# Translation Quality Assessment: Evaluation and Estimation

Lucia Specia

University of Sheffield
l.specia@sheffield.ac.uk

MTM - Prague, 12 September 2016

# Outline

# Outline

# Why do we care?

... or why is this the first lecture of the Marathon?

In the business of **developing MT**, we need to

- measure progress over new/alternative versions
- compare different MT systems
- decide whether a translation is good enough for something
- optimise parameters of MT systems
- understand where systems go wrong (diagnosis)

# Why do we care?

- One should optimise a system using the same metric that will be used to evaluate it
- **Issue**: how to choose a metric? Choice should be related to the **purpose** of the system will be used (not the case in practice)
- Other aspects are important for **tuning** (sentence/corpus-level, fast, cheap, differentiable, ...)

# Complex problem

"MT evaluation is better understood than MT"
(Carbonell and Wilks, 1991)

# Complex problem

"MT evaluation is better understood than MT"
(Carbonell and Wilks, 1991)

# Complex problem

"MT evaluation is better understood than MT"
(Carbonell and Wilks, 1991)

"There are more MT evaluation metrics than MT approaches"
(Specia, 2016)

# Complex problem

- What does **quality** mean?
  - Fluent? Adequate? Both?
  - Easy to post-edit?
  - System A better than system B?
  - ...

# Complex problem

- What does **quality** mean?
  - Fluent? Adequate? Both?
  - Easy to post-edit?
  - System A better than system B?
  - ...
- Quality for **whom**/**what**?
  - End-user (gisting vs dissemination)
  - Post-editor (light vs heavy post-editing)
  - Other applications (e.g. CLIR)
  - MT-system (tuning or diagnosis for improvement)
  - ...

# Complex problem

MT Do buy this product, it's their craziest invention!

# Complex problem

MT Do buy this product, it's their craziest invention!

HT Do **not** buy this product, it's their craziest invention!

# Complex problem

MT Do buy this product, it's their craziest invention!

HT Do **not** buy this product, it's their craziest invention!

- **Severe** if end-user does not speak source language
- **Trivial** to post-edit by translators

# Complex problem

MT Six-hours battery, **30 minutes** to **full charge** last.

# Complex problem

MT Six-hours battery, **30 minutes** to **full charge** last.

HT The battery lasts 6 hours and it can be **fully recharged** in **30 minutes**.

# Complex problem

MT Six-hours battery, **30 minutes** to **full charge** last.

HT The battery lasts 6 hours and it can be **fully recharged** in **30 minutes**.

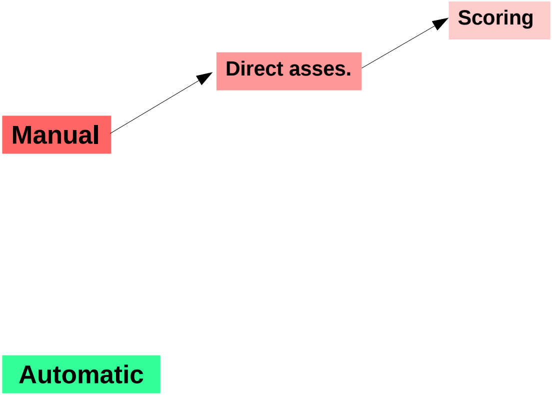
- **Ok** for gisting - meaning preserved
- **Very costly** for post-editing if style is to be preserved

# A taxonomy of MT evaluation methods

**Manual**

**Automatic**

# A taxonomy of MT evaluation methods



# A taxonomy of MT evaluation methods

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

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

| Translation | Adequacy | Fluency |
|---|---|---|
| both countries are rather a necessary laboratory the internal operation of the eu . | 1 2 3 4 5 | 1 2 3 4 5 |

Is this translation correct?

Read the text below. How much do you agree with the following statement:

**The black text adequately expresses the meaning of the gray text in English.**

To snobs like me who declare that they'd rather play sports than watch them, it's hard to see the appeal of watching games rather than taking up a controller myself.

**Snob like me, who say that it is better to be in sports than watching him, it is hard to understand the appeal of having to watch the game, rather than to take a joystick in hand.**

