# Introduction to Machine Translation Evaluation

**Cristina España-Bonet**
UdS & DFKI

Summer Semester 2018
30th May 2018

# *Wait!*

Thanks to

**Meritxell Gonzàlez** and **Lluís Màrquez**

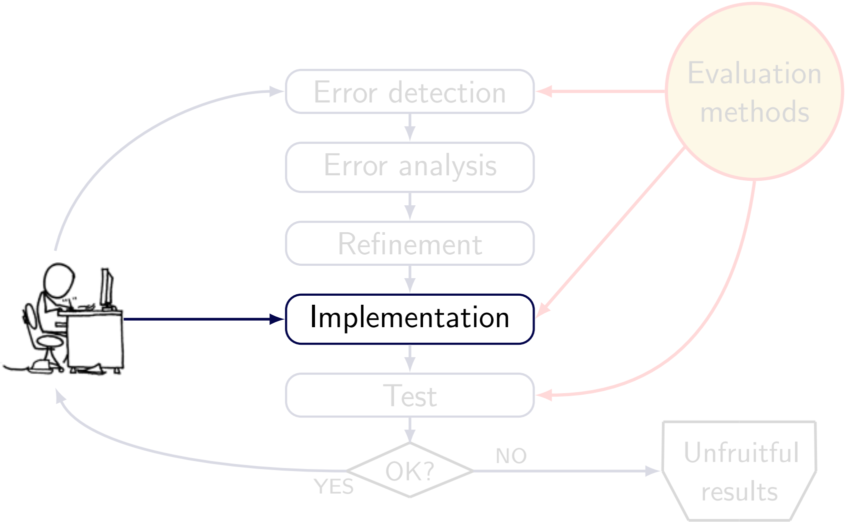
for some of the slides

# *Outline*

1. Basics
2. Manual Evaluation
3. Automatic Evaluation
4. Tools
5. References

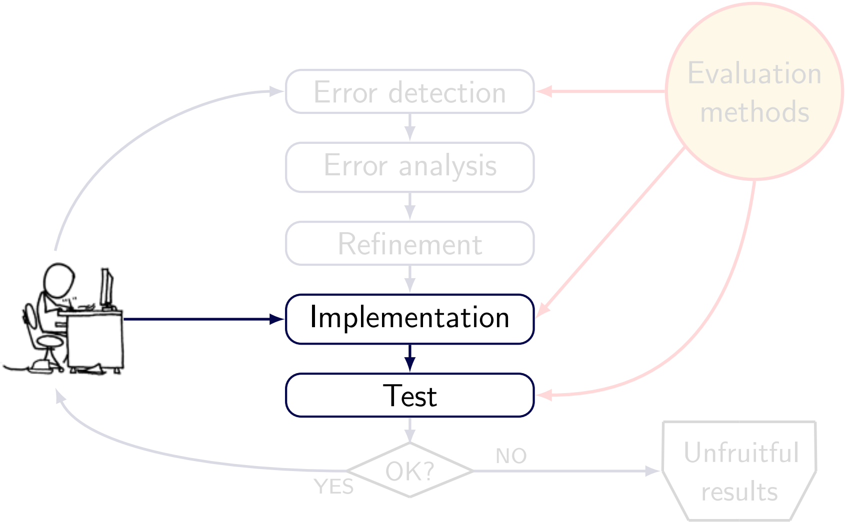
# MT Evaluation

## Importance for system development



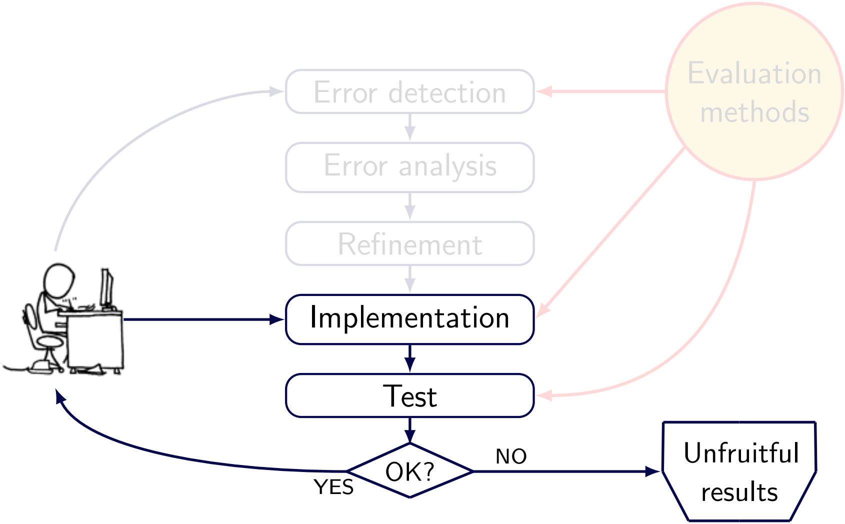
# MT Evaluation

## Importance for system development



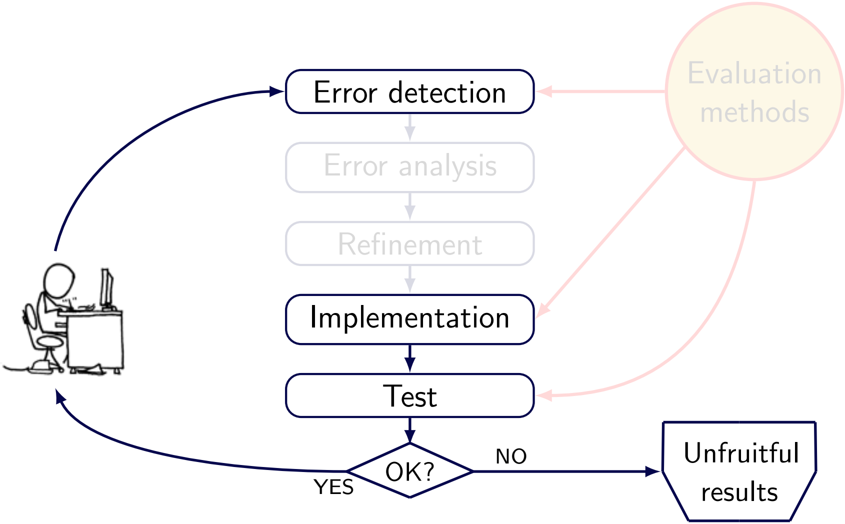
# MT Evaluation

## Importance for system development



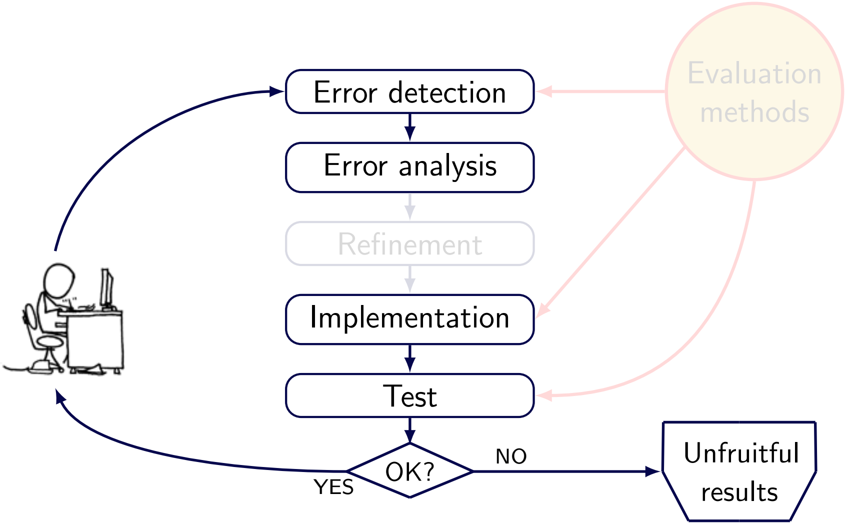
# MT Evaluation

## Importance for system development



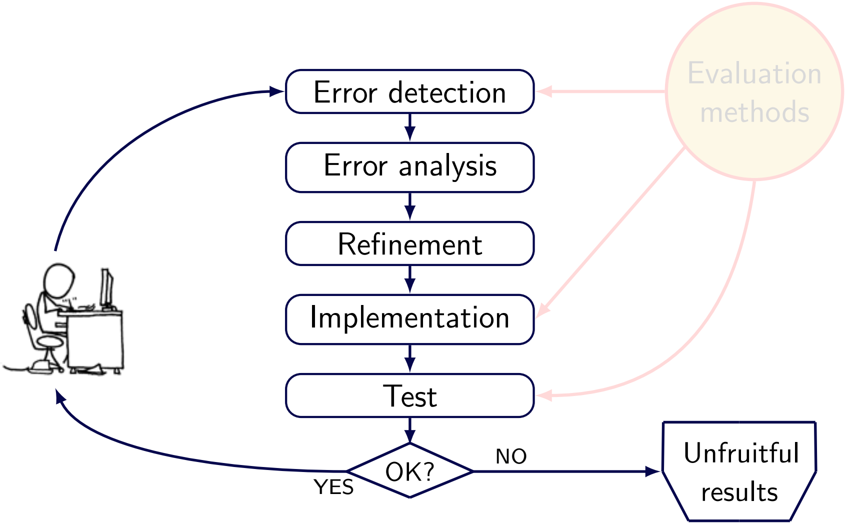
# MT Evaluation

## Importance for system development



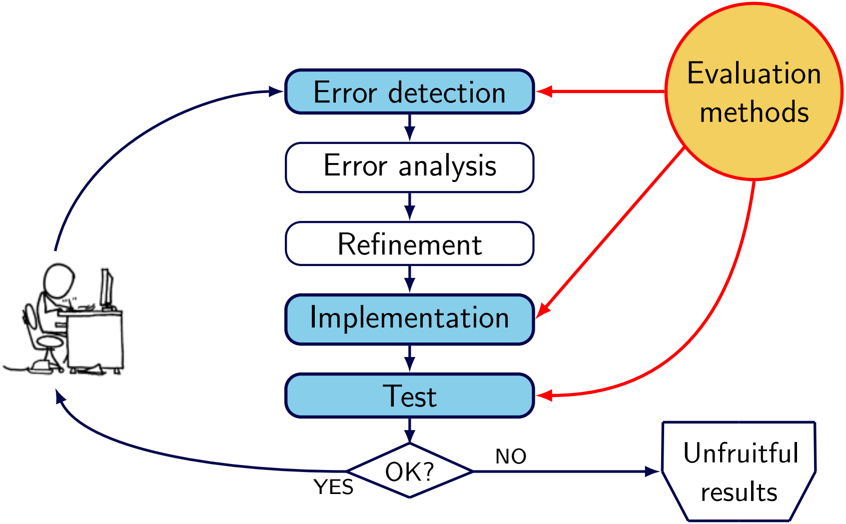
