Wait!

Thanks to

Meritxell Gonzàlez and Lluís Màrquez

for some of the slides
Outline

1. Basics
3. Automatic Evaluation
4. Tools
MT Evaluation
Importance for system development

Error detection → Evaluation methods
Error analysis → Refinement → Implementation → Test

OK? YES → Unfruitful results
NO
MT Evaluation
Importance for system development

Error detection
Error analysis
Refinement
Implementation
Test
Evaluation methods

Unfruitful results
MT Evaluation
Importance for system development

Implementation → Error detection → Error analysis → Refinement → Evaluation methods

Implementation → Test

Test

YES → OK?

NO → Unfruitful results
MT Evaluation
Importance for system development

- Error detection
  - Error analysis
  - Refinement
  - Implementation
  - Test

OK?

Evaluation methods

Unfruitful results

YES  NO
MT Evaluation
Importance for system development

- Error detection
- Error analysis
- Refinement
- Implementation
- Test

Evaluation methods

YES OK?

NO Unfruitful results
MT Evaluation
Importance for system development

- Error detection
- Error analysis
- Refinement
- Implementation
- Test
  - OK?
    - YES
    - NO
      - Unfruitful results

Evaluation methods
MT Evaluation
Importance for system development

Evaluation methods

Error detection → Error analysis → Refinement → Implementation → Test

OK? YES → OK? NO → Unfruitful results
Automatic metrics notably accelerate the development cycle of MT systems:

- Error analysis
- System optimisation
- System comparison

Besides, they are

- costless (vs. costly),
- objective (vs. subjective),
- reusable (vs. non-reusable)
MT Evaluation
Automatic vs. Manual evaluation

Automatic metrics notably accelerate the development cycle of MT systems:

- Error analysis
- System optimisation
- System comparison

Besides, they are

- **costless** (vs. costly),
- **objective** (vs. subjective),
- **reusable** (vs. non-reusable)
MT Evaluation
Automatic vs. Manual evaluation

Risks of Automatic Evaluation

- **System overtuning**: when system parameters are adjusted towards a given metric

- **Blind system development**: when metrics are unable to capture actual system improvements

- **Unfair system comparisons**: when metrics are unable to reflect difference in quality between MT systems
Machine Translation is an open NLP task

- The correct translation is not unique
- The set of valid translations is not small
- Translation correctness is not black and white
- Quality aspects are heterogeneous
MT Evaluation

Quality aspects

**Adequacy** (or Fidelity) Does the output convey the same meaning as the input sentence? Is part of the message lost, added, or distorted?

**Fluency** (or Intelligibility) Is the output fluent? This involves both grammatical correctness and idiomatic word choices.

**Post–edition effort** Time required to *repair* the translation, number of key strokes, etc.
Outline

1 Basics

2 Manual Evaluation
   • Likert scales
   • Rankings
   • Pros, cons and agreements

3 Automatic Evaluation

4 Tools
Manual Evaluation

Human annotations

**Likert scales** – TAUS recommendation

**Adequacy**  How much of the meaning expressed in the gold-standard translation or the source is also expressed in the target translation?

4  Everything
3  Most
2  Little
1  None

**Fluency**  To what extent is a target side translation grammatically well informed, without spelling errors and experienced as using natural/intuitive language by a native speaker?

4  Flawless
3  Good
2  Disfluent
1  Incomprehensible

Manual Evaluation
Human annotations

Likert scales – NIST example

**Adequacy I**  How much of the meaning expressed in the Reference translation is also expressed in the System translation?

**Adequacy II**  Does the Machine translation mean essentially the same as the Reference translation?

7-point scale ranging from 1 (None) to 7 (All)

Yes/No, Adequacy I > 4
No, Adequacy II ≤ 4

http://www.itl.nist.gov/iad/mig/tests/metricsmatr/2008/results/correlationResults.html
Manual Evaluation
Human annotations

**Ranking** — Pair-wise comparison

Annotators chose the best system, given the source and target sentence, and 2 anonymised random systems.