0 % 100 %

# A taxonomy of MT evaluation methods

**Scoring**

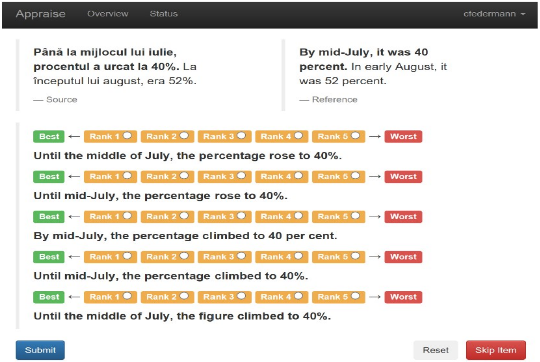

**Direct asses.**

**Ranking**

**Manual**

**Automatic**

# A taxonomy of MT evaluation methods



# A taxonomy of MT evaluation methods

**Manual**

**Direct asses.**

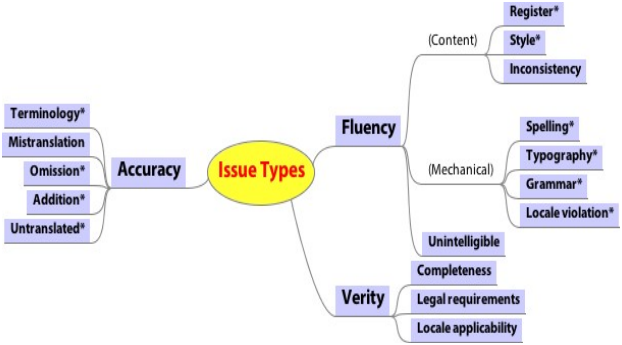
**Scoring**  

**Ranking** 

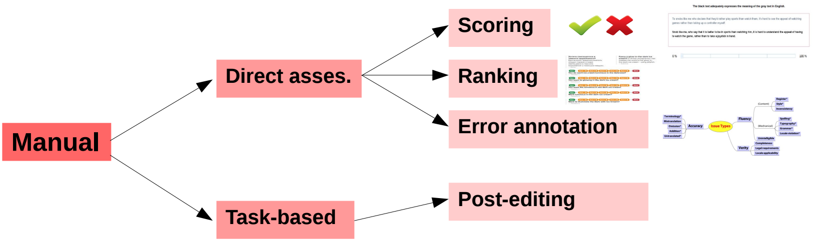
**Error annotation**

**Automatic**

# A taxonomy of MT evaluation methods

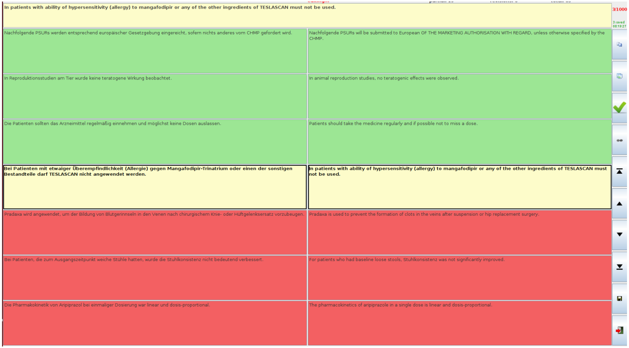


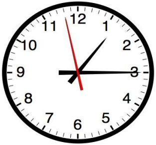
# A taxonomy of MT evaluation methods




Automatic

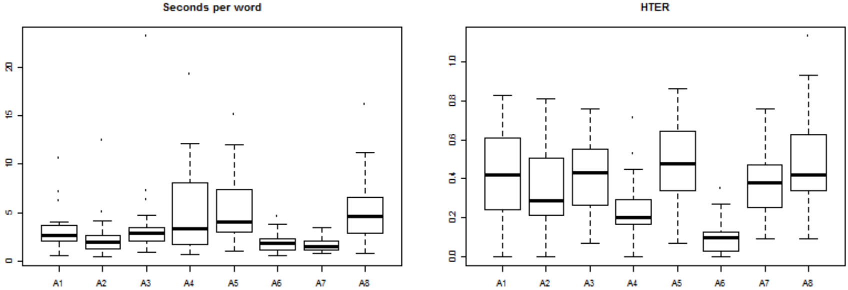
# A taxonomy of MT evaluation methods



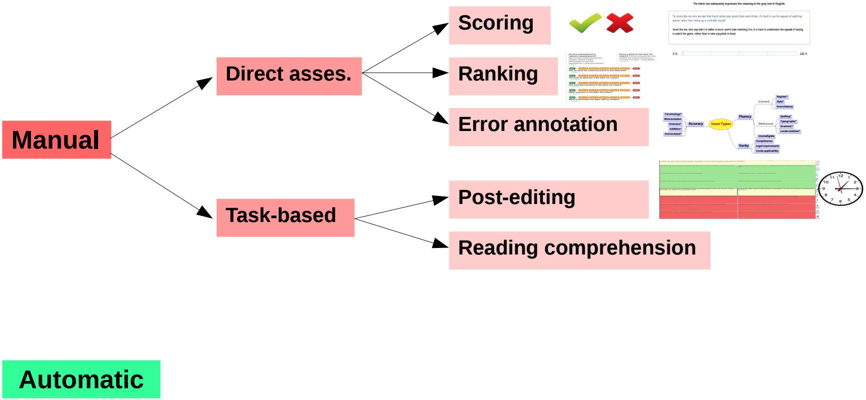



**HTER**

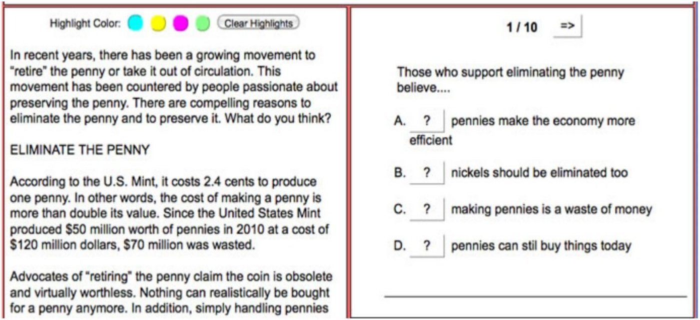
# A taxonomy of MT evaluation methods



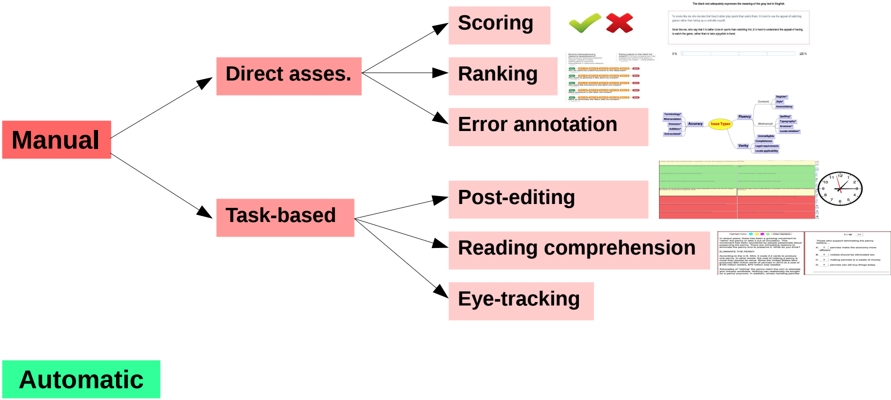
# A taxonomy of MT evaluation methods



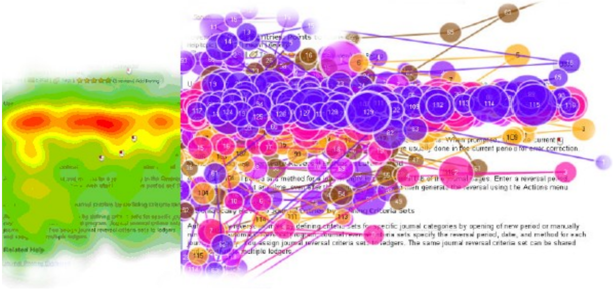
# A taxonomy of MT evaluation methods



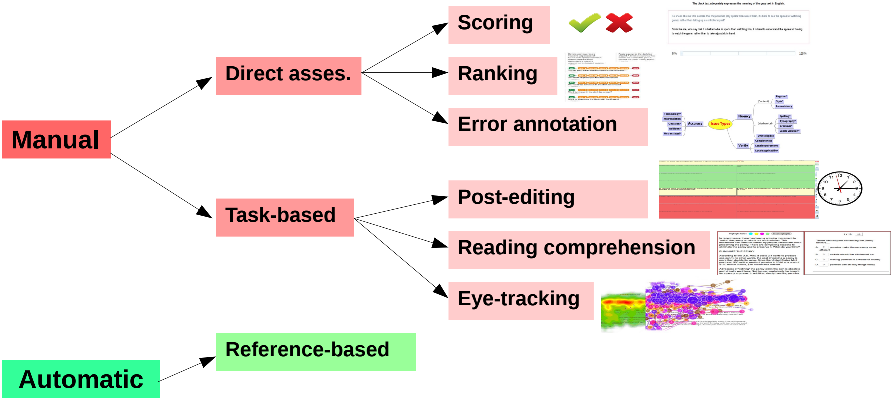
# A taxonomy of MT evaluation methods



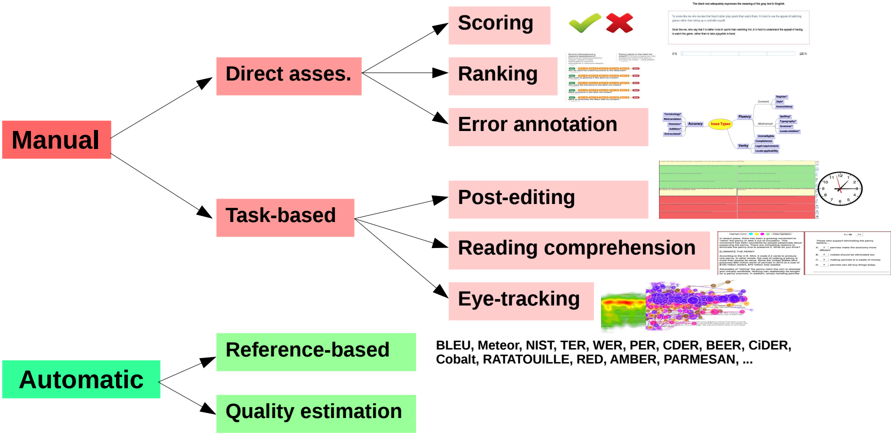
# A taxonomy of MT evaluation methods



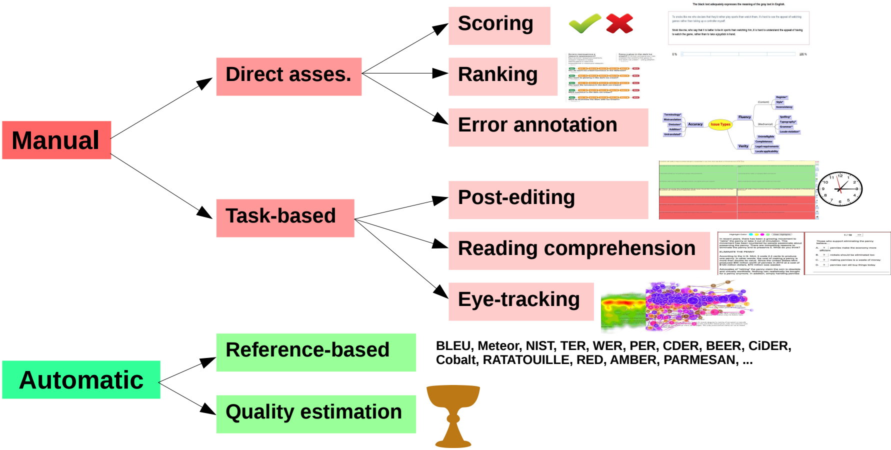
# A taxonomy of MT evaluation methods



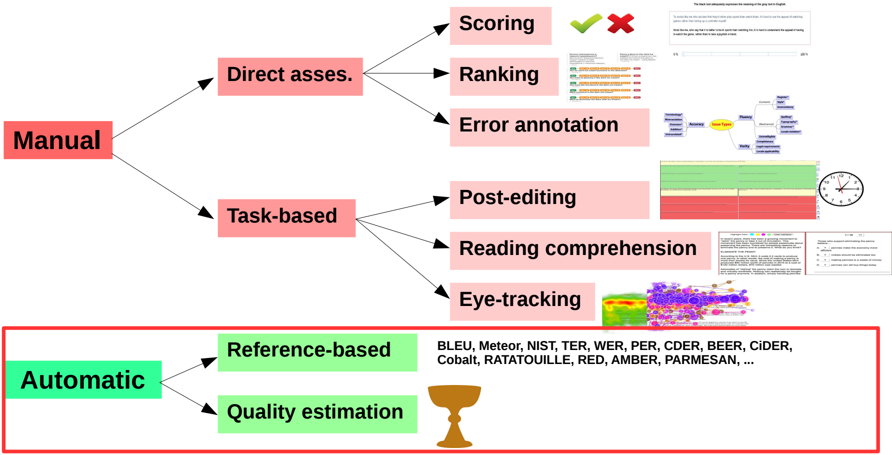
# A taxonomy of MT evaluation methods



# A taxonomy of MT evaluation methods



# A taxonomy of MT evaluation methods



# Outline

# Assumption

The closer an MT system output is to a human translation (HT = **reference**), the better it is.

Which system is better?

$MT_1$ **Indignation** in front of photos **of a veiled woman** controlled **on** the beach in Nice.