# MT Evaluation

## Importance for system development



# MT Evaluation

## Importance for system development



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)

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)

## 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

**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*

## **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 |

https://www.taus.net/think-tank/best-practices/evaluate-best-practices/adequacy-fluency-guidelines

## **Likert scales** – NIST example

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

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

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

**Yes/No**, Adequacy I $> 4$
**No**, Adequacy II $\leq 4$

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

**Ranking** – Pair-wise comparison

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

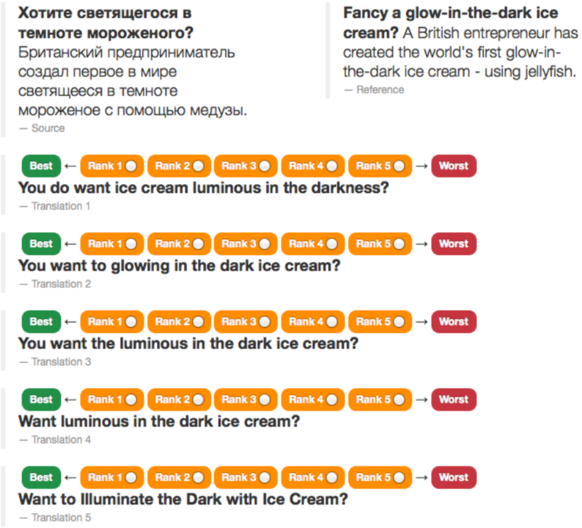
**Ranking**

Annotators rank *n* anonymised systems, randomly selected and randomly ordered.

