Machine Translation

June 13, 2013

UNIVERSITÄT DES SAARLANDES

Christian Federmann
Saarland University
cfedermann@coli.uni-saarland.de

Language Technology II
SS 2013

Evaluation of Translations

- We want translations that are:
  - *equivalent* in meaning to the source text
  - *fluent* in the target language
- Evaluation is:
  - comparing source text and translation
  - examining translation
  - checking the MT system to find out where errors come from

Requirements

- What do we need for evaluation?
  - Source text
  - Translation
  - Reference (sample translation)?

- Who should evaluate?
  - Linguists?
  - Professional translators?
  - Anyone who knows both source and target language?
  - Speakers of the target language?

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 et al.]

- 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

Evaluation for SMT development

Development cycle of an SMT system [Och 2000]

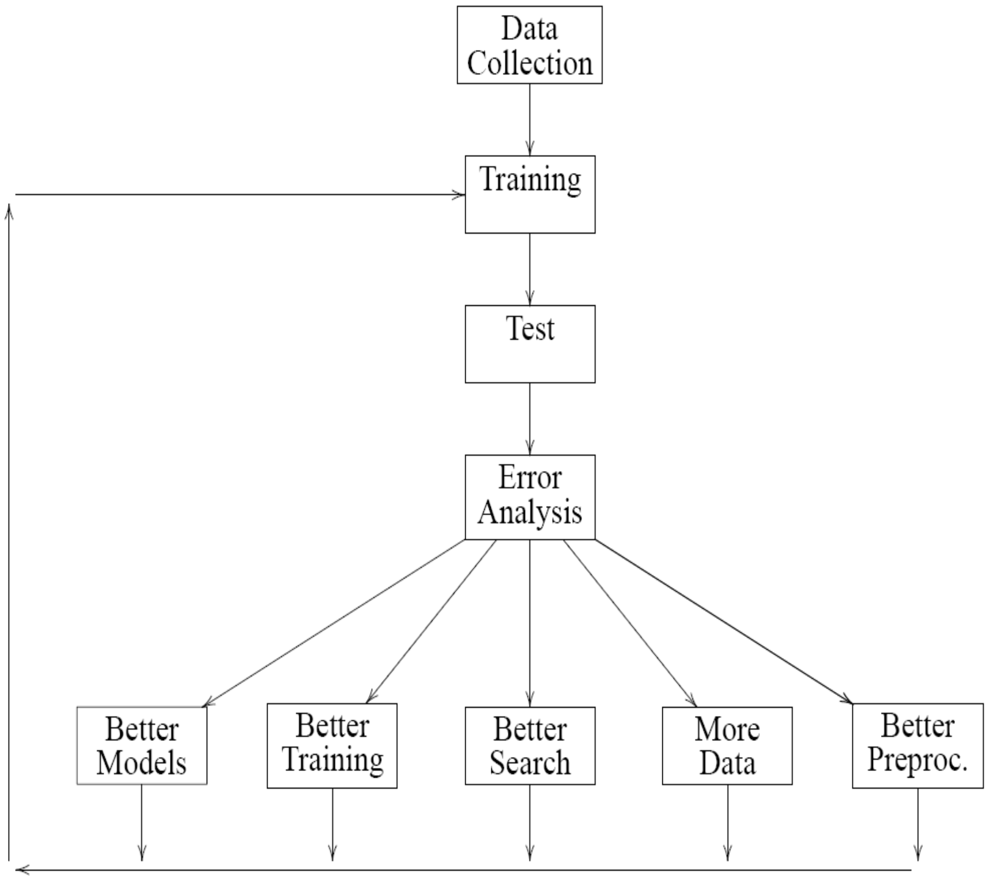

Figure 3.1: Development cycle of a statistical MT system.

Evaluation of MT systems

- Two types of MT evaluation (with different requirements):
  - Manual („subjective“)
  - Automatic („objective“)
- Manual evaluation requires a certain amount of knowledge (of the source/target language, of linguistics, …).
- Automatic evaluation requires a reference translation to compare the translation to.