$MT_2$ Outrage **at pictures of a veiled woman** controlled **on** the beach in Nice.

---

$HT_a$ **Indignation at pictures of a veiled woman** being checked **on** a beach in Nice.

# Assumption

The closer an MT system output is to a human translation (HT = **reference**), the better it is.

Which system is better?

$MT_1$ **Indignation** in front of photos **of a veiled woman** controlled **on** the beach in Nice.

$MT_2$ Outrage **at pictures of a veiled woman** controlled **on** the beach in Nice.

---

$HT_a$ **Indignation at pictures of a veiled woman** being checked **on** a beach in Nice.

Or, simply, how good is the $MT_1$ system output?

# Assumption

Which system is better?

$MT_1$ Indignation in front of photos of a veiled woman controlled on the beach in Nice.

$MT_2$ Outrage at pictures of a veiled woman controlled on the beach in Nice.

---

$HT_a$ Indignation at pictures of a veiled woman being checked on a beach in Nice.

$HT_b$ **Photos** of a veiled woman checked by the police **on the beach in Nice** cause **outrage**.

# Assumption

Which system is better?

$MT_1$ Indignation in front of photos of a veiled woman controlled on the beach in Nice.

$MT_2$ Outrage at pictures of a veiled woman controlled on the beach in Nice.

---

$HT_a$ Indignation at pictures of a veiled woman being checked on a beach in Nice.

$HT_b$ **Photos** of a veiled woman checked by the police **on the beach in Nice** cause **outrage**.

Or, again, how good is the $MT_1$ system output?

# BLEU

## BLEU: BiLingual Evaluation Understudy

- **Most widely used metric**, both for MT system evaluation/comparison and SMT tuning
- Matching of n-grams between MT and HT: rewards **same words** in **equal order**