# Manual Evaluation

## Appraise

**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 |

**More details:**

https://staff.fnwi.uva.nl/r.fernandezrovira/teaching/MoLProject2011/annotation-reliability.pdf

(slides 6-14)

# Manual Evaluation

## Interanotator Agreement

Workshop on statistical machine translation, **WMT13**

- Inter-$\kappa$ only slight or fair
- Even Intra-$\kappa$ only fair or moderate

| | Inter-$\kappa$ | Intra-$\kappa$ |
|---|---|---|
| CZ–EN | 0.244 | 0.479 |
| EN–CZ | 0.168 | 0.290 |
| DE–EN | 0.299 | 0.535 |
| EN–DE | 0.267 | 0.498 |
| ES–EN | 0.277 | 0.575 |
| EN–ES | 0.206 | 0.492 |
| FR–EN | 0.275 | 0.578 |
| EN–FR | 0.231 | 0.495 |
| RU–EN | 0.278 | 0.450 |
| EN–RU | 0.243 | 0.513 |

**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

$$HTER = \frac{\text{Substitutions} + \text{Insertions} + \text{Deletions} + \text{Shifts}}{\text{ReferenceWords}}$$

# Wait!

# Outline

**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**)

**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

**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.

**BLEU: a Method for Automatic Evaluation of Machine Translation**

Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu
IBM Research Division

"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."

# 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.`

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.`

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.
```

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: $\text{Prec.} = \frac{1+}{7}$

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.`

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

Precision-based measure, but: $\text{Prec.} = \frac{2+}{7}$

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.`

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

Precision-based measure, but: $\text{Prec.} = \frac{3+}{7}$

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

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.`

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

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

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.`

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

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

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.`

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

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

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)

A reference word should only be matched once.

Algorithm:

1. Count number of times $w_i$ occurs in each reference.
2. Keep the minimun 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. $w_i \rightarrow \mathrm{The}$
   $\#w_{i,R1} = 2$
   $\#w_{i,R2} = 1$
2. $\mathrm{Max}_{(1)}=2$, $\#w_{i,C} = 7$
   $\Rightarrow$ Min=2
3. No more distinct words

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

Modified 1-gram precision: $\mathrm{P}_1 =$

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. $w_i \rightarrow \mathrm{The}$
   $\#_{w_i,R1} = 2$
   $\#_{w_i,R2} = 1$
2. $\mathrm{Max}_{(1)}$=2, $\#_{w_i,C} = 7$
   $\Rightarrow$ Min=2
3. No more distinct words

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

Modified 1-gram precision: $P_1 = \frac{2}{}$

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. $w_i \rightarrow$ The
   $\#_{w_i,R1} = 2$
   $\#_{w_i,R2} = 1$
2. $\text{Max}_{(1)}=2$, $\#_{w_i,C} = 7$
   $\Rightarrow$ 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.`

Reference 1:

`The cat is on the mat.`

Reference 2:

`There is a cat on the mat.`

1. $w_i \to \text{The}$
   $\#_{w_i, R1} = 2$
   $\#_{w_i, R2} = 1$
2. $\text{Max}_{(1)}$=2, $\#_{w_i, C} = 7$
   $\Rightarrow$ Min=2
3. No more distinct words

## Modified n-gram precision

- Straightforward generalisation to $n$-grams, $\mathrm{P}_n$.
- Generalisation to multiple sentences:

$$\mathrm{P}_n = \frac{\sum_{C \in \{\text{candidates}\}} \sum_{n\text{gram} \in C} \mathit{Count}_{\text{clipped}}(n\text{gram})}{\sum_{C \in \{\text{candidates}\}} \sum_{n\text{gram} \in C} \mathit{Count}(n\text{gram})}$$

| | |
|---|---|
| low $n$ | high $n$ |
| adequacy | 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.`

## 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**

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

$c$ candidate length, $r$ reference length

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

**BiLingual Evaluation Understudy, BLEU**

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

- Geometric average of $\mathrm{P}_n$ (empirical suggestion)
- $w_n$ positive weights summing to one
- Brevity penalty

## 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.

# Wait!

**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.

## 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.

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.

## Metric for Evaluation of Translation with Explicit ORdering

$$METEOR = (1 - Pen)F_\alpha$$

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

*P*__recision__ and *R*__ecall__ weighted harmonic mean

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

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

**Matches**: exact, lemma, synonym, paraphrase

## Metric for Evaluation of Translation with Explicit ORdering

$$METEOR = (1 - Pen)F_\alpha$$

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

*P***recision** and *R***ecall** weighted harmonic mean

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

**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

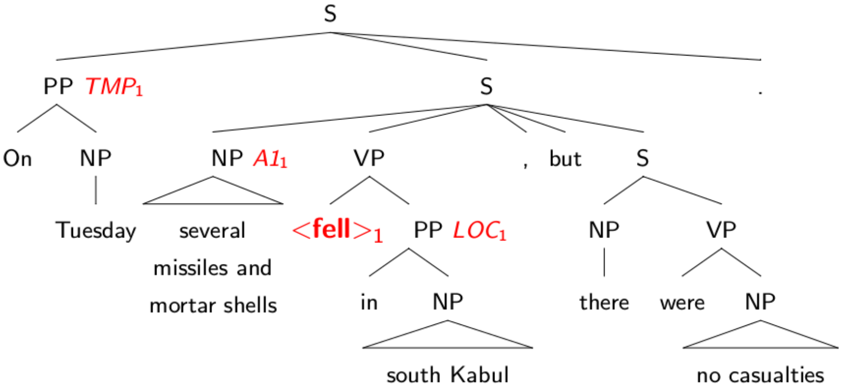
## Comparing other linguistic features than words

Candidate:

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

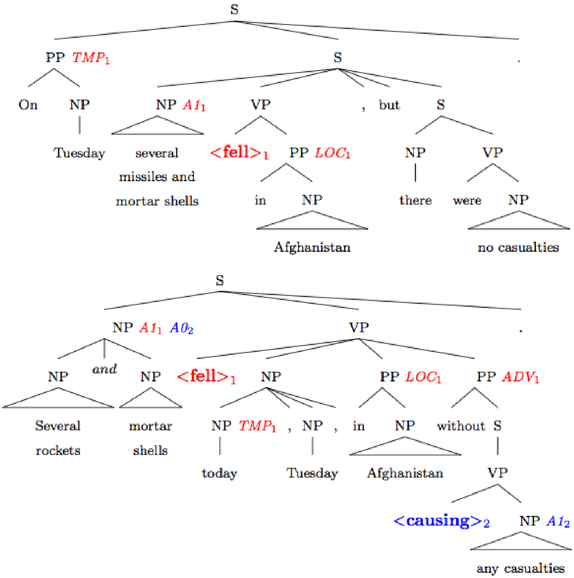
Reference:

```
 Several rockets and mortar shells fell today,
Tuesday, in south Kabul without causing any
casualties.
```


S
PP TMP1
S
.
On
NP
NP A11
VP
,
but
S
Tuesday
several missiles and mortar shells
<fell>1
PP LOC1
NP
VP
in
NP
there
were
NP
south Kabul
no casualties

# 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

$$\mathrm{O}(t) = \frac{\displaystyle\sum_{i \in (\mathrm{items}_t(\mathrm{cand}) \,\cap\, \mathrm{items}_t(\mathrm{ref}))} \mathrm{count}_{\mathrm{cand}}(i, t)}{\displaystyle\sum_{i \in (\mathrm{items}_t(\mathrm{cand}) \,\cup\, \mathrm{items}_t(\mathrm{ref}))} \max(\mathrm{count}_{\mathrm{cand}}(i, t), \mathrm{count}_{\mathrm{ref}}(i, t))}$$

$t$ is the LE type

'cand': candidate translation
'ref': reference translation
$\mathrm{items}_t(s)$: set of items occurring inside LEs of type $t$
$\mathrm{count}_s(i, t)$: occurrences of item $i$ in $s$ inside a LE of type $t$

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

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

$T$: 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.

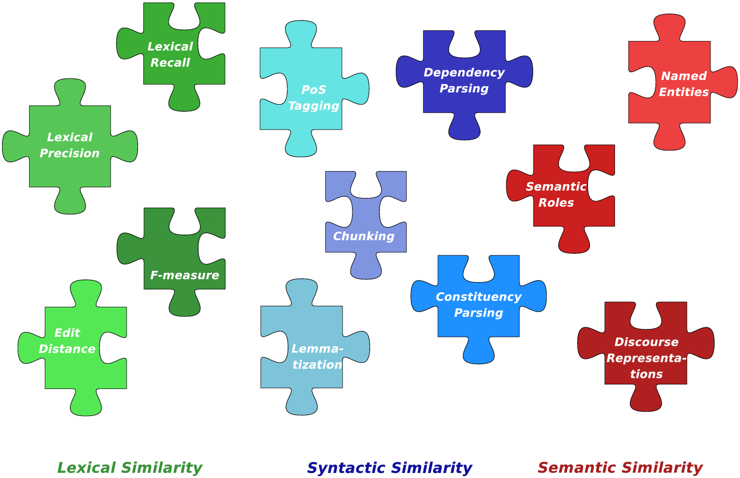
Compare other linguistic features than words:

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

Combination of the existing metrics.

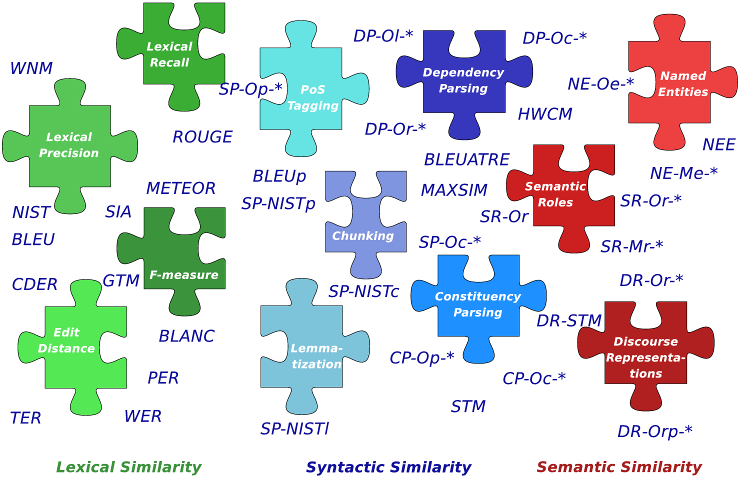
# Limits of lexical similarity

## Combination of the existing metrics



# Limits of lexical similarity

## Combination of the existing metrics



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

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

$$\mathrm{ULC}_M(\mathrm{cand}, \mathrm{ref}) = \frac{1}{|M|} \sum_{m \in M} m(\mathrm{cand}, \mathrm{ref})$$

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

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

$$\mathrm{ULC}_M(\mathrm{cand}, \mathrm{ref}) = \frac{1}{|M|} \sum_{m \in M} m(\mathrm{cand}, \mathrm{ref})$$

- Evaluation is important in the system development cycle. Automatic evaluation accelerates significatively 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.

# Wait!

# Outline

**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://asiya.cs.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://asiya.cs.upc.edu/demo/asiya_online.php

1. With BLEU scoring tool in `Moses`:

```
moses/scripts/generic/multi-bleu.perl references.en <
testset.translated.en
```

2. 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

2. 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

```
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-Ol-* 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-Oe-*
NE-Oe-** NIST-5 RG-L RG-S* RG-SU* RG-W-1.2 SP-Oc-* SP-Op-* SP-cNIST-5
SP-iobNIST-5 SP-lNIST-5 SP-pNIST-5 SR-Mr-* SR-Mrv-* SR-Or SR-Or-* SR-Orv
```

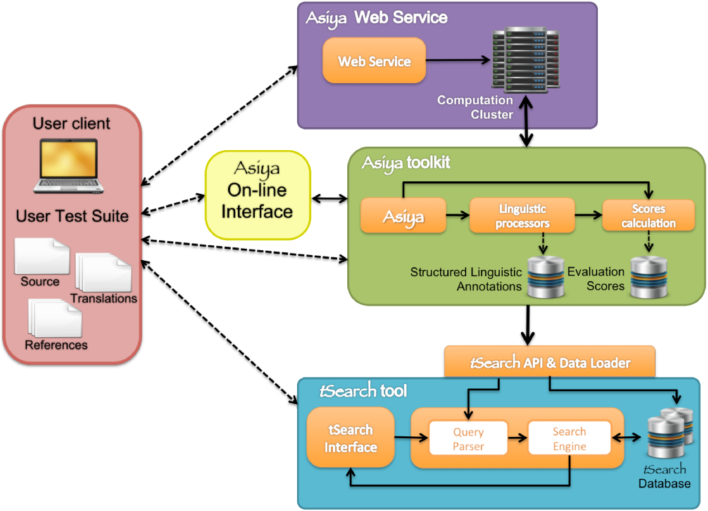
# Tools

## In practice

```
--------------------------------------------------------------------------------------------------------------------
METRIC NAMES
--------------------------------------------------------------------------------------------------------------------
668 metrics are available for language 'en'