**Ranking**

Annotators rank $n$ anonymised systems, randomly selected and randomly ordered.
Manual Evaluation

Appraise

(Federmann 2012)
“Appraise is an open-source tool for manual evaluation of Machine Translation output.”

Appraise allows to collect **human judgments** on translation output, implementing annotation tasks such as

- translation quality checking;
- ranking of translations;
- error classification;
- manual post-editing.
Manual Evaluation

Pros & Cons

- Likert scales have to be defined
- 4-, 5-, 7, 10-point likert scales have been used
- The concept of ranking is easy
- Ranks provide less information
- Agreement among annotators (common!)
Cohen’s kappa coefficient, $\kappa$ (Cohen, 1960)

$$
\kappa = \frac{Pr(\text{agreement}) - Pr(\text{expected})}{1 - Pr(\text{expected})}
$$

Kappa interpretation (Landis & Kogh, 1977)

- 0.0–0.2 slight
- 0.2–0.4 fair
- 0.4–0.6 moderate
- 0.6–0.8 substantial
- 0.8–1.0 almost perfect
### Manual Evaluation

Interanotator Agreement Workshop on statistical machine translation, WMT13

<table>
<thead>
<tr>
<th>Language Pairs</th>
<th>Inter-$\kappa$</th>
<th>Intra-$\kappa$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CZ-EN</td>
<td>0.244</td>
<td>0.479</td>
</tr>
<tr>
<td>RU-EN</td>
<td>0.278</td>
<td>0.513</td>
</tr>
<tr>
<td>EN-FR</td>
<td>0.231</td>
<td>0.495</td>
</tr>
<tr>
<td>FR-EN</td>
<td>0.275</td>
<td>0.578</td>
</tr>
<tr>
<td>ES-EN</td>
<td>0.206</td>
<td>0.492</td>
</tr>
<tr>
<td>EN-DE</td>
<td>0.267</td>
<td>0.575</td>
</tr>
<tr>
<td>DE-EN</td>
<td>0.299</td>
<td>0.535</td>
</tr>
<tr>
<td>EN-CZ</td>
<td>0.168</td>
<td>0.290</td>
</tr>
</tbody>
</table>

- Even Inter-$\kappa$ only slight or fair
- Even Intra-$\kappa$ only fair or moderate
Human-targeted Translation Error Rate, HTER

**Annotator** Post-edition of the candidate translation to have the same meaning as a reference translation with as few edits as possible

**Evaluation** TER with the candidate translation and the post-edited reference

\[
HTER = \frac{\text{Substitutions} + \text{Insertions} + \text{Deletions} + \text{Shifts}}{\text{ReferenceWords}}
\]
Outline

1 Basics

2 Manual Evaluation
   • Likert scales
   • Rankings
   • Pros, cons and agreements

3 Automatic Evaluation
   • Lexical metrics
     • BLEU
   • Limits of lexical similarity
     • METEOR

4 Tools
   • Software
   • Demo
**Setting**  Compute *similarity* between system’s output and one or several reference translations

**Challenge**  The similarity measure should be able to discriminate whether the two sentences convey the same meaning (*semantic equivalence*)
Automatic evaluation

Lexical similarity

Metrics based on lexical similarity
(most of the metrics!)

- **Edit Distance**: WER, PER, TER
- **Precision**: BLEU, NIST, WNM
- **Recall**: ROUGE, CDER
- **Precision/Recall**: GTM, METEOR, BLANC, SIA

Nowadays, BLEU is accepted as the standard metric.
Automatic evaluation
Lexical similarity

Metrics based on lexical similarity
(most of the metrics!)

- **Edit Distance:** WER, PER, TER
- **Precision:** BLEU, NIST, WNM
- **Recall:** ROUGE, CDER
- **Precision/Recall:** GTM, METEOR, BLANC, SIA

Nowadays, BLEU is accepted as *the standard* metric.
The main idea is to use a weighted average of variable length phrase matches against the reference translations. This view gives rise to a family of metrics using various weighting schemes. We have selected a promising baseline metric from this family.”
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.
Automatic evaluation
IBM BLEU: Papineni, Roukos, Ward and Zhu (2001)

Candidate 1:
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.
Candidate 1:

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.
Automatic evaluation
IBM BLEU: Papineni, Roukos, Ward and Zhu (2001)

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.
Modified $n$-gram precision (1-gram)

Precision-based measure, but:

Candidate:
   The the the the the the the the the.

