avoid analyzing long (awful) MT outputs
- Focus on specific parts of the sentence:
  – a portion, clause, or syntactic constituent
- Use large test sets to be able to extract interesting parts only
  – count and skip bad translations, don't waste time

This may require **re-thinking the whole evaluation protocol**