METRICS = { -PER, -TER, -TERbase, -TERp, -TERp-A, -WER, BLEU, BLEU-1, BLEU-2, BLEU-3, BLEU-4, BLEUi-2, BLEUi-3, BLEUi-4, CP-Oc(*), CP-Oc(ADJP), CP-Oc(ADVP), CP-Oc(CONJP), CP-Oc(FRA
G), CP-Oc(INTJ), CP-Oc(LST), CP-Oc(NAC), CP-Oc(NP), CP-Oc(NX), CP-Oc(O), CP-Oc(PP), CP-Oc(PRN), CP-Oc(PRT), CP-Oc(QP), CP-Oc(RRC), CP-Oc(S), CP-Oc(SBAR), CP-Oc(SINV), CP-Oc(SQ), CP
-Oc(UCP), CP-Oc(VP), CP-Oc(WHADJP), CP-Oc(WHADVP), CP-Oc(WHNP), CP-Oc(WHPP), CP-Oc(X), CP-Op(#), CP-Op($), CP-Op(''), CP-Op((), CP-Op()), CP-Op(*), CP-Op(,), CP-Op(.), CP-Op(:), CP
-Op(CC), CP-Op(CD), CP-Op(DT), CP-Op(EX), CP-Op(F), CP-Op(FW), CP-Op(IN), CP-Op(J), CP-Op(JJ), CP-Op(JJR), CP-Op(JJS), CP-Op(LS), CP-Op(MD), CP-Op(N), CP-Op(NN), CP-Op(NNP), CP-Op(
NNPS), CP-Op(NNS), CP-Op(P), CP-Op(PDT), CP-Op(POS), CP-Op(PRP$), CP-Op(PRP), CP-Op(R), CP-Op(RB), CP-Op(RBR), CP-Op(RBS), CP-Op(RP), CP-Op(SYM), CP-Op(TO), CP-Op(UH), CP-Op(V), CP
-Op(VB), CP-Op(VBD), CP-Op(VBG), CP-Op(VBN), CP-Op(VBP), CP-Op(VBZ), CP-Op(W), CP-Op(WDT), CP-Op(WP$), CP-Op(WP), CP-Op(WRB), CP-Op(``), CP-STM-1, CP-STM-2, CP-STM-3, CP-STM-4, CP-
STM-5, CP-STM-6, CP-STM-7, CP-STM-8, CP-STM-9, CP-STMi-2, CP-STMi-3, CP-STMi-4, CP-STMi-5, CP-STMi-6, CP-STMi-7, CP-STMi-8, CP-STMi-9, DP-HWCM_c-1, DP-HWCM_c-2, DP-HWCM_c-3, DP-HWC
M_c-4, DP-HWCM_r-1, DP-HWCM_r-2, DP-HWCM_r-3, DP-HWCM_r-4, DP-HWCM_w-1, DP-HWCM_w-2, DP-HWCM_w-3, DP-HWCM_w-4, DP-HWCMi_c-2, DP-HWCMi_c-3, DP-HWCMi_c-4, DP-HWCMi_r-2, DP-HWCMi_r-3,
 DP-HWCMi_r-4, DP-HWCMi_w-2, DP-HWCMi_w-3, DP-HWCMi_w-4, DP-Oc(*), DP-Oc(a), DP-Oc(as), DP-Oc(aux), DP-Oc(be), DP-Oc(c), DP-Oc(comp), DP-Oc(det), DP-Oc(have), DP-Oc(n), DP-Oc(postd
et), DP-Oc(ppspec), DP-Oc(predet), DP-Oc(prep), DP-Oc(saidx), DP-Oc(sentadjunct), DP-Oc(subj), DP-Oc(that), DP-Oc(u), DP-Oc(v), DP-Oc(vbe), DP-Oc(xsaid), DP-Ol(*), DP-Ol(1), DP-Ol(
2), DP-Ol(3), DP-Ol(4), DP-Ol(5), DP-Ol(6), DP-Ol(7), DP-Ol(8), DP-Ol(9), DP-Or(*), DP-Or(amod), DP-Or(amount-value), DP-Or(appo), DP-Or(appo-mod), DP-Or(as-arg), DP-Or(as1), DP-Or
(as2), DP-Or(aux), DP-Or(be), DP-Or(being), DP-Or(by-subj), DP-Or(c), DP-Or(cn), DP-Or(comp1), DP-Or(conj), DP-Or(desc), DP-Or(dest), DP-Or(det), DP-Or(else), DP-Or(fc), DP-Or(gen)
, DP-Or(guest), DP-Or(have), DP-Or(head), DP-Or(i), DP-Or(inv-aux), DP-Or(inv-have), DP-Or(lex-dep), DP-Or(lex-mod), DP-Or(mod), DP-Or(mod-before), DP-Or(neg), DP-Or(nn), DP-Or(num
), DP-Or(num-mod), DP-Or(obj), DP-Or(obj1), DP-Or(obj2), DP-Or(p), DP-Or(p-spec), DP-Or(pcomp-c), DP-Or(pcomp-n), DP-Or(person), DP-Or(pnmod), DP-Or(poss), DP-Or(post), DP-Or(pre),
 DP-Or(pred), DP-Or(punc), DP-Or(rel), DP-Or(s), DP-Or(sc), DP-Or(subcat), DP-Or(subclass), DP-Or(subj), DP-Or(title), DP-Or(vrel), DP-Or(wha), DP-Or(whn), DP-Or(whp), DPm-HWCM_c-1
, DPm-HWCM_c-2, DPm-HWCM_c-3, DPm-HWCM_c-4, DPm-HWCM_r-1, DPm-HWCM_r-2, DPm-HWCM_r-3, DPm-HWCM_r-4, DPm-HWCM_w-1, DPm-HWCM_w-2, DPm-HWCM_w-3, DPm-HWCM_w-4, DPm-HWCMi_c-2, DPm-HWCMi
_c-3, DPm-HWCMi_c-4, DPm-HWCMi_r-2, DPm-HWCMi_r-3, DPm-HWCMi_r-4, DPm-HWCMi_w-2, DPm-HWCMi_w-3, DPm-HWCMi_w-4, DPm-Oc(*), DPm-Oc(......), DPm-Ol(*), DPm-Ol(1), DPm-Ol(2), DPm-Ol(3)
, DPm-Ol(4), DPm-Ol(5), DPm-Ol(6), DPm-Ol(7), DPm-Ol(8), DPm-Ol(9), DPm-Or(*), DPm-Or(.......), DR-Fr(*), DR-Frp(*), DR-Ol, DR-Or(*), DR-Or(*)_b, DR-Or(*)_i, DR-Or(alfa), DR-Or(car
d), DR-Or(drs), DR-Or(eq), DR-Or(imp), DR-Or(merge), DR-Or(named), DR-Or(not), DR-Or(or), DR-Or(pred), DR-Or(prop), DR-Or(rel), DR-Or(smerge), DR-Or(timex), DR-Or(whq), DR-Or-(dr),
 DR-Orp(*), DR-Orp(*)_b, DR-Orp(*)_i, DR-Orp(alfa), DR-Orp(card), DR-Orp(dr), DR-Orp(drs), DR-Orp(eq), DR-Orp(imp), DR-Orp(merge), DR-Orp(named), DR-Orp(not), DR-Orp(or), DR-Orp(pr
ed), DR-Orp(prop), DR-Orp(rel), DR-Orp(smerge), DR-Orp(timex), DR-Orp(whq), DR-Pr(*), DR-Prp(*), DR-Rr(*), DR-Rrp(*), DR-STM-1, DR-STM-2, DR-STM-3, DR-STM-4, DR-STM-4_b, DR-STM-4_i
