Machine Translation IV: MT Evaluation

January 23, 2008

Andreas Eisele
UdS Computerlinguistik & DFKI
eisele@dfki.de

Language Technology I
WS 2007/8

MT Evaluation

UNIVERSITÄT DES SAARLANDES

Overview

- Reasons for MT evaluation
- Manual error analysis
- Automatic error analysis
- A closer look at the BLEU score

Reasons for MT Evaluation

UNIVERSITÄT DES SAARLANDES

"More has been written about MT evaluation over the past 50 years than about MT itself"

[Y. Wilks, according to Hovy e.a.]

MT evaluation may serve different purposes

It may help to decide

- whether to apply MT at all
- which of a set of systems to use for a given task
- which problems/error to focus on in further development of one system
- how to combine systems in a hybrid architecture

Types of MT Evaluation

UNIVERSITÄT DES SAARLANDES

- Relative vs. absolute evaluation
  - which system is better? vs.
  - rate system X on a scale from 0 (useless) to 100 (perfect)

- Adequacy evaluation
  - will system X fit a given purpose?
- Task-based evaluation
  - can users of system X achieve a given task?
- Diagnostic evaluation
  - which phenomena are/aren‘t handled correctly?
- Performance evaluation
  - measure performance in specific areas in more detail

- Black-Box vs. Glass-Box
  - does evaluation see only in-/output or also the internal representations?

Subjective Evaluation

UNIVERSITÄT DES SAARLANDES

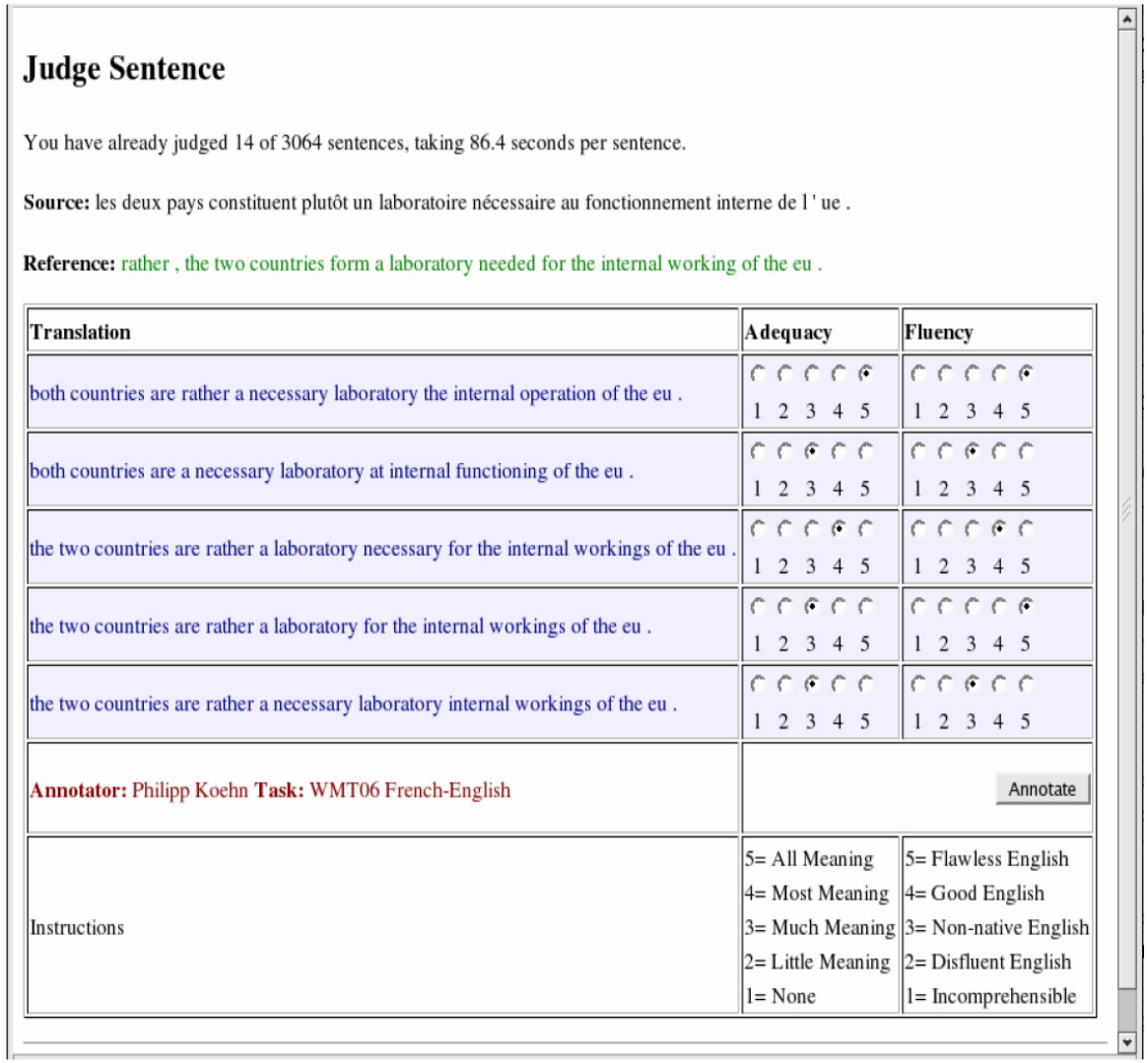
Main focus traditionally on two aspects:

- Adequacy
  - „Is the output equivalent to the input“ (in what sense?)
- Fluency
  - „Is the output well-formed in the target language?“

Subjective MT Evaluation in Practice

UNIVERSITÄT DES SAARLANDES

Koehn/Monz 2006 distributed the burden of manual evaluation over the participants in the shared MT task, using a web-based evaluation interface



Judge Sentence

You have already judged 14 of 3064 sentences, taking 86.4 seconds per sentence.

Source: les deux pays constituent plutôt un laboratoire nécessaire au fonctionnement interne de l ' ue .

Reference: rather , the two countries form a laboratory needed for the internal working of the eu .

Translation	Adequacy	Fluency
both countries are rather a necessary laboratory the internal operation of the eu .	1 2 3 4 5	1 2 3 4 5
both countries are a necessary laboratory at internal functioning of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a laboratory necessary for the internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a laboratory for the internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a necessary laboratory internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
Annotator: Philipp Koehn **Task:** WMT06 French-English		Annotate
Instructions	5= All Meaning 4= Most Meaning 3= Much Meaning 2= Little Meaning 1= None	5= Flawless English 4= Good English 3= Non-native English 2= Disfluent English 1= Incomprehensible

Problems of Subjective Evaluation

UNIVERSITÄT DES SAARLANDES

- Task is very tedious
- Inter-annotator agreement could be better
- Long sentences are particularly hard to judge
- Linguistic expertise of the evaluators not exploited

Manual Error Analysis

Human evaluators may give more specific diagnosis of problems [Vilar e.a. 2006]

