A Brief Introduction to Machine Translation Evaluation

Francisco Guzmán
ALT Research Group
Qatar Computing Research Institute (QCRI)

MT Marathon of the Americas
Urbana-Champaign, IL, USA
May 12, 2015
to **Cristina España i Bonet** and **Lluís Màrquez**
for some of the slides

to the **QCRI-ALT** group
for their feedback
What to Expect Today

- Why is evaluating MT a hard task?
- How do we (humans) evaluate translations?
- What are different approaches for automatic MT eval?
- What are (dis-)advantages of automatic MT eval?
Can You Evaluate This Translation?

**Source:**
Renzi logra una nueva ley electoral para dar estabilidad a Italia

**Candidate/Hypothesis:**
Renzi achieved a new electoral law to give stability to Italy
What Makes a Good Translation?

According to professional translators, it all depends...

- guidelines (i.e. client requirements)
- genre (e.g. news, blog)
- style (e.g. humorous, wordy, scientific)
- localization (e.g. tailored for target audience)
- ...

Not an easy task!
Difficulties of MT Evaluation

- **Machine Translation** is an *open* NLP task
  - the *correct translation* is not unique
  - the set of admissible translations can be large
  - translation correctness is not black and white
Difficulties of MT Evaluation

Machine Translation is an open NLP task

- the correct translation is not unique
- the set of admissible translations can be large
- translation correctness is not black and white

Evaluation is necessary in the MT system development cycle
Motivation

What Makes a Good Automatic Translation?

Idea: Compare MT output to a human reference

Source:
Renzi logra una nueva ley electoral para dar estabilidad a Italia

Candidate/Hypothesis:
Renzi achieved a new electoral law to give stability to Italy

Reference:
Renzi passed new electoral law aimed to stabilize Italy
What Makes a Good Automatic Translation?

Idea: Compare MT output to a human reference.

Source:
Renzi logra una nueva ley electoral para dar estabilidad a Italia

Candidate/Hypothesis:
Renzi achieved a new electoral law to give stability to Italy

Reference:
Renzi passed new electoral law aimed to stabilize Italy

This is a simpler task
Motivation 10

MT Evaluation

**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
**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.

Two possibilities: *manual* and *automatic evaluation*.
Talk Overview

1. Motivation
3. Automatic Evaluation
4. Recent advances
5. Conclusions
6. Extra slides
Different Views on Quality

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.
Manual Evaluation: 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

Other examples: NIST
Ranking

**Pairwise**

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

**N-way**

Annotators rank $n$ anonymised systems, randomly selected and randomly ordered.
Ranking with Appraise

(Federmann, 2012)

Appraise

You do want ice cream luminous in the darkness?
— Translation 1

You want to glowing in the dark ice cream?
— Translation 2

You want the luminous in the dark ice cream?
— Translation 3

Want luminous in the dark ice cream?
— Translation 4

Want to illuminate the Dark with Ice Cream?
— Translation 5
Ranking is better

**Advantages:**

- Conceptually easier to rank
- Higher agreement among annotators
  (Callison-Burch et al., 2007)
- No scales to be defined

**Disadvantages:**

- Less information is provided
Manual Evaluation

HTER

Human-targeted Translation Error Rate, HTER

**Annotation**  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

\[
HTER = \frac{\text{Substitutions} + \text{Insertions} + \text{Deletions} + \text{Shifts}}{\text{ReferenceWords}}
\]
Evaluation matters!

Progress in the field is measured by evaluation campaigns:

- **NIST**  Open Machine Translation Evaluation
- **WMT**  Workshop Machine Translation
- **IWSLT**  International Workshop on Spoken Language Translation
Human Evaluation Shortcomings

- Subjective
- Costly
- Non-reusable
Human Evaluation Shortcomings

- Subjective
- Costly
- Non-reusable
Talk Overview

1. Motivation
3. Automatic Evaluation
4. Recent advances
5. Conclusions
6. Extra slides
Reference-based Automatic Evaluation (RAE)

**Setting**

⇒ Compute similarity between MT *system’s output* (Hyp) and one or several *reference* translations (Ref)

**Source**

Es un plan de acción que asegura que el Ejército siempre cumpla las órdenes del partido

**Hypothesis**

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-based Automatic Evaluation (RAE)

Setting

⇒ Compute similarity between MT system’s output (Hyp) and one or several reference translations (Ref)

Source  Es un plan de acción que asegura que el Ejército siempre cumpla las órdenes del partido

Hypothesis  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.