Reference 1:
   The cat is on the mat.

Reference 2:
   There is a cat on the mat.
Automatic evaluation
IBM BLEU: Papineni, Roukos, Ward and Zhu (2001)

Modified n-gram precision (1-gram)

Precision-based measure, but:

\[
\text{Prec.} = \frac{1 + \frac{7}{7}}{7}
\]

Candidate:

The the the the the the the the.

Reference 1:

The cat is on the mat.

Reference 2:

There is a cat on the mat.
Automatic evaluation
IBM BLEU: Papineni, Roukos, Ward and Zhu (2001)

Modified n-gram precision (1-gram)

Precision-based measure, but:

\[ \text{Prec.} = \frac{2 + 7}{7} \]

Candidate:
\[ \text{The the the the the the the the} \]

Reference 1:
\[ \text{The cat is on the mat.} \]

Reference 2:
\[ \text{There is a cat on the mat.} \]
Automatic evaluation
IBM BLEU: Papineni, Roukos, Ward and Zhu (2001)

Modified n-gram precision (1-gram)

Precision-based measure, but:

\[ \text{Prec.} = \frac{3 + 7}{7} \]

Candidate:

**The the the the the the the.**

Reference 1:

**The cat is on the mat.**

Reference 2:

**There is a cat on the mat.**
Automatic evaluation

IBM BLEU: Papineni, Roukos, Ward and Zhu (2001)

**Modified n-gram precision** (1-gram)

Precision-based measure, but:

$$\text{Prec.} = \frac{4 + 7}{7}$$

Candidate:

The the the the the the the the.

Reference 1:

The cat is on the mat.

Reference 2:

There is a cat on the mat.
Modified n-gram precision (1-gram)

Precision-based measure, but: \[ \text{Prec.} = \frac{5 + 7}{7} \]

Candidate:

\text{The the the the the the the the.}

Reference 1:

\text{The cat is on the mat.}

Reference 2:

\text{There is a cat on the mat.}
Modified n-gram precision (1-gram)

Precision-based measure, but: \[ \text{Prec.} = \frac{6 + 7}{7} \]

Candidate:

The the the the the the the.

Reference 1:

The cat is on the mat.

Reference 2:

There is a cat on the mat.
Automatic evaluation
IBM BLEU: Papineni, Roukos, Ward and Zhu (2001)

Modified n-gram precision (1-gram)

Precision-based measure, but: \[ \text{Prec.} = \frac{7}{7} \]

Candidate:
\text{The the the the the the the the.}

Reference 1:
\text{The cat is on the mat.}

Reference 2:
\text{There is a cat on the mat.}
Automatic evaluation
IBM BLEU: Papineni, Roukos, Ward and Zhu (2001)

Modified n-gram precision (1-gram)

A reference word should only be matched once.

Algorithm:

1. Count number of times \( w_i \) occurs in each reference.
2. Keep the minimum between the maximum of (1) and the number of times \( w_i \) appears in the candidate (clipping).
3. Add these values and divide by candidate’s number of words.
Modified n-gram precision (1-gram)

Modified 1-gram precision:

Candidate:

The the the the the the the.

Reference 1:

The cat is on the mat.

Reference 2:

There is a cat on the mat.

1. \( w_i \rightarrow \text{The} \)
   
2. \( \# w_{i,R1} = 2 \)
   \( \# w_{i,R2} = 1 \)
   
3. \( \text{Max}(1)=2, \# w_{i,C} = 7 \)
   \( \Rightarrow \text{Min}=2 \)

No more distinct words
Automatic evaluation

IBM BLEU: Papineni, Roukos, Ward and Zhu (2001)

Modified n-gram precision (1-gram)

Modified 1-gram precision: \( P_1 = \)

Candidate:

\textcolor{red}{The} the the the the the the the.