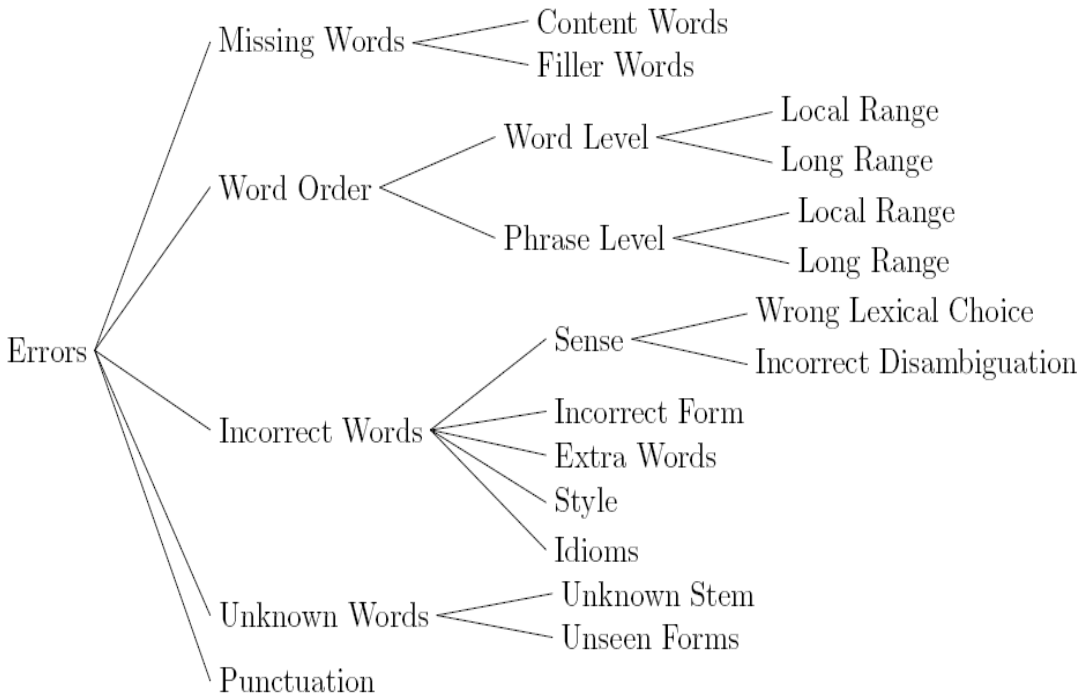


Figure 1: Classification of translation errors.

Automatic Scoring of MT Quality

UNIVERSITÄT DES SAARLANDES

Main Idea:

Given a “good” (reference) translation, quality of machine translation output boils down to the question of similarity

This is a monolingual problem, may be easier than the original question

Textual similarity may be measured automatically

Various simple error metrics have been successfully used in speech recognition (Word error rate, …)

Evaluation for SMT development

UNIVERSITÄT DES SAARLANDES

Development cycle of an SMT system [Och 2000]

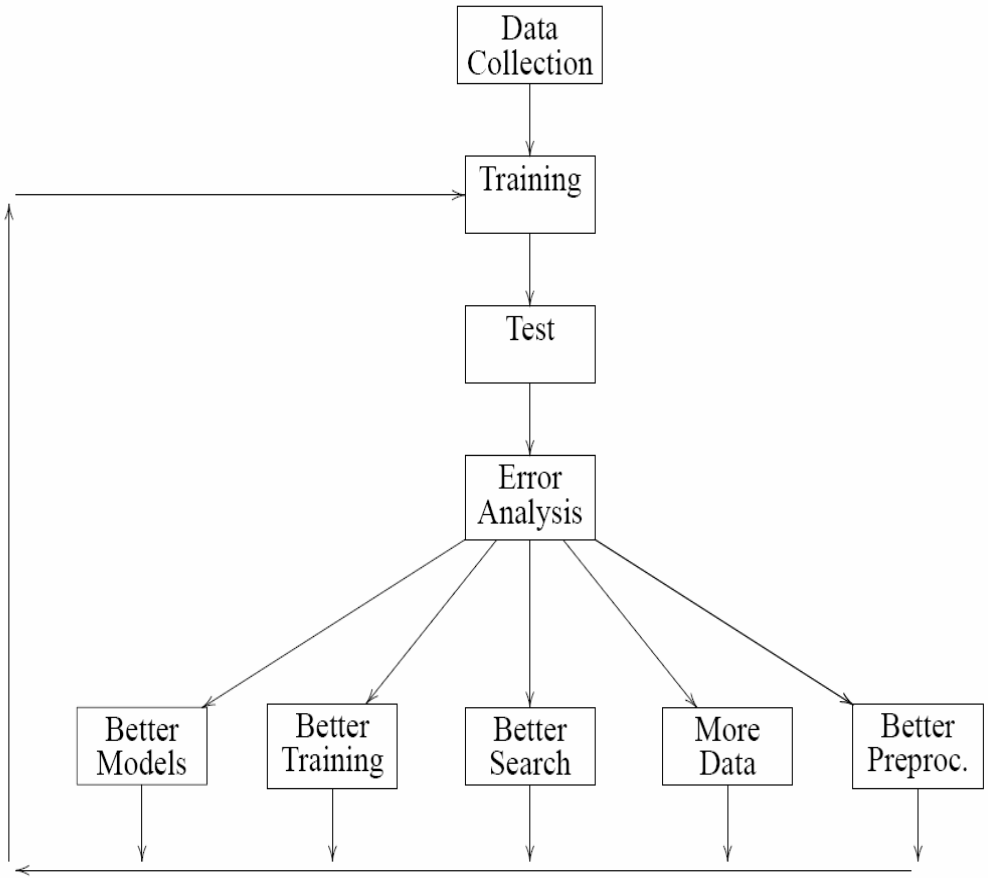


Figure 3.1: Development cycle of a statistical MT system.