- $\#clip(g)$ count of reference n-grams $g$ which happen in a MT sentence $h$ clipped by the number of times $g$ appears in the HT sentence for $h$; $\#(g') =$ number of n-grams in MT output
- n-gram precision $p_n$ for a set of translations in $C$:

$$p_n = \frac{\sum_{c \in C} \sum_{g \in ngrams(c)} \#clip(g)}{\sum_{c \in C} \sum_{g' \in ngrams(c)} \#(g')}$$

# BLEU

- Combine (mean of the log) 1-$n$ $n$-gram precisions

$$\sum_n \log p_n$$

# BLEU

- Combine (mean of the log) 1-$n$ $n$-gram precisions

$$\sum_n \log p_n$$

- Bias towards translations with fewer words
- **Brevity penalty** to penalise MT sentences that are shorter than reference
  - Compares the overall number of words $w_h$ of the entire hypotheses set with ref length $w_r$:

$$BP = \begin{cases} 1 & \text{if } w_c \geq w_r \\ e^{(1-w_r/w_c)} & \text{otherwise} \end{cases}$$

# BLEU

- Combine (mean of the log) 1-$n$ $n$-gram precisions

$$\sum_n \log p_n$$

- Bias towards translations with fewer words
- **Brevity penalty** to penalise MT sentences that are shorter than reference
  - Compares the overall number of words $w_h$ of the entire hypotheses set with ref length $w_r$:

$$BP = \begin{cases} 1 & \text{if } w_c \geq w_r \\ e^{(1-w_r/w_c)} & \text{otherwise} \end{cases}$$

$$BLEU = BP * \exp\left(\sum_n \log p_n\right)$$

# BLEU

- Scale: 0-1, but highly **dependent on the test set**
- Rewards **fluency** by matching high n-grams (up to 4)
- Rewards **adequacy** by unigrams and brevity penalty – poor model of recall
- **Synonyms and paraphrases** only handled if in one of reference translations
- All tokens are **equally weighted**: incorrect content word = incorrect determiner