The Evaluation Dilemma (I)

- Manual evaluation is:
  - meaningful

    We get error types that we can re-use.
  - expensive

    Requires expert knowledge & takes some time to complete.
  - tedious

    Errors might be repetitive/very common.
  - error-prone

    Different evaluators use different scales.
  - not useful for regression testing

    Too expensive to run for many tasks.

The Evaluation Dilemma (II)

- Automatic evaluation is:
  - repeatable

    Each run gets the same result.
  - objective

    Only based on reference translation(s), doesn't take into account personal preferences.
  - not necessarily relevant

    What does an automatic score mean?
  - → better systems may have worse scores
  - → rule-based systems are usually punished by automatic scores

The Evaluation Dilemma (III)

- We want reliable, meaningful results in a quick turnaround.

- We need to
  - lower the effort for manual evaluation,
  - increase the quality of automatic evaluation,
  - or do both.

Types of Manual MT Evaluation (I)

- Absolute evaluation
  - Only looks at one system at a time
  - Rate system X on a scale, e.g. from 0 (useless) to 10 (perfect)
- Relative evaluation
  - Compares up to *n* systems
  - Rank systems 1 to n (with/without ties allowed)
- Adequacy evaluation
  - Purpose: assimilation/dissemination, …
  - Will system X fit a given purpose?

Types of Manual MT Evaluation (II)

UNIVERSITÄT DES SAARLANDES

- Task-based evaluation
  - Can users of system X achieve a given task?
  - Difference to adequacy: task is clearly defined, i.e. answer questions based on translation
- Diagnostic evaluation
  - Which phenomena are/aren‘t handled correctly?
  - Requires expert knowledge
- Performance evaluation
  - Measure performance in specific areas in more detail
  - Difference to diagnostics: less concerned with finding out why something was translated incorrectly

Types of Manual MT Evaluation (III)

- Black Box vs. Glass Box
  - Black Box: we only see input and output
  - Glass Box: we have access to the internal representations in the system (search graph, analysis trees, …)
  - We can evaluate only the output
  - We can evaluate all intermediary steps (lexicon entries, analysis tree(s), transfer rules, phrase table, language model, search graph, …)
- Most RBMT systems are black boxes, but here we could get a lot of information from the intermediary steps.
- SMT systems are mainly open source, but evaluating a search graph?

Manual Evaluation

UNIVERSITÄT DES SAARLANDES

- To get fast results, usually use ranking tasks.

- Either split up adequacy and fluency, or have only one score for both?

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 Manual Evaluation

- Task is very tedious:
  - You always need to compare all *n* translations with each other
  - How do you weigh problems in different parts of the sentence?
- Long sentences are particularly hard to judge.
- Interannotator agreement could be better:
  - Different evaluators have different (internal) guidelines.
  - If we publish guidelines, we get more streamlined results, but we also lose information.
- Linguistic expertise of the evaluators not exploited:
  - You don’t say *why* system X is best.