The BLEU score

UNIVERSITÄT DES SAARLANDES

BLEU = Bilingual Evaluation Understudy

Goals:

- Measure the similarity of an MT result with reference translation(s)
- Can deal with multiple reference translations
- Take word order into account (more informed than position-independent word error rate)
- Allow for major reordering (less strict than word error rate/ Levenshtein distance)

Main ideas:

Combine n-gram **precision** for multiple n (typically 1..4)

Approximate **recall** via so-called **brevity penalty**

BLEU score

UNIVERSITÄT DES SAARLANDES

See http://www1.cs.columbia.edu/nlp/sgd/bleu.pdf for details, the main formulas are as follows:

We first compute the geometric average of the modified n-gram precisions, p_n, using n-grams up to length N and positive weights w_n summing to one.

Next, let c be the length of the candidate translation and r be the effective reference corpus length. We compute the brevity penalty BP,

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

Then,

$$\textsc{Bleu} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right).$$

The ranking behavior is more immediately apparent in the log domain,

$$\log \textsc{Bleu} = \min(1 - \frac{r}{c}, 0) + \sum_{n=1}^{N} w_n \log p_n.$$

In our baseline, we use $N = 4$ and uniform weights $w_n = 1/N$.

See http://www.statmt.org/wmt06/shared-task/multi-bleu.perl for a practical implementation.

Why BLEU is popular

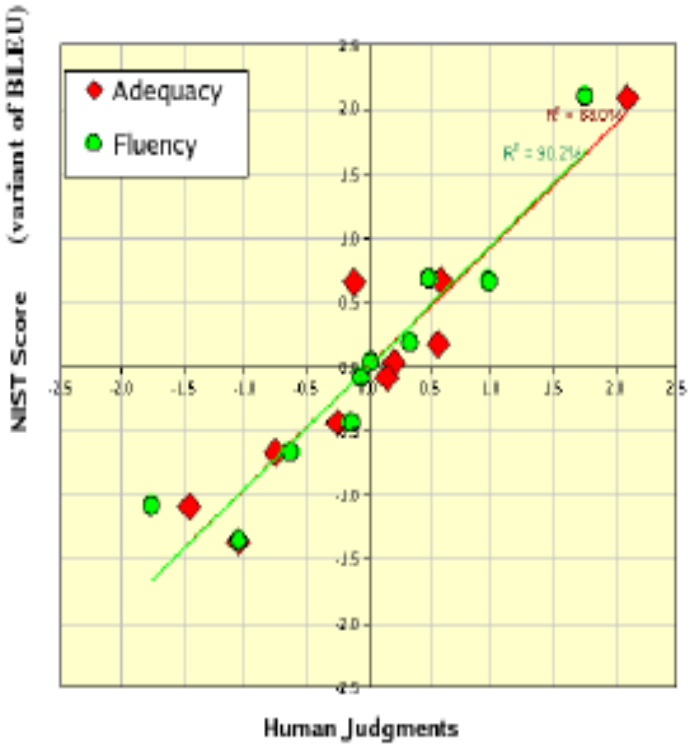


Figure 8.8: Correlation between an automatic metric (here: NIST) and human judgment (fluency, adequacy). Illustration by George Doddington.