# BLEU

- Scale: 0-1, but highly **dependent on the test set**
- Rewards **fluency** by matching high n-grams (up to 4)
- Rewards **adequacy** by unigrams and brevity penalty – poor model of recall
- **Synonyms and paraphrases** only handled if in one of reference translations
- All tokens are **equally weighted**: incorrect content word = incorrect determiner
- Better for evaluating changes in the same system than comparing **different MT architectures**

# BLEU

Example:

**MT:** in **two weeks** Iraq's **weapons** will give **army**

**HT:** the Iraqi **weapons** are to be handed over to the **army** within **two weeks**

- 1-gram precision: 4/8
- 2-gram precision: 1/7
- 3-gram precision: 0/6
- 4-gram precision: 0/5

# Edit distance metrics

## TER: Translation Error Rate

- Levenshtein edit distance
- Minimum proportion of **insertions**, **deletions**, and **substitutions** to transform MT sentence into HT
- Adds **shift** operation

# Edit distance metrics

## TER: Translation Error Rate

- Levenshtein edit distance
- Minimum proportion of **insertions**, **deletions**, and **substitutions** to transform MT sentence into HT
- Adds **shift** operation

```
REF:           SAUDI ARABIA denied  this week
information published in the AMERICAN new york times

HYP: [this week] the saudis denied
   information published in the *****    new york times
```

# Edit distance metrics

## TER: Translation Error Rate

- Levenshtein edit distance
- Minimum proportion of **insertions**, **deletions**, and **substitutions** to transform MT sentence into HT
- Adds **shift** operation

```
REF:            SAUDI ARABIA denied  this week
information published in the AMERICAN new york times

HYP: [this week] the saudis denied
  information published in the *****     new york times
```

1 shift, 2 substit., 1 deletion: $\text{TER} = \frac{4}{13} = 0.31$

# Edit distance metrics

## TER: Translation Error Rate

- Levenshtein edit distance
- Minimum proportion of **insertions**, **deletions**, and **substitutions** to transform MT sentence into HT
- Adds **shift** operation

```
REF:           SAUDI ARABIA denied  this week
information published in the AMERICAN new york times

HYP: [this week] the saudis denied
  information published in the *****     new york times
```

1 shift, 2 substit., 1 deletion: $\text{TER} = \frac{4}{13} = 0.31$

### Human-targeted TER (HTER)

TER between MT and its post-edited version

# Alignment-based metrics

**METEOR**:

- Unigram **Precision** and **Recall**
- Align MT & HT
- Matching considers **inflection variants** (stems), **synonyms**, **paraphrases**
- **Fluency** addressed via a direct penalty: fragmentation of the matching
- METEOR score = F-mean score discounted for fragmentation = F-mean * (1 - DF)
- Parameters can be trained

# Alignment-based metrics

**MT:** in **two weeks** **Iraq's weapons** will give **army**

**HT:** the **Iraqi weapons** are to be handed over to the **army** within **two weeks**

# Alignment-based metrics

**MT:** in **two weeks** **Iraq's weapons** will give **army**

**HT:** the **Iraqi weapons** are to be handed over to the **army** within **two weeks**

- Matching:

**MT** **two weeks** **Iraq's weapons** **army**

**HT:** **Iraqi weapons** **army** **two weeks**

# Alignment-based metrics

**MT:** in **two weeks Iraq's weapons** will give **army**

**HT:** the **Iraqi weapons** are to be handed over to the **army** within **two weeks**

- Matching:

**MT** **two weeks Iraq's weapons army**

**HT:** **Iraqi weapons army two weeks**

- P = 5/8 =0.625
- R = 5/14 = 0.357
- F-mean = 10*P*R/(9P+R) = 0.373

# Alignment-based metrics

**MT:** in **two weeks Iraq's weapons** will give **army**

**HT:** the **Iraqi weapons** are to be handed over to the **army** within **two weeks**

- Matching:

**MT** **two weeks Iraq's weapons army**

**HT:** **Iraqi weapons army two weeks**

- P = 5/8 =0.625
- R = 5/14 = 0.357
- F-mean = 10*P*R/(9P+R) = 0.373
- Fragmentation: 3 frags for 5 words = (3)/(5) = 0.6
- Discounting factor: DF = 0.5 * (0.6**3) = 0.108
- **METEOR**: F-mean * (1 - DF) = 0.373 * 0.892 = 0.333

# BEER

## BEER: BEtter Evaluation as Ranking

- **Trained metric**
  $score(h, r) = \sum_i w_i \times \phi_i(h, r) = \vec{\mathbf{w}} \cdot \vec{\phi}$
- Learns from pairwise rankings
- Various features between MT output and reference translation
  - Precision, Recall and F1 over character n-grams (1-6)
  - Idem for word unigrams: content vs function separately
  - Reordering through permutation trees and distance to ideal monotone permutation

# Dozens more....

**Some - WMT metrics task**:

- CharacTer
- chrF/wordF
- TerroCat
- MEANT and TINE
- TESLA
- LEPOR
- ROSE
- AMBER
- Many other linguistically motivated metrics where matching goes beyond word forms
- ...