Reference 1:

\textcolor{red}{The} cat is on \textcolor{red}{the} mat.

Reference 2:

There is a cat on \textcolor{red}{the} mat.

1. \( w_i \rightarrow \text{The} \)
   \#\(w_i,R_1 = 2\)
   \#\(w_i,R_2 = 1\)

2. Max\(_(1) = 2\), \#\(w_i,c = 7\)
   \Rightarrow Min=2

3. No more distinct words
Modified n-gram precision (1-gram)

Modified 1-gram precision: \( P_1 = \frac{2}{7} \)

Candidate:

The the the the the the the the.

Reference 1:

The cat is on the mat.

Reference 2:

There is a cat on the mat.

1. \( w_i \rightarrow \text{The} \)
   \( \# w_i, R_1 = 2 \)
   \( \# w_i, R_2 = 1 \)

2. Max\(_{(1)}\) = 2, \( \# w_i, C = 7 \)
   \( \Rightarrow \) Min = 2

3. No more distinct words
Automatic evaluation
IBM BLEU: Papineni, Roukos, Ward and Zhu (2001)

Modified n-gram precision (1-gram)

Modified 1-gram precision: \[ P_1 = \frac{2}{7} \]

Candidate:
The the the the the the the.

Reference 1:
The cat is on the mat.

Reference 2:
There is a cat on the mat.

1. \( w_i \rightarrow \text{The} \)
   \#_{w_i,R1} = 2
   \#_{w_i,R2} = 1

2. \( \text{Max}_{(1)} = 2, \ #_{w_i,C} = 7 \)
   \Rightarrow \text{Min}=2

3. No more distinct words
Automatic evaluation
IBM BLEU: Papineni, Roukos, Ward and Zhu (2001)

Modified n-gram precision

- Straightforward generalisation to $n$-grams, $P_n$.
- Generalisation to multiple sentences:

$$P_n = \frac{\sum_{C \in \{\text{candidates}\}} \sum_{ngram \in C} Count_{\text{clipped}}(ngram)}{\sum_{C \in \{\text{candidates}\}} \sum_{ngram \in C} Count(ngram)}$$

low $n$ adequacy

high $n$ fluency
Brevity penalty

Candidate:
  of the

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.
Automatic evaluation
IBM BLEU: Papineni, Roukos, Ward and Zhu (2001)

Brevity penalty

Candidate: of the $P_1 = 2/2, P_2 = 1/1$

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.
Automatic evaluation
IBM BLEU: Papineni, Roukos, Ward and Zhu (2001)

Brevity penalty

\[ BP = \begin{cases} 
  1 & \text{if } c > r \\
  e^{1-r/c} & \text{if } c \leq r
\end{cases} \]

- Multiplicative factor
- At sentence level, huge punishment for short sentences
- Estimated at document level

c candidate length, r reference length
Automatic evaluation
IBM BLEU: Papineni, Roukos, Ward and Zhu (2001)

BiLingual Evaluation Understudy, BLEU

\[
\text{BLEU} = \text{BP} \cdot \exp \left( \sum_{n=1}^{N} w_n \log P_n \right)
\]

- Geometric average of \( P_n \) (empirical suggestion)
- \( w_n \) positive weights summing to one
- Brevity penalty
Automatic evaluation
IBM BLEU: Papineni, Roukos, Ward and Zhu (2001)

Paper’s Conclusions

- BLEU correlates with human judgements.
- It can distinguish among similar systems.
- Need for multiple references or a big test with heterogeneous references.
- More parametrisation in the future.
Watch out with BLEU implementations!

There are several widely used implementations of BLEU.
(Moses multi-bleu.perl script, NIST mteval-vXX.pl script, etc.)

Results differ because of:

- Different tokenisation approach.
- Different definition of closest reference in the brevity penalty estimation.
**NIST** is based on BLEU but:

- Arithmetic average of $n$-gram counts rather than a geometric average.
- Informative $n$-grams are given more weight.
- Different definition of brevity penalty.
Limits of lexical similarity

The reliability of lexical metrics depends very strongly on the heterogeneity/representativity of reference translations.

e:  This sentence *is* going to be difficult to evaluate.