Challenge

⇒ The similarity measure should be able to discriminate whether the two sentences convey the same meaning
Desiderata for MT Metrics
(Lavie, 2009)

- **Human-like**: High-levels of correlation with quantified human notions of translation quality
- **Fine-grained**: Sensitivity to small differences in MT quality between systems and versions of systems
- **Consistency**: Same MT system on similar texts should produce similar scores
- **Reliability**: MT systems that score similarly will perform similarly
- **Lightweight**: Fast, easy to run
Desiderata for MT Metrics (Lavie, 2009)

**Human-like** High-levels of correlation with quantified human notions of translation quality

**Fine-grained** Sensitivity to small differences in MT quality between systems and versions of systems

**Consistency** Same MT system on similar texts should produce similar scores

**Reliability** MT systems that score similarly will perform similarly

**Lightweight** Fast, easy to run
Different Levels of Analysis

- Lexical (words)
- Syntactic
- Semantic
- Pragmatic (discourse)
Lexical Matching

First approaches

⇒ **Lexical similarity** as a measure of quality
  - word $n$-gram matching, edit distance, etc.
  - **BLEU**, NIST, TER, Meteor, Rouge, etc.
  - (Papineni et al., 2002; Doddington, 2002; Snover et al., 2006; Lavie & Agarwal 2007; Lin, 2004; etc.)
Lexical Matching

First approaches

⇒ **Lexical similarity** as a measure of quality
  - word \(n\)-gram matching, edit distance, etc.
  - **BLEU**, NIST, TER, Meteor, Rouge, etc.
  - (Papineni et al., 2002; Doddington, 2002; Snover et al., 2006; Lavie & Agarwal 2007; Lin, 2004; etc.)

Nowadays, BLEU is accepted as *the de-facto* 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.”
Automatic evaluation
IBM BLEU: Papineni, Roukos, Ward and Zhu (2001)

BiLingual Evaluation Understudy, BLEU

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

- Precision at different levels (n=1: unigrams, n=2: bigrams, etc)
- Geometric average of \( P_n \) (empirical suggestion)
- \( w_n \) positive weights summing to one (typically \( 1/N \))
- Brevity penalty
Hypothesis:

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.
Hypothesis:

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.
**IBM BLEU**

**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.
```
IBM BLEU

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

Precision-based measure, but:

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

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

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.
IBM BLEU

Modified n-gram precision (1-gram)

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

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

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: \[ \text{Prec.} = \frac{5 + 7}{7} \]

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 \textit{n-gram precision} (1-gram)

Precision-based measure, but:

\[
\text{Prec.} = \frac{6 + 7}{7}
\]

Candidate:

\textit{The the the the the the the the.}

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

Reference 2:
\textit{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 the.

Reference 1:

The cat is on the mat.

Reference 2:

There is a cat on the mat.
**IBM BLEU**

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

A reference word should only be matched once.

**Algorithm:**

1. Count number of times $w_i$ occurs in the candidate.
2. Keep the minimum of (1) and the maximum number of times $w_i$ appears in any reference (*clipping*).
3. Add these values and divide by candidate’s number of words.
**IBM BLEU**

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

**Modified 1-gram precision:**

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

1. \( w_i \rightarrow \text{The} \)
   \#\( w_i,R_1 \) = 2
   \#\( w_i,R_2 \) = 1
   \#\( w_i,C \) = 7

2. \( \max(R^*)=2, \)
   \( \Rightarrow \ min(R^*,C)=2 \)

3. No more distinct words
**IBM BLEU**

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

Modified 1-gram precision: \[ P_1 = \]

**Candidate:**

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

**Reference 1:**

\[ \text{The cat is on the mat} \]

**Reference 2:**

\[ \text{There is a cat on the mat} \]

1. \( w_i \rightarrow \text{The} \)
   \( \# w_i,R_1 = 2 \)
   \( \# w_i,R_2 = 1 \)
   \( \# w_i,c = 7 \)

2. \( \text{Max}(R^*),c = 2 \)
   \( \Rightarrow \text{Min}(R^*,c) = 2 \)

3. \( \text{No more distinct words} \)
IBM BLEU

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. \( w_i \rightarrow \text{The} \)
   \# \( w_i, R_1 = 2 \)
   \# \( w_i, R_2 = 1 \)
   \# \( w_i, c = 7 \)

2. \( \text{Max}(R^*) = 2, \Rightarrow \text{Min}(R^*, c) = 2 \)

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. \( w_i \rightarrow \text{The} \)
   \( \# w_i, R_1 = 2 \)
   \( \# w_i, R_2 = 1 \)
   \( \# w_i, C = 7 \)

2. \( \text{Max}(R^*) = 2 \)
   \( \Rightarrow \text{Min}(R^*, C) = 2 \)

3. No more distinct words
Generalisation to multiple sentences:

\[
P_n = \frac{\sum_{C \in \{\text{candidates}\}} \sum_{\text{ngram} \in C} \text{Count}_{\text{clipped}}(\text{ngram})}{\sum_{C \in \{\text{candidates}\}} \sum_{\text{ngram} \in C} \text{Count}(\text{ngram})}
\]

low \(n\) adequacy

high \(n\) fluency
Automatic evaluation
IBM BLEU: Papineni, Roukos, Ward and Zhu (2001)

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

\[ P_1 = \frac{2}{2}, \quad P_2 = \frac{1}{1} \]

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
Brevity penalty

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

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

*c* candidate length, *r* reference length
Sometimes we want to evaluate BLEU at the sentence level. This can lead to trouble:

- **Problem**
  - Precision: Zero matches = Zero score

- **Solution**
  - Smooth Precision: Add +1 to precision counts
  - Smooth BP: Add +1 to reference component
Limits of lexical similarity

Hyp: 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.
Limits of lexical similarity

Hyp: 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.
Limits of lexical similarity

Hyp: 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.
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
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
Problems of Lexical Similarity Measures

- Lexical similarity is not a sufficient nor a necessary condition so that two sentences express the same meaning (Culy and Riehemann, 2003; Coughlin, 2003; Callison-Burch et al., 2006)

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

- Lexical metrics have problems distinguishing MT output from fully fluent and adequate translations obtained from them through professional postediting (Denkowski and Lavie, 2012)