**Asiya toolkit - up until ∼2014**

# Dozens more....

WMT16 metrics task (by Bojar et al.):

| Metric | # Wins | Language Pairs |
|---|---|---|
| BEER | 11 | csen, encs, ende, enfi, enro, enru, entr, fien, roen, ruen, tren |
| UoW.ReVal | 6 | csen, deen, fien, roen, ruen, tren |
| chrF2 | 6 | csen, encs, enro, entr, fien, ruen |
| chrF1 | 5 | encs, enro, fien, ruen, tren |
| chrF3 | 4 | deen, enfi, entr, ruen |
| mosesCDER | 4 | csen, enfi, enru, entr |
| CharacTer | 3 | csen, deen, roen |
| mosesBLEU | 3 | csen, encs, enfi |
| mosesPER | 3 | enro, ruen, tren |
| mtevalBLEU | 3 | csen, encs, enro |
| wordF1 | 3 | csen, encs, enro |
| wordF2 | 3 | csen, encs, enro |
| mosesTER | 2 | csen, encs |
| mtevalNIST | 2 | encs, tren |
| wordF3 | 2 | csen, entr |
| mosesWER | 1 | csen |

# Problems with reference-based evaluation

- Reference(s): subset of good translations, usually **one**
  Some metrics expand matching, e.g. synonyms in **Meteor**

- Huge **variation** in reference translations. E.g.

| | | | |
|---|---|---|---|
| **Source** | 不过这一切都由不得你<br>*However these all totally beyond the control of you.* | | |
| **MT** | But all this is beyond the control of you. | **Human score** | **BLEU score** |
| **$HT_1$** | But all this is beyond your control. | 3.4 | 0.427 |
| **$HT_2$** | However, you cannot choose yourself. | 2 | 0.049 |
| **$HT_3$** | However, not everything is up to you to decide. | 2 | 0.050 |
| **$HT_4$** | But you can't choose that. | 2.8 | 0.055 |

- Metrics completely disregard **source segment**
- **Cannot** be applied for MT systems in use

# Problems with reference-based evaluation

- Reference(s): subset of good translations, usually **one**
  Some metrics expand matching, e.g. synonyms in **Meteor**

- Huge **variation** in reference translations. E.g.

| | | | |
|---|---|---|---|
| **Source** | 不过这一切都由不得你<br>*However these all totally beyond the control of you.* | | |
| **MT** | But all this is beyond the control of you. | **Human score** | **BLEU score** |
| **$HT_1$** | But all this is beyond your control. | 3.4 | 0.427 |
| **$HT_2$** | However, you cannot choose yourself. | 2 | 0.049 |
| **$HT_3$** | However, not everything is up to you to decide. | 2 | 0.050 |
| **$HT_4$** | But you can't choose that. | 2.8 | 0.055 |

- Metrics completely disregard **source segment**
- **Cannot** be applied for MT systems in use

# Problems with reference-based evaluation

- Reference(s): subset of good translations, usually **one**
  Some metrics expand matching, e.g. synonyms in **Meteor**
- Huge **variation** in reference translations. E.g.

| **Source** | 不过这一切都由不得你<br>*However these all totally beyond the control of you.* | | |
|---|---|---|---|
| **MT** | But all this is beyond the control of you. | **Human score** | **BLEU score** |
| **HT$_1$** | But all this is beyond your control. | 3.4 | 0.427 |
| **HT$_2$** | However, you cannot choose yourself. | 2 | 0.049 |
| **HT$_3$** | However, not everything is up to you to decide. | 2 | 0.050 |
| **HT$_4$** | But you can't choose that. | 2.8 | 0.055 |

- Metrics completely disregard **source segment**
- **Cannot** be applied for MT systems in use

# Outline

# QE - Overview

- **Quality estimation** (QE): metrics that provide an **estimate** on the **quality** of translations *on the fly*

# QE - Overview

- **Quality estimation** (QE): metrics that provide an **estimate** on the **quality** of translations *on the fly*
- Quality defined by the **data**: **purpose** is clear, no comparison to **references**, **source** considered

# QE - Overview

- **Quality estimation** (QE): metrics that provide an **estimate** on the **quality** of translations *on the fly*
- Quality defined by the **data**: **purpose** is clear, no comparison to **references**, **source** considered

Quality = **Can we publish it as is?**

# QE - Overview

- **Quality estimation** (QE): metrics that provide an **estimate** on the **quality** of translations *on the fly*
- Quality defined by the **data**: **purpose** is clear, no comparison to **references**, **source** considered

Quality = **Can we publish it as is?**

Quality = **Can a reader get the gist?**

# QE - Overview

- **Quality estimation** (QE): metrics that provide an **estimate** on the **quality** of translations *on the fly*
- Quality defined by the **data**: **purpose** is clear, no comparison to **references**, **source** considered

Quality = **Can we publish it as is?**

Quality = **Can a reader get the gist?**

Quality = **Is it worth post-editing it?**

# QE - Overview

- **Quality estimation** (QE): metrics that provide an **estimate** on the **quality** of translations *on the fly*
- Quality defined by the **data**: **purpose** is clear, no comparison to **references**, **source** considered