Ref1: The evaluation of the clause *is* complicated.
Ref2: The sentence will be hard to qualify.
Ref3: The translation is going to be hard to evaluate.
Ref4: It will be difficult to punctuate the output.

Lexical similarity is nor a sufficient neither a necessary condition so that two sentences convey the same meaning.
Limits of lexical similarity

The reliability of lexical metrics depends very strongly on the heterogeneity/representativity of reference translations.

e: This sentence is going to be difficult to evaluate.

Ref1: The evaluation of the clause is complicated.
Ref2: The sentence will be hard to qualify.
Ref3: The translation is going to be hard to evaluate.
Ref4: It will be difficult to punctuate the output.

Lexical similarity is nor a sufficient neither a necessary condition so that two sentences convey the same meaning.
The reliability of lexical metrics depends very strongly on the heterogeneity/representativeness of reference translations.

\[
\text{e: This sentence is going to be difficult to evaluate.}
\]

Ref1: The evaluation of the clause is complicated.
Ref2: The sentence will be hard to qualify.
Ref3: The translation is going to be hard to evaluate.
Ref4: It will be difficult to punctuate the output.

Lexical similarity is nor a sufficient neither a necessary condition so that two sentences convey the same meaning.
Limits of lexical similarity

Beyond lexical similarity

Extend the reference material:

- Using lexical variants such as morphological variations or synonymy lookup or using paraphrasing support.

Compare other linguistic features than words:

- Syntactic similarity: shallow parsing, full parsing (constituents/dependencies).
- Semantic similarity: named entities, semantic roles, discourse representations.

Combination of the existing metrics.
Extending the reference material
METEOR, Banerjee and Lavie (2005)

Metric for Evaluation of Translation with Explicit ORdering

\[
METEOR = (1 - Pen) F_\alpha
\]

\[
F_\alpha = \frac{PR}{\alpha P + (1 - \alpha) R}
\]

\[
Pen = \gamma \left( \frac{\text{chunks}}{\text{mapped unigrams}} \right)^\beta
\]

*Precision* and *Recall* weighted harmonic mean

*Penalty* factor, penalises non-contiguous matches

Matches: exact, lemma, synonym, paraphrase
Extending the reference material
METEOR, Banerjee and Lavie (2005)

Metric for Evaluation of Translation with Explicit ORdering

\[ \text{METEOR} = (1 - \text{Pen}) F_\alpha \]

\[ F_\alpha = \frac{PR}{\alpha P + (1 - \alpha) R} \]

\[ \text{Pen} = \gamma \left( \frac{\text{chunks}}{\text{mapped unigrams}} \right)^\beta \]

**Precision** and **Recall**
weighted harmonic mean

**Penalty** factor, penalises
non-contiguous matches

**Matches**: exact, lemma, synonym, paraphrase
Limits of lexical similarity

Beyond lexical similarity

Extend the reference material:

- Using lexical variants such as morphological variations or synonymy lookup or using paraphrasing support.

Compare other linguistic features than words:

- **Syntactic** similarity: shallow parsing, full parsing (constituents /dependencies).
- **Semantic** similarity: named entities, semantic roles, discourse representations.

Combination of the existing metrics.
Limits of lexical similarity
Comparing other linguistic features than words

Candidate:
On Tuesday several missiles and mortar shells fell in south Kabul, but there were no casualties.

Reference:
Several rockets and mortar shells fell today, Tuesday, in south Kabul without causing any casualties.
Limits of lexical similarity
Comparing other linguistic features than words
Limits of lexical similarity
Comparing other linguistic features than words
Limits of lexical similarity
Comparing other linguistic features than words

Overlap
Generic similarity measure among Linguistic Elements. Inspired by the Jaccard similarity coefficient.

Linguistic element (LE): abstract reference to any possible type of linguistic unit, structure, or relationship among them.