Manual Error Analysis

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

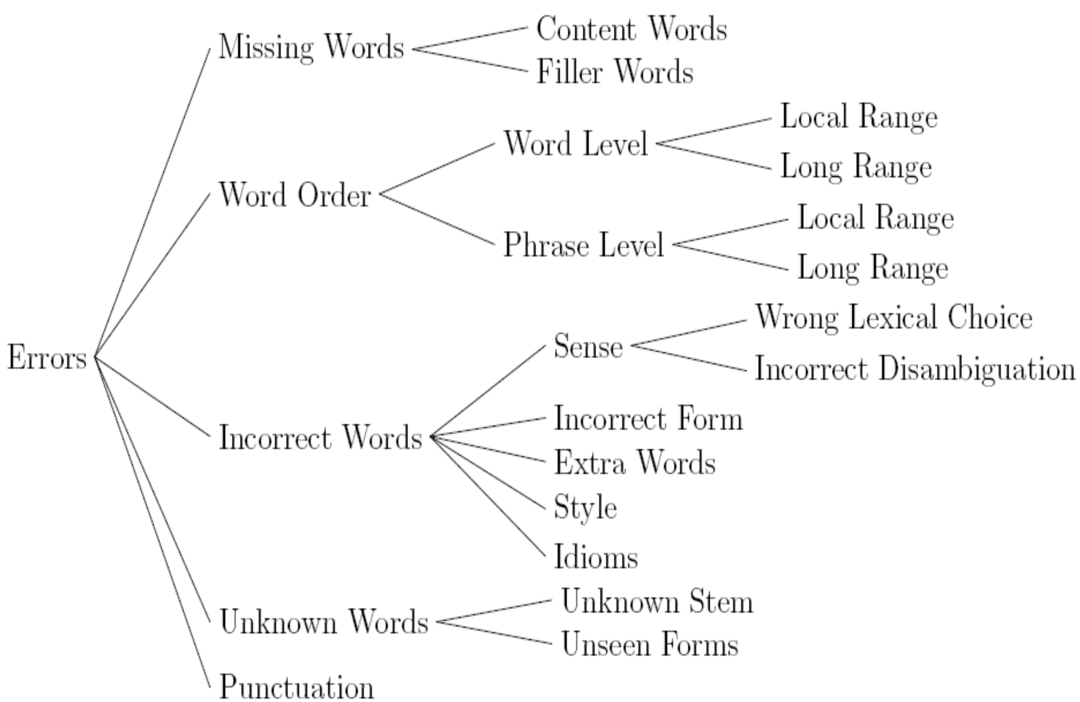

Figure 1: Classification of translation errors.

Automatic Evaluation of MT Quality

- 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 → doesn’t require knowledge in both source and target language.
- Textual similarity may be measured automatically
- Various simple error metrics have been successfully used in speech recognition (Word error rate, …).

TER – Translation Error Rate

- Derived from Levenshtein Distance.
- Counts number of edits necessary to turn translation into references.
- Uses:
  - Deletions
  - Substitutions
  - Insertions
- Very simple.

BLEU – Bilingual Evaluation Understudy

- Idea:
  - 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

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,

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

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

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

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

See ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v13a.pl for a practical implementation.

Why BLEU is popular

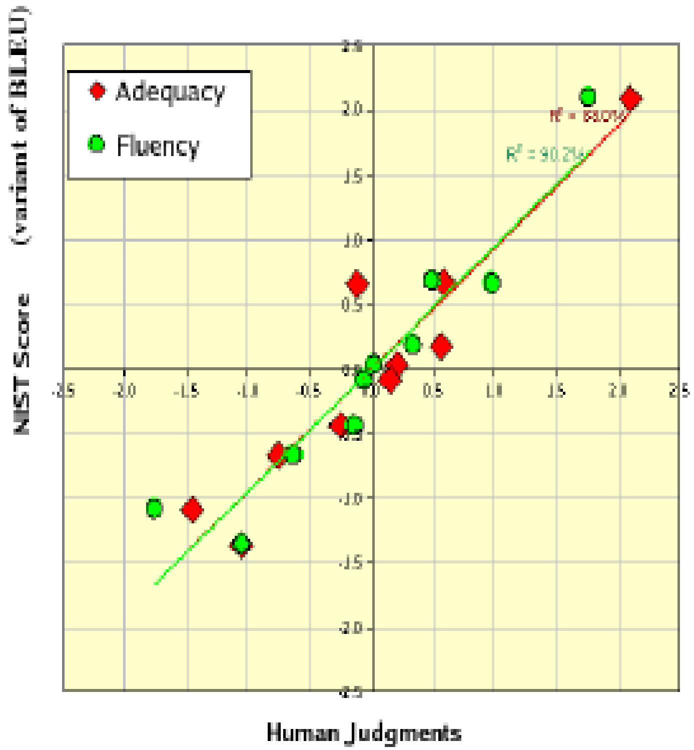

Figure 8.8: Correlation between an automatic metric (here: NIST) and human judgment (fluency, adequacy). Illustration by George Doddington.