Quality = **Can we publish it as is?**

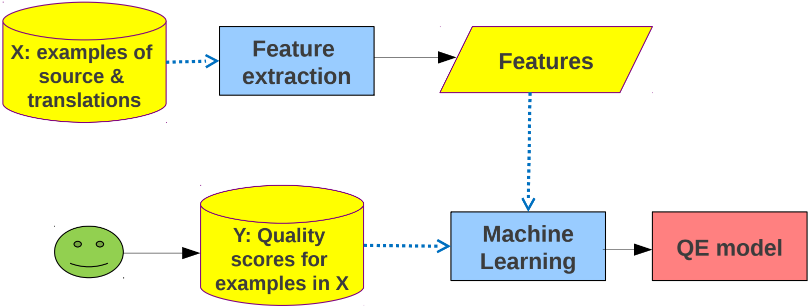
Quality = **Can a reader get the gist?**

Quality = **Is it worth post-editing it?**

Quality = **How much effort to fix it?**

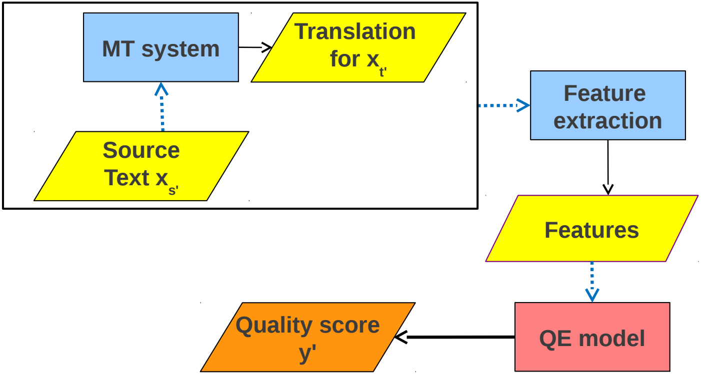
# QE - Framework

Building a model:



# QE - Framework

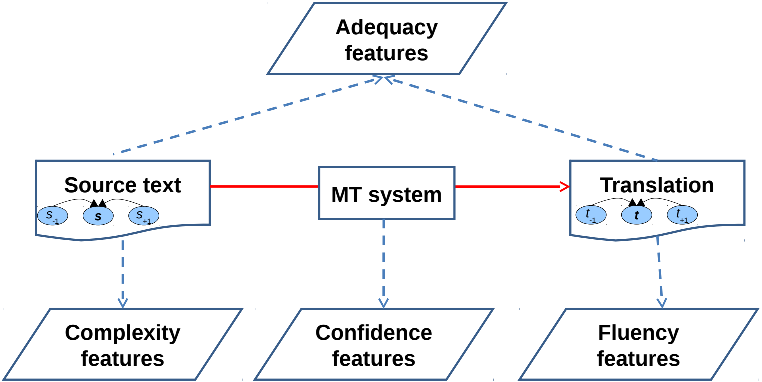
Applying the model:



# Data and levels of granularity

- **Sentence level**: 1-5 subjective scores, PE time, PE edits
- **Word level**: good/bad, good/delete/replace, MQM
- **Phrase level**: good/bad
- **Document level**: PE effort

# Features and algorithms




**Algorithms** can be used off-the-shelf

# QE - baseline setting

**Features**:

- number of tokens in the source and target sentences
- average source token length
- average number of occurrences of words in the target
- number of punctuation marks in source and target sentences
- LM probability of source and target sentences
- average number of translations per source word
- % of seen source n-grams

# QE - baseline setting

**Features**:

- number of tokens in the source and target sentences
- average source token length
- average number of occurrences of words in the target
- number of punctuation marks in source and target sentences
- LM probability of source and target sentences
- average number of translations per source word
- % of seen source n-grams

**SVM regression** with RBF kernel

# QE - baseline setting

**Features**:

- number of tokens in the source and target sentences
- average source token length
- average number of occurrences of words in the target
- number of punctuation marks in source and target sentences
- LM probability of source and target sentences
- average number of translations per source word
- % of seen source n-grams

**SVM regression** with RBF kernel

QuEst: `http://www.quest.dcs.shef.ac.uk/`

# QE - SoA sentence-level

Predicting HTER (WMT16)

| System ID | **Pearson** ↑ | Spearman ↑ |
|---:|:---:|:---:|
| **English-German** | | |
| • YSDA/SNTX+BLEU+SVM | 0.525 | – |
| POSTECH/SENT-RNN-QV2 | 0.460 | 0.483 |
| SHEF-LIUM/SVM-NN-emb-QuEst | 0.451 | 0.474 |
| POSTECH/SENT-RNN-QV3 | 0.447 | 0.466 |
| SHEF-LIUM/SVM-NN-both-emb | 0.430 | 0.452 |
| UGENT-LT3/SCATE-SVM2 | 0.412 | 0.418 |
| UFAL/MULTIVEC | 0.377 | 0.410 |
| RTM/RTM-FS-SVR | 0.376 | 0.400 |
| UU/UU-SVM | 0.370 | 0.405 |
| UGENT-LT3/SCATE-SVM1 | 0.363 | 0.375 |
| RTM/RTM-SVR | 0.358 | 0.384 |
| Baseline SVM | 0.351 | 0.390 |
| SHEF/SimpleNets-SRC | 0.182 | – |
| SHEF/SimpleNets-TGT | 0.182 | – |