- For instance: POS tags, word lemmas, NPs, syntactic phrases
- A sentence can be seen as a bag (or a sequence) of LEs of a certain type
- LEs may embed
Overlap

Generic similarity measure among Linguistic Elements. Inspired by the Jaccard similarity coefficient.

Linguistic element (LE): abstract reference to any possible type of linguistic unit, structure, or relationship among them.

- For instance: POS tags, word lemmas, NPs, syntactic phrases
- A sentence can be seen as a bag (or a sequence) of LEs of a certain type
- LEs may embed
Limits of lexical similarity
Comparing other linguistic features than words

\[ O(t) = \frac{\sum_{i \in (\text{items}_t(\text{cand}) \cap \text{items}_t(\text{ref}))} \text{count}_{\text{cand}}(i, t)}{\sum_{i \in (\text{items}_t(\text{cand}) \cup \text{items}_t(\text{ref}))} \max(\text{count}_{\text{cand}}(i, t), \text{count}_{\text{ref}}(i, t))} \]

\( t \) is the LE type
‘cand’: candidate translation
‘ref’: reference translation
\( \text{items}_t(s) \): set of items occurring inside LEs of type \( t \)
\( \text{count}_s(i, t) \): occurrences of item \( i \) in \( s \) inside a LE of type \( t \)
Limits of lexical similarity
Comparing other linguistic features than words

Coarser variant: **micro-averaged overlap over all types**

\[
O(\star) = \frac{\sum_{t \in T} \sum_{i \in (\text{items}_t(\text{cand}) \cap \text{items}_t(\text{ref}))} \text{count}_{\text{cand}}(i, t)}{\sum_{t \in T} \sum_{i \in (\text{items}_t(\text{cand}) \cup \text{items}_t(\text{ref}))} \max(\text{count}_{\text{cand}}(i, t), \text{count}_{\text{ref}}(i, t))}
\]

\(T\): set of all LE types associated to the given LE class
Limits of lexical similarity

Beyond lexical similarity

Extend the reference material:

- Using lexical variants such as morphological variations or synonymy lookup or using paraphrasing support.

Compare other linguistic features than words:

- Syntactic similarity: shallow parsing, full parsing (constituents /dependencies).
- Semantic similarity: named entities, semantic roles, discourse representations.

Combination of the existing metrics.
Limits of lexical similarity
Combination of the existing metrics
Limits of lexical similarity

Combination of the existing metrics
Limits of lexical similarity

Combination of the existing metrics

- Different measures capture **different aspects** of similarity suitable for combination

- The most simple approach: **ULC**

**Uniformly averaged linear combination** of measures (ULC):

$$\text{ULC}_M(\text{cand}, \text{ref}) = \frac{1}{|M|} \sum_{m \in M} m(\text{cand}, \text{ref})$$
Limits of lexical similarity
Combination of the existing metrics

- Different measures capture **different aspects** of similarity suitable for combination
- The most simple approach: **ULC**

**Uniformly** averaged **linear combination** of measures (ULC):

\[
ULC_M(cand, ref) = \frac{1}{|M|} \sum_{m \in M} m(cand, ref)
\]
Evaluation is important in the system development cycle. Automatic evaluation accelerates significantly the process.

Manual evaluation is still necessary but shows low agreements among annotators.

Up to now, most (common) metrics rely on lexical similarity, but it cannot assure a correct evaluation.

Current work is being devoted to go beyond lexical similarity.
Outline

1. Basics
3. Automatic Evaluation
4. Tools
   - Software
   - Demo
Evaluate your translations

1. With BLEU scoring tool. Available as a Moses script or from NIST:
   ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v13a.pl

2. With Asiya package:
   http://nlp.lsi.upc.edu/asiya/
ASiya has been designed to assist both system and metric developers by offering a rich repository of metrics and meta-metrics.