, DR-STM-5, DR-STM-6, DR-STM-7, DR-STM-8, DR-STM-9, DR-STMi-2, DR-STMi-3, DR-STMi-4, DR-STMi-5, DR-STMi-6, DR-STMi-7, DR-STMi-8, DR-STMi-9, DRdoc-Ol, DRdoc-Or(*), DRdoc-Or(*)_b, DR
doc-Or(*)_i, DRdoc-Or(alfa), DRdoc-Or(card), DRdoc-Or(dr), DRdoc-Or(drs), DRdoc-Or(eq), DRdoc-Or(imp), DRdoc-Or(merge), DRdoc-Or(named), DRdoc-Or(not), DRdoc-Or(or), DRdoc-Or(pred)
, DRdoc-Or(prop), DRdoc-Or(rel), DRdoc-Or(smerge), DRdoc-Or(timex), DRdoc-Or(whq), DRdoc-Orp(*), DRdoc-Orp(*)_b, DRdoc-Orp(*)_i, DRdoc-Orp(alfa), DRdoc-Orp(card), DRdoc-Orp(dr), DR
doc-Orp(drs), DRdoc-Orp(eq), DRdoc-Orp(imp), DRdoc-Orp(merge), DRdoc-Orp(named), DRdoc-Orp(not), DRdoc-Orp(or), DRdoc-Orp(pred), DRdoc-Orp(prop), DRdoc-Orp(rel), DRdoc-Orp(smerge),
 DRdoc-Orp(timex), DRdoc-Orp(whq), DRdoc-STM-1, DRdoc-STM-2, DRdoc-STM-3, DRdoc-STM-4, DRdoc-STM-4_b, DRdoc-STM-4_i, DRdoc-STM-5, DRdoc-STM-6, DRdoc-STM-7, DRdoc-STM-8, DRdoc-STM-9
, DRdoc-STMi-2, DRdoc-STMi-3, DRdoc-STMi-4, DRdoc-STMi-5, DRdoc-STMi-6, DRdoc-STMi-7, DRdoc-STMi-8, DRdoc-STMi-9, Fl, GTM-1, GTM-2, GTM-3, METEOR-ex, METEOR-pa, METEOR-st, METEOR-s
y, NE-Me(*), NE-Me(ANGLE_QUANTITY), NE-Me(DATE), NE-Me(DISTANCE_QUANTITY), NE-Me(LANGUAGE), NE-Me(LOC), NE-Me(MEASURE), NE-Me(METHOD), NE-Me(MISC), NE-Me(MONEY), NE-Me(NUM), NE-Me(
ORG), NE-Me(PER), NE-Me(PERCENT), NE-Me(PROJECT), NE-Me(SIZE_QUANTITY), NE-Me(SPEED_QUANTITY), NE-Me(SYSTEM), NE-Me(TEMPERATURE_QUANTITY), NE-Me(TIME), NE-Me(WEIGHT_QUANTITY), NE-O
e(*), NE-Oe(**), NE-Oe(ANGLE_QUANTITY), NE-Oe(DATE), NE-Oe(DISTANCE_QUANTITY), NE-Oe(LANGUAGE), NE-Oe(LOC), NE-Oe(MEASURE), NE-Oe(METHOD), NE-Oe(MISC), NE-Oe(MONEY), NE-Oe(NUM), NE
-Oe(O), NE-Oe(ORG), NE-Oe(PER), NE-Oe(PERCENT), NE-Oe(PROJECT), NE-Oe(SIZE_QUANTITY), NE-Oe(SPEED_QUANTITY), NE-Oe(SYSTEM), NE-Oe(TEMPERATURE_QUANTITY), NE-Oe(TIME), NE-Oe(WEIGHT_Q
UANTITY), NIST, NIST-1, NIST-2, NIST-3, NIST-4, NIST-5, NISTi-2, NISTi-3, NISTi-4, NISTi-5, Ol, Pl, ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4, ROUGE-L, ROUGE-S*, ROUGE-SU*, ROUGE-W, Rl, S
P-Oc(*), SP-Oc(ADJP), SP-Oc(ADVP), SP-Oc(CONJP), SP-Oc(INTJ), SP-Oc(LST), SP-Oc(NP), SP-Oc(O), SP-Oc(PP), SP-Oc(PRT), SP-Oc(SBAR), SP-Oc(UCP), SP-Oc(VP), SP-Op(#), SP-Op($), SP-Op(
''), SP-Op((), SP-Op()), SP-Op(*), SP-Op(,), SP-Op(.), SP-Op(:), SP-Op(CC), SP-Op(CD), SP-Op(DT), SP-Op(EX), SP-Op(F), SP-Op(FW), SP-Op(IN), SP-Op(J), SP-Op(JJ), SP-Op(JJR), SP-Op(
JJS), SP-Op(LS), SP-Op(MD), SP-Op(N), SP-Op(NN), SP-Op(NNP), SP-Op(NNPS), SP-Op(NNS), SP-Op(P), SP-Op(PDT), SP-Op(POS), SP-Op(PRP$), SP-Op(PRP), SP-Op(R), SP-Op(RB), SP-Op(RBR), SP
-Op(RBS), SP-Op(RP), SP-Op(SYM), SP-Op(TO), SP-Op(UH), SP-Op(V), SP-Op(VB), SP-Op(VBD), SP-Op(VBG), SP-Op(VBN), SP-Op(VBP), SP-Op(VBZ), SP-Op(W), SP-Op(WDT), SP-Op(WP$), SP-Op(WP),
 SP-Op(WRB), SP-Op(``), SP-cNIST, SP-cNIST-1, SP-cNIST-2, SP-cNIST-3, SP-cNIST-4, SP-cNIST-5, SP-cNISTi-2, SP-cNISTi-3, SP-cNISTi-4, SP-cNISTi-5, SP-iobNIST, SP-iobNIST-1, SP-iobNI
ST-2, SP-iobNIST-3, SP-iobNIST-4, SP-iobNIST-5, SP-iobNISTi-2, SP-iobNISTi-3, SP-iobNISTi-4, SP-iobNISTi-5, SP-lNIST, SP-lNIST-1, SP-lNIST-2, SP-lNIST-3, SP-lNIST-4, SP-lNIST-5, SP
-lNISTi-2, SP-lNISTi-3, SP-lNISTi-4, SP-lNISTi-5, SP-pNIST, SP-pNIST-1, SP-pNIST-2, SP-pNIST-3, SP-pNIST-4, SP-pNIST-5, SP-pNISTi-2, SP-pNISTi-3, SP-pNISTi-4, SP-pNISTi-5, SR-Fr(*)
, SR-MFr(*), SR-MPr(*), SR-MRr(*), SR-Mr(*), SR-Mr(*)_b, SR-Mr(*)_i, SR-Mr(A0), SR-Mr(A1), SR-Mr(A2), SR-Mr(A3), SR-Mr(A4), SR-Mr(A5), SR-Mr(AA), SR-Mr(AM-ADV), SR-Mr(AM-CAU), SR-M
r(AM-DIR), SR-Mr(AM-DIS), SR-Mr(AM-EXT), SR-Mr(AM-LOC), SR-Mr(AM-MNR), SR-Mr(AM-MOD), SR-Mr(AM-NEG), SR-Mr(AM-PNC), SR-Mr(AM-PRD), SR-Mr(AM-REC), SR-Mr(AM-TMP), SR-Mrv(*), SR-Mrv(*
)_b, SR-Mrv(*)_i, SR-Mrv(A0), SR-Mrv(A1), SR-Mrv(A2), SR-Mrv(A3), SR-Mrv(A4), SR-Mrv(A5), SR-Mrv(AA), SR-Mrv(AM-ADV), SR-Mrv(AM-CAU), SR-Mrv(AM-DIR), SR-Mrv(AM-DIS), SR-Mrv(AM-EXT)