# Outline

# SMT vs NMT

Pearson correlation with DA scores for popular metrics on 200 sentences from WMT16's **uedin** SMT and NMT systems:

| | uedin-pbmt | uedin-nmt |
|---|---|---|
| BLEU | 0.4433 | 0.5126 |
| Meteor | 0.5123 | 0.5781 |
| TER | -0.4042 | -0.5592 |
| chrF2 | 0.4959 | 0.5826 |
| BEER | 0.5034 | 0.6140 |
| UPF-Cobalt | 0.5365 | 0.5511 |
| CobaltF-comp | 0.5306 | 0.6064 |
| DPMFcomb | 0.5757 | 0.6507 |

(Work with **Marina Fomicheva**)

# Are metrics better for NMT because systems are better?

Correlation with DA scores on 840 **low-quality** (Q1-human) & 840 **high-quality** (Q4-human) sentences (all systems)

| | Q1 - low quality | Q4 - high quality |
|---|---|---|
| BLEU | 0.0338 | 0.4561 |
| Meteor | 0.1985 | 0.5143 |
| TER | -0.0870 | -0.3710 |
| UPF-Cobalt | 0.1499 | 0.4035 |
| CobaltF-comp | 0.0918 | 0.4691 |
| DPMFcomb | 0.2035 | 0.4426 |
| BEER | 0.2277 | 0.3840 |
| chrF2 | 0.2177 | 0.3749 |

(Work with **Marina Fomicheva**)

# Or was it a feature of the **uedin** systems?

Correlation of various MT systems on 400 sentences per group:

| | PBMT | PBMT + NMT | Syntax |
|---|---|---|---|
| BLEU | 0.5662 | 0.4676 | **0.4521** |
| Meteor | 0.6178 | 0.5462 | 0.5560 |
| TER | -0.5277 | **-0.4177** | **-0.3929** |
| chrF2 | 0.5549 | 0.5093 | 0.4602 |
| BEER | 0.5445 | 0.4913 | 0.4598 |
| UPF-Cobalt | 0.6510 | **0.5400** | **0.5221** |
| CobaltF-comp | 0.6328 | 0.5788 | 0.5693 |
| MetricsF | 0.6575 | 0.5840 | 0.5803 |
| DPMFcomb | 0.6700 | 0.5876 | **0.5815** |

These NMT systems only use neural models for rescoring. Also, average DA scores not higher for the PMT+NMT group

(Work with **Marina Fomicheva**)

# Conclusions

- (Machine) Translation evaluation is still an open problem

# Conclusions

- (Machine) Translation evaluation is still an open problem
- **Quality estimation** and other trained metrics can learn different "versions" of quality

# Conclusions

- (Machine) Translation evaluation is still an open problem
- **Quality estimation** and other trained metrics can learn different "versions" of quality
- Which metrics are used in practice?
  - BLEU + your favourite other
  - And same metric for tuning

# Conclusions

- (Machine) Translation evaluation is still an open problem
- **Quality estimation** and other trained metrics can learn different "versions" of quality
- Which metrics are used in practice?
  - BLEU + your favourite other
  - And same metric for tuning
- And for **official** comparisons?
  - WMT: manual ranking and direct assessment
  - IWSLT: manual post-editing

# Conclusions

- (Machine) Translation evaluation is still an open problem
- **Quality estimation** and other trained metrics can learn different "versions" of quality
- Which metrics are used in practice?
  - BLEU + your favourite other
  - And same metric for tuning
- And for **official** comparisons?
  - WMT: manual ranking and direct assessment
  - IWSLT: manual post-editing
- Are our metrics good at assessing NMT systems?
- Are these metrics good to optimise NMT systems?

# Translation Quality Assessment: Evaluation and Estimation

Lucia Specia

University of Sheffield
l.specia@sheffield.ac.uk

MTM - Prague, 12 September 2016

The University Of Sheffield.

# Conclusions

| MT system | Type | Average score | Segments |
|---|---|---|---|
| AFRL-MITLL-Phrase | PBMT + NMT | 0.0118 | 56 |
| AFRL-MITLL-contrast | PBMT + NMT | -0.1423 | 72 |
| AMU-UEDIN | PBMT + NMT | 0.1981 | 61 |
| KIT | PBMT + NMT | 0.1431 | 73 |
| LIMSI | PBMT | -0.1482 | 84 |
| NRC | PBMT | 0.0877 | 58 |
| PJATK | PBMT | 0.0137 | 132 |
| PROMT-Rule-based | RBMT | 0.0107 | 56 |
| PROMT-SMT | PBMT | -0.1163 | 154 |
| UH-factored | PBMT | -0.1138 | 70 |
| UH-opus | PBMT | -0.0059 | 72 |
| cu-mergedtrees | Syntax PBMT | -0.4976 | 106 |
| dvorkanton | PBMT + NMT | -0.1548 | 72 |
| jhu-pbmt | PBMT | -0.0985 | 446 |
| jhu-syntax | Syntax PBMT | -0.2491 | 125 |
| online-B | PBMT | 0.0793 | 430 |
| online-F | PBMT | -0.2447 | 125 |
| online-G | PBMT | 0.0186 | 272 |
| tbtk-syscomb | PBMT | -0.0594 | 85 |
| uedin-nmt | NMT | 0.0774 | 342 |
| uedin-pbmt | PBMT | 0.0391 | 231 |
| uedin-syntax | Syntax PBMT | 0.0121 | 238 |