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

Why BLEU is controversial

UNIVERSITÄT DES SAARLANDES

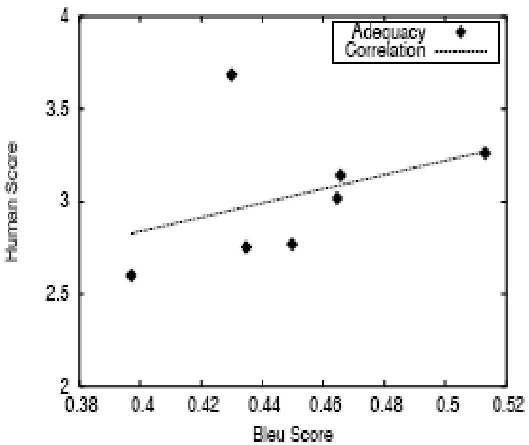

Figure 2: Bleu scores plotted against human judgments of adequacy, with $R^2 = 0.14$ when the outlier entry is included

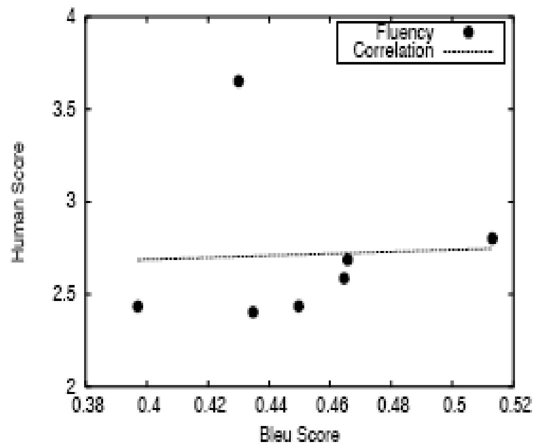

Figure 3: Bleu scores plotted against human judgments of fluency, with $R^2 = 0.002$ when the outlier entry is included

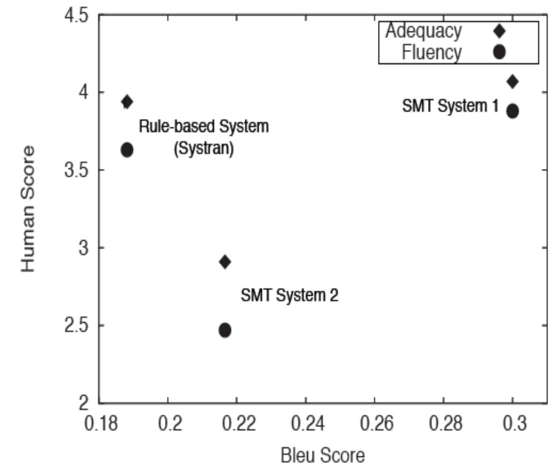

Figure 4: Bleu scores plotted against human judgments of fluency and adequacy, showing that Bleu vastly underestimates the quality of a non-statistical system

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

METEOR

- METEOR uses precision and recall: calculates alignment between translation and reference.
- But it also makes use of different matching modules:
  - exact

    translation: house

    reference: house
  - stemmer (lemmatiser)

    translation: houses

    reference: house
  - synonymy (wordnet)

    translation: building

    reference: house

Scores

UNIVERSITÄT DES SAARLANDES

- We want a score that correlates with human judgment.
- To get best results, use several scores.
- But still each score is just a number: is a system with a BLEU score of 16 really worse than a system with a score of 20? How about 17.9 and 18.5?
- We would like to know *error types* (cf. manual evaluation).
  - POS-BLEU, …

Other Uses for Evaluation

- We usually evaluate to improve our systems.

→ global evaluation for entire text (document-level)

- Evaluation at run-time: quality estimation.
  - Based on a number of features determine how good the MT quality is on the *sentence-level*.
  - Can be useful for e.g. post-editing (if the text is too bad, don't show it to the translator).

Summary

- Manual evaluation is meaningful, but tedious.
- Automatic scoring is fast, but how do we get the meaning out of the scores?
- Evaluation ties in with quality estimation.