, SR-Mrv(AM-LOC), SR-Mrv(AM-MNR), SR-Mrv(AM-MOD), SR-Mrv(AM-NEG), SR-Mrv(AM-PNC), SR-Mrv(AM-PRD), SR-Mrv(AM-REC), SR-Mrv(AM-TMP), SR-Nv, SR-Ol, SR-Or, SR-Or(*), SR-Or(*)_b, SR-Or(*
)_i, SR-Or(A0), SR-Or(A1), SR-Or(A2), SR-Or(A3), SR-Or(A4), SR-Or(A5), SR-Or(AA), SR-Or(AM-ADV), SR-Or(AM-CAU), SR-Or(AM-DIR), SR-Or(AM-DIS), SR-Or(AM-EXT), SR-Or(AM-LOC), SR-Or(AM
-MNR), SR-Or(AM-MOD), SR-Or(AM-NEG), SR-Or(AM-PNC), SR-Or(AM-PRD), SR-Or(AM-REC), SR-Or(AM-TMP), SR-Or_b, SR-Or_i, SR-Orv, SR-Orv(*), SR-Orv(*)_b, SR-Orv(*)_i, SR-Orv(A0), SR-Orv(A
1), SR-Orv(A2), SR-Orv(A3), SR-Orv(A4), SR-Orv(A5), SR-Orv(AA), SR-Orv(AM-ADV), SR-Orv(AM-CAU), SR-Orv(AM-DIR), SR-Orv(AM-DIS), SR-Orv(AM-EXT), SR-Orv(AM-LOC), SR-Orv(AM-MNR), SR-O
rv(AM-MOD), SR-Orv(AM-NEG), SR-Orv(AM-PNC), SR-Orv(AM-PRD), SR-Orv(AM-REC), SR-Orv(AM-TMP), SR-Orv_b, SR-Orv_i, SR-Ov, SR-Pr(*), SR-Rr(*) }
```

## Asiya interfaces



**Evaluate the results on-line**

1. Asiya Interface

   http://asiya.lsi.upc.edu/demo/asiya_online.php

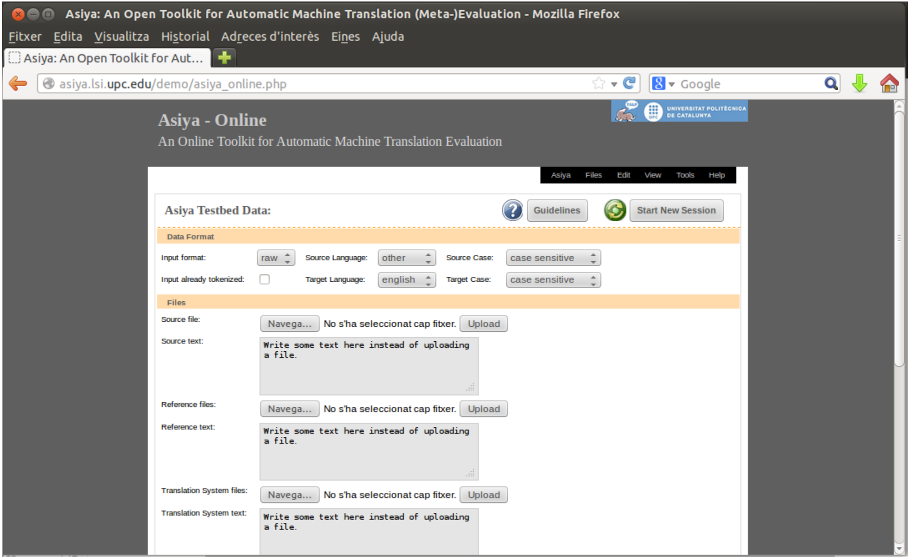
**Analise the results on-line**

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



# Introduction to Machine Translation Evaluation Outline

1. Basics
2. Manual Evaluation
3. Automatic Evaluation
4. Tools
   - Software
   - Demo
5. References

**Machine Translation Evaluation Resources and Methods: A Survey**

Lifeng Han, Derek F. Wong, Lidia S. Chao

https://arxiv.org/pdf/1605.04515v7.pdf

# References

**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]

# References

**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]

# References I


Satanjeev Banerjee and Alon Lavie.
METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments.
In *Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, 2005.

Jacob Cohen.
A coefficient of agreement for nominal scales.
*Educational and Psychological Measurement*, 20(1):37–46, 1960.

George Doddington.
Automatic evaluation of machine translation quality using n-gram co-occurrence statistics.
In *Proceedings of the 2nd Internation Conference on Human Language Technology*, pages 138–145, 2002.

Christian Federmann.
Appraise: An open-source toolkit for manual evaluation of machine translation output.
*The Prague Bulletin of Mathematical Linguistics*, 98:25–35, September 2012.

# References II


Jesús Giménez and Enrique Amigó.
IQMT: A Framework for Automatic Machine Translation Evaluation.
In *Proceedings of the 5th LREC*, pages 685–690, 2006.

J. R. Landis and G. G. Koch.
The measurement of observer agreement for categorical data.
*Biometrics*, 33(1):159–174, 1977.

Chin-Yew Lin and Franz Josef Och.
Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statics.
In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2004.

Gregor Leusch, Nicola Ueffing, and Hermann Ney.
CDER: Efficient MT Evaluation Using Block Movements.
In *Proceedings of EACL*, pages 241–248, 2006.

I. Dan Melamed, Ryan Green, and Joseph P. Turian.
Precision and Recall of Machine Translation.
In *Proceedings of the Joint Conference on Human Language Technology and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, 2003.

# References III

Sonja Nießen, Franz Josef Och, Gregor Leusch, and Hermann Ney.
An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research.
In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, 2000.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu.
Bleu: a method for automatic evaluation of machine translation.
In *Proceedings of the Association of Computational Linguistics*, pages 311–318, 2002.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, , and John Makhoul.
A Study of Translation Edit Rate with Targeted Human Annotation.
In *Proceedings of AMTA*, pages 223–231, 2006.

C. Tillmann, S. Vogel, H. Ney, A. Zubiaga, and H. Sawaf.
Accelerated DP based Search for Statistical Translation.
In *Proceedings of European Conference on Speech Communication and Technology*, 1997.