Problems of Lexical Similarity Measures

NIST 2005 Arabic-to-English Exercise
(Callison-Burch et al., 2006; Koehn and Monz, 2006)
Problems of Lexical Similarity Measures

NIST 2005 Arabic-to-English Exercise
(Callison-Burch et al., 2006; Koehn and Monz, 2006)
Problems of Lexical Similarity Measures

NIST 2005 Arabic-to-English Exercise
(Callison-Burch et al., 2006; Koehn and Monz, 2006)

⇒ \( n \)-gram based metrics favor MT systems which closely replicate the lexical realization of the references

⇒ Test sets tend to be similar (domain, register, sublanguage) to training materials

⇒ Statistical MT systems heavily rely on the training data

⇒ Statistical MT systems tend to share the reference sublanguage and be favored by \( n \)-gram based measures
Problems of Lexical Similarity Measures

NIST 2005 Arabic-to-English Exercise
(Callison-Burch et al., 2006; Koehn and Monz, 2006)

⇒ $n$-gram based metrics favor MT systems which closely replicate the lexical realization of the references

⇒ Test sets tend to be similar (domain, register, sublanguage) to training materials

⇒ Statistical MT systems heavily rely on the training data

⇒ Statistical MT systems tend to share the reference sublanguage and be favored by $n$-gram based measures
Linguistic Generalization

Active area of research

⇒ Generalization over lexical matching and usage of more complex linguistic information to compute similarity

▷ stemming, synonymy, paraphrasing, etc.

▷ shallow parsing, constituency and dependency parsing, named entities, semantic roles, textual entailment, etc.

▷ discourse trees
Existing Metrics

Lexical Precision
Lexical Recall
F-measure
Edit Distance

Lexical Similarity

PoS Tagging
Dependency Parsing
Named Entities

Syntactic Similarity

Semantic Roles
Constituency Parsing

Semantic Similarity

Discourse Representations
Lemma-tization
Chunking

Francisco Guzmán
A Brief Introduction to MT Evaluation

MT Marathon 2015
Existing Metrics

Lexical Precision
- WNM
- NIST
- BLEU
- CDER
- SIA
- METEOR
- SP-NISTp
- BLEUp
- SP-NISTc
- BLANC
- Edit Distance
- PER
- WER

Lexical Recall
- ROUGE
- SP-Op*
- PoS Tagging
- DP-Or*
- CHunking
- Lemma-ization
- CP-Op*
- SP-NISTI

Syntactic Similarity
- Dependency Parsing
- DP-OI*
- DP-Oc*
- Constituency Parsing
- SP-Oc*
- CP-Oc*
- STM

Semantic Similarity
- NEE
- NE-Or*
- NE-Me*
- NER
- HWCM
- MAXSIM
- SR-Or
- SR-Mr*
- DR-Or*
- DR-Orp*

Francisco Guzmán
A Brief Introduction to MT Evaluation
MT Marathon 2015
Talk Overview

1. Motivation
3. Automatic Evaluation
4. Recent advances
5. Conclusions
6. Extra slides
Which one is better?

**Idea:** Measure the correlation of evaluation metrics with human judgments (e.g. Appraise)

Campaigns:

- metricsMATR (NIST)
- WMT metrics

<table>
<thead>
<tr>
<th>Metric</th>
<th>Orig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEMPOS</td>
<td>.902</td>
</tr>
<tr>
<td>AMBER</td>
<td>.857</td>
</tr>
<tr>
<td>METEOR</td>
<td>.834</td>
</tr>
<tr>
<td>TERRORCAT</td>
<td>.831</td>
</tr>
<tr>
<td>SIMPBLEU</td>
<td>.823</td>
</tr>
<tr>
<td>TER</td>
<td>.812</td>
</tr>
<tr>
<td>BLEU</td>
<td>.810</td>
</tr>
<tr>
<td>POSF</td>
<td>.754</td>
</tr>
<tr>
<td>NIST</td>
<td>.817</td>
</tr>
<tr>
<td>Asiya-LEX</td>
<td>.879</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>
Which one is better?

**Idea**: Measure the correlation of evaluation metrics with human judgments (e.g., Appraise)

**Campaigns**:

- **metricsMATR** (NIST)
- **WMT metrics**

<table>
<thead>
<tr>
<th>Metric</th>
<th>Orig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEMPOS</td>
<td>.902</td>
</tr>
<tr>
<td>AMBER</td>
<td>.857</td>
</tr>
<tr>
<td><strong>METEOR</strong></td>
<td>.834</td>
</tr>
<tr>
<td>TERRORCAT</td>
<td>.831</td>
</tr>
<tr>
<td>SIMBLEU</td>
<td>.823</td>
</tr>
<tr>
<td>TER</td>
<td>.812</td>
</tr>
<tr>
<td>BLEU</td>
<td>.810</td>
</tr>
<tr>
<td>POSF</td>
<td>.754</td>
</tr>
<tr>
<td></td>
<td>...</td>
</tr>
<tr>
<td>NIST</td>
<td>.817</td>
</tr>
<tr>
<td></td>
<td>...</td>
</tr>
<tr>
<td>Asiya-LEX</td>
<td>.879</td>
</tr>
</tbody>
</table>
In Germany voices, suggesting that ECB be the last resort creditor.

In Germany the ECB should be for the creditors of last resort.
Setting

- Discourse structures: computed at sentence level with the RST-based parser from Joty et al. (2012)

- Similarity: computed with STK kernel (Collins & Duffy, 2001)
  \[ \Rightarrow \text{the similarity is the sum of all common sub-trees} \]
Recent advances

Setting

- Discourse structures: computed at sentence level with the RST-based parser from Joty et al. (2012)

- Similarity: computed with STK kernel (Collins & Duffy, 2001)