http://nlp.lsi.upc.edu/asiya/
In practice

With BLEU scoring tool in Moses:

```
moses/scripts/generic/multi-bleu.perl references.en < testset.translated.en
```
With the Asiya toolkit:

```
Asiya.pl -eval single,ulc -g sys Asiya.config
```

input=raw

SRCLANG=de
TRGLANG=en
SRCCASE=cs
TRGCASE=cs

#SRC __________________________________________________________
src=./data/patsA61P.test.de
#REF _________________________________________________________
ref=./data/patsA61P.test.en
#OUT _________________________________________________________
sys=./data/patsA61P.test.trans.de2en
sys=./data/patsA61P.test.trad.google.de2en
sys=./data/patsA61P.test.trad.bing.de2en
#-------------------------------------------------------
Tools

In practice

With the Asiya toolkit:

Asiya.pl -eval single,ulc -g sys Asiya.config

input=raw

SRCLANG=de
TRGLANG=en
SRCCASE=cs
TRGCASE=cs

#SRC
src=./data/patsA61P.test.de
#REF
ref=./data/patsA61P.test.en
#OUT
sys=./data/patsA61P.test.trans.de2en
sys=./data/patsA61P.test.trad.google.de2en
sys=./data/patsA61P.test.trad.bing.de2en
In practice

Asiya.pl -eval single,ulc -m metrSet Asiya.config

SRCLANG=de
TRGLANG=en

#SRC ================================================
src=./data/patsA61P.test.de
#REF ================================================
ref=./data/patsA61P.test.en
#OUT ================================================
sys=./data/patsA61P.test.trans.de2en
#-----------------------------------------------------
metrSet=1-PER 1-TER 1-WER BLEU-4 CP-Oc-* CP-Op-* CP-STM-9 DP-HWC-c-4
DP-HWC-r-4 DP-HWC-w-4 DP-Oc-* DP-01-* DP-Or-* DR-Or-* DR-Orp-* DR-STM-9
GTM-1 GTM-2 GTM-3 MTR-exact MTR-stem MTR-wnstm MTR-wnsyn NE-Me-* NE-0e-*
NE-0e-** NIST-5 RG-L RG-S* RG-SU* RG-W-1.2 SP-Oc-* SP-Op-* SP-cNIST-5
SP-iobNIST-5 SP-1NIST-5 SP-pNIST-5 SR-Mr-* SR-Mrv-* SR-Or SR-0r-* SR-Orv
Asiya interfaces
Evaluate the results on-line

1. Asiya Interface
   http://asiya.lsi.upc.edu/demo/asiya_online.php
Tools

On-line evaluation

Analise the results on-line

1. t-Search Interface
   http://asiya.lsi.upc.edu/demo/tsearch_upload.php
MT Evaluation
Demo: http://asiya.lsi.upc.edu/demo/asiya_online.php
Manual Evaluation

- Cohen, 1960 [Coh60]
- Landis & Koch, 1977 [LK77]
- Federmann 2012 [Fed12]
References

Automatic Evaluation

- Papineni, 2002 [PRWZ02]
- Doddington, 2002 [Dod02]
- Banerjee & Alon Lavie, 2005 [BL05]
- Giménez & Amigó, 2006 [GA06]
Metrics I

- WER [NOLN00]
- PER [TVN+97]
- TER [SDS+06]
References

Metrics II

- BLEU [PRWZ02]
- NIST [Dod02]
- METEOR [BL05]
- ROUGE [LO04]
References

Metrics III

- GTM [MGT03]
- BLANC [Dod02]
- CDER [LUN06]
- ULC [GA06]
Satanjeev Banerjee and Alon Lavie.
METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments.

Jacob Cohen.
A coefficient of agreement for nominal scales.  

George Doddington.
Automatic evaluation of machine translation quality using n-gram co-occurrence statistics.  

Christian Federmann.
Jesus Gimenez and Enrique Amigo.

J. R. Landis and G. G. Koch.
The measurement of observer agreement for categorical data.

Chin-Yew Lin and Franz Josef Och.

Gregor Leusch, Nicola Ueffing, and Hermann Ney.
CDER: Efficient MT Evaluation Using Block Movements.

I. Dan Melamed, Ryan Green, and Joseph P. Turian.
Precision and Recall of Machine Translation.