From http://cio.nist.gov/esd/emaildir/lists/mt_list/msg00065.html

Why BLEU is controversial

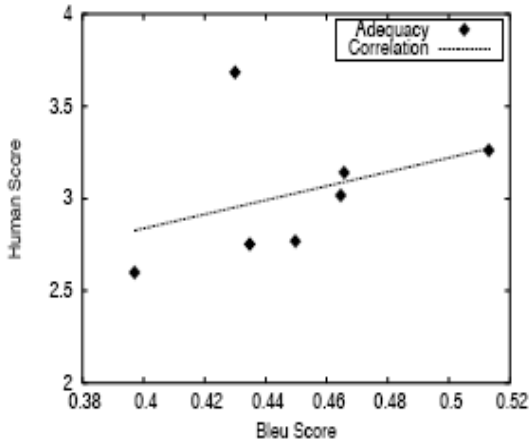


Figure 2: Bleu scores plotted against human judgments of adequacy, with $R^2 = 0.14$ when the outlier entry is included

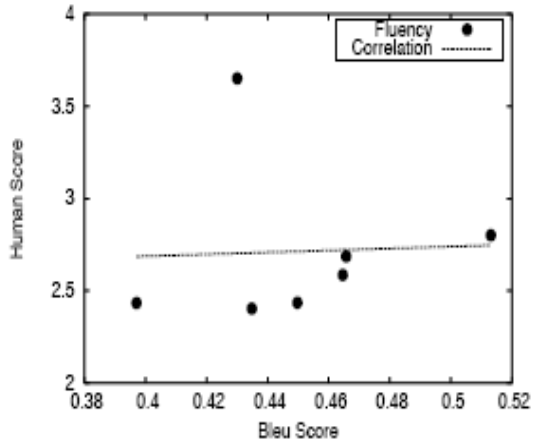


Figure 3: Bleu scores plotted against human judgments of fluency, with $R^2 = 0.002$ when the outlier entry is included

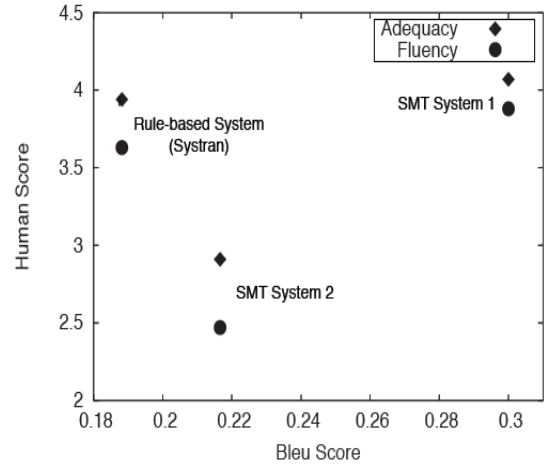


Figure 4: Bleu scores plotted against human judgments of fluency and adequacy, showing that Bleu vastly underestimates the quality of a non-statistical system

From: Re-evaluating the Role of BLEU in Machine Translation Research, Chris Callison-Burch, Miles Osborne, Philipp Koehn, EACL 2006
http://www.iccs.inf.ed.ac.uk/~pkoehn/publications/bleu2006.pdf

BLEU Score Example

UNIVERSITÄT DES SAARLANDES

Three machine translation systems generate the following texts:

SYS1: She cannot be used as a basis for the installation of a European constitution .
SYS2: It cannot a basis for the establishment of a European constitution .
SYS3: It can form the basis for a European constitution .

Assume that for automatic evaluation we also have access to the following two reference translations:

REF1: It cannot serve as a basis for the establishment of a European constitution .
REF2: It can not serve as a basis for the introduction of a European constitution .

Sketch how the BLEU-4 score for the given translation candidates will be computed. What are the 1- ... 4-gram accuracies that will enter into the computation? Insert appropriate numbers into the slots in the following lines. You do not need to compute the brevity penalty for this exercise.

	1-grams	2-grams	3-grams	4-grams
SYS1:	__/15	__/14	__/13	__/12
SYS2:	__/12	__/11	__/10	__/9
SYS3:	__/10	__/9	__/8	__/7

Evaluation of MT systems

UNIVERSITÄT DES SAARLANDES

Two types of MT evaluation

- Human („subjective“)
- Automatic („objective“)

The evaluation dilemma:

- Manual evaluation is meaningful, but expensive, tedious, and error-prone, not useful for regression testing
- Automatic evaluation is repeatable, objective, but not necessarily relevant; better systems may have worse scores

We need to

- lower the effort for manual evaluation,
- increase the quality of automatic evaluation,
- or do both