  ⇒ the similarity is the sum of all common sub-trees
Recent advances

Untuned combinations

[WMT12, into-en, system-level, ρ]

- Combination with other existing evaluation metrics
- Other smarter ways are possible.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Orig.</th>
<th>+DR-LEX</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>DR-LEX</td>
<td>0.876</td>
</tr>
<tr>
<td>II</td>
<td>SEMPOS</td>
<td>0.902</td>
</tr>
<tr>
<td>II</td>
<td>AMBER</td>
<td>0.857</td>
</tr>
<tr>
<td>II</td>
<td>METEOR</td>
<td>0.834</td>
</tr>
<tr>
<td>II</td>
<td>TERRORCAT</td>
<td>0.831</td>
</tr>
<tr>
<td>II</td>
<td>SIMPBLEU</td>
<td>0.823</td>
</tr>
<tr>
<td>II</td>
<td>TER</td>
<td>0.812</td>
</tr>
<tr>
<td>II</td>
<td>BLEU</td>
<td>0.810</td>
</tr>
<tr>
<td>II</td>
<td>POSF</td>
<td>0.754</td>
</tr>
<tr>
<td>III</td>
<td>NIST</td>
<td>0.817</td>
</tr>
<tr>
<td>IV</td>
<td>Asiya-LEX</td>
<td>0.879</td>
</tr>
<tr>
<td>average</td>
<td>.839</td>
<td>.874</td>
</tr>
<tr>
<td>diff.</td>
<td></td>
<td>+.035</td>
</tr>
</tbody>
</table>
Talk Overview

1. Motivation
3. Automatic Evaluation
4. Recent advances
5. Conclusions
6. Extra slides
Virtues and curse

⇒ Automatic evaluation metrics have notably accelerated the development cycle of MT systems
  ◦ Cheap, objective and reusable
  ◦ Used for error analysis, system optimization, system comparison, etc.

⇒ Risks of Automatic Evaluation
  ◦ System over-tuning
  ◦ Blind system development
  ◦ Unfair system comparisons
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.
Thank you!

A Brief Introduction to Machine Translation Evaluation

Francisco Guzmán
ALT Research Group
Qatar Computing Research Institute (QCRI)

MT Marathon of the Americas
Urbana-Champaign, IL, USA
May 12, 2015
**Goal** Instead of adjusting weights of already existing metrics, we want to work in a unified learning framework, able to represent many layers of linguistic information and able to learn from fine-grained features.

- Two alternatives for the input representation:
  - *Structured* (with kernel-based learning)
  - *Distributed* (with ANN learning)

- Common setting: pairwise quality comparison
Differentiating *better* from *worse* translation

- **Input:** \( \langle t_1, t_2, r \rangle \)
  
  \[ \Rightarrow \text{"Is } t_1 \text{ a better translation than } t_2, \text{ given } r\"? \]

- **Pairwise ranking setting**
  
  \[ \Rightarrow \text{closer to the evaluation that humans do better} \]
  
  \[ \Rightarrow \text{valid for most MT comparison/ranking tasks} \]
  
  \[ \Rightarrow \text{not an absolute quality score} \]
Learning with preference kernels

- Tree-based representation of all layers of information
- Pairwise ranking with the preference kernel (Shen & Joshi, 2003)
- Learning example: $\langle h_1, h_2 \rangle = \langle \phi_M(t_1, r), \phi_M(t_2, r) \rangle$
  \[ \Rightarrow \quad \phi_M \text{ makes a structured and relational representation of } t \text{ and } r \]
  \[ \Rightarrow \quad \phi_M(t_1, r) = \langle t_1^r, r^{t_1} \rangle \]
  \[ \Rightarrow \quad \text{two separate trees instead of a graph} \]
Learning with preference kernels: $\phi_M(t, r)$

Guzmán et al, EMNLP2014

---

Extra slides

---

Francisco Guzmán
A Brief Introduction to MT Evaluation

MT Marathon 2015
Learning with preference kernels (II)

Guzmán et al, EMNLP2014

- Learning example: \( \langle h_1, h_2 \rangle = \langle \phi_M(t_1, r), \phi_M(t_2, r) \rangle \)

- Preference kernel

  \[
  PK(\langle h_1, h_2 \rangle, \langle h'_1, h'_2 \rangle) = K(h_1, h'_1) + K(h_2, h'_2) - K(h_1, h'_2) - K(h_2, h'_1)
  \]

  \[
  K(h_1, h'_1) = PTK(t'_1, t'_1) + PTK(r_{t_1}, r_{t'_1})
  \]

  - PTK = Partial Tree Kernel (Shen & Joshi, 2003)

  - PTK = Partial Tree Kernel (Moschitti, 2006)
Learning with preference kernels (II)

- Learning example: $\langle h_1, h_2 \rangle = \langle \phi_M(t_1, r), \phi_M(t_2, r) \rangle$

- Preference kernel (Shen & Joshi, 2003)

\[
PK(\langle h_1, h_2 \rangle, \langle h'_1, h'_2 \rangle) = \\
K(h_1, h'_1) + K(h_2, h'_2) - K(h_1, h'_2) - K(h_2, h'_1)
\]

\[
K(h_1, h'_1) = PTK(t'_1, t'_1) + PTK(r^{t_1}, r^{t'_1})
\]

- PTK = Partial Tree Kernel (Moschitti, 2006)
Input mapped to fixed-length vectors \([x_{t1}, x_{t2}, x_r]\) using syntactic (Stanford’s parser) and semantic embeddings (a la ‘word2vec’).
Hidden layer to compute three types of interactions: \( \text{sim}(t_1, r) \), \( \text{sim}(t_2, r) \), and \( \text{sim}(t_1, t_2) \).
External sources of information as direct features (*skip arcs*). We plug in BLEU, NIST, TER, and METEOR scores.