

# Translation Quality Assessment: Evaluation and Estimation

Lucia Specia

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

9 September 2013



# Overview

“MT Evaluation is better understood than MT” (Carbonell and Wilks, 1991)

# Outline

- 1 Translation Quality
- 2 QTLaunchPad Project
- 3 Quality Estimation
- 4 State of the art in QE
- 5 Open issues
- 6 Conclusions

# Outline

- 1 Translation Quality
- 2 QTLaunchPad Project
- 3 Quality Estimation
- 4 State of the art in QE
- 5 Open issues
- 6 Conclusions

# Overview

- What does **quality** mean?
  - Fluent?
  - Adequate?
  - Easy to post-edit?

# Overview

- What does **quality** mean?
  - Fluent?
  - Adequate?
  - Easy to post-edit?
- Quality for **whom**?
  - End-user
  - MT-system (tuning)
  - Post-editor
  - Other applications (e.g. CLIR)

# Overview

- What does **quality** mean?
  - Fluent?
  - Adequate?
  - Easy to post-edit?
- Quality for **whom**?
  - End-user
  - MT-system (tuning)
  - Post-editor
  - Other applications (e.g. CLIR)
- Quality for **what**?
  - Internal communications
  - Dissemination (publishing)
  - Gisting (Google Translate)
  - Draft translations (light vs heavy post-editing)
  - MT system improvement (diagnosis)

# Overview

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

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



# Overview

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

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

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

# Overview

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

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

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

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

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

# Overview

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

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

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

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

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

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

# Overview

How do we **measure** quality?

- **Human metrics**: error counts (which?), ranking, acceptability, 1-N fluency/adequacy
- **Automatic metrics** based on human **references**: (BLEU, METEOR, TER, etc.
- **Semi-automatic metrics** based on **post-editions**: HTER, PE time, eye-tracking, etc.

# Overview

How do we **measure** quality?

- **Human metrics**: error counts (which?), ranking, acceptability, 1-N fluency/adequacy
- **Automatic metrics** based on human **references**: (BLEU, METEOR, TER, etc.
- **Semi-automatic metrics** based on **post-editions**: HTER, PE time, eye-tracking, etc.
- **Automatic metrics** without references: **quality estimation**

# Outline

- 1 Translation Quality
- 2 QTLaunchPad Project**
- 3 Quality Estimation
- 4 State of the art in QE
- 5 Open issues
- 6 Conclusions

# QTLaunchPad project

<http://www.qt21.eu/launchpad/>

- Multidimensional Quality Metrics (**MQM**) based on a **specification**
- Machine and human translation quality
- Manual and (semi-)automatic assessment
- Takes quality of **source text** into account
- MT system improvement, gisting, dissemination, etc.



# Multidimensional Quality Metrics (MQM)

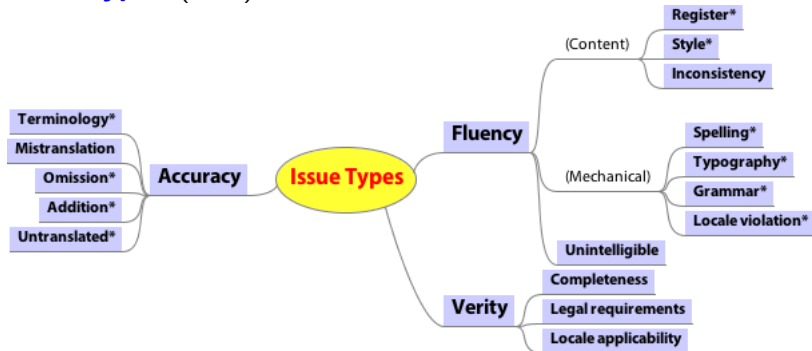
Issues selected based on a given **specification** (dimensions):

- Language/locale
- Subject field/domain
- Terminology
- Text Type
- Audience
- Purpose
- Register
- Style
- Content correspondence
- Output modality, ...



# Multidimensional Quality Metrics (MQM)

**Issue types** (core):



# Multidimensional Quality Metrics (MQM)

**Issue types:** <http://www.qt21.eu/launchpad/content/high-level-structure-0>

**Combining issue types:**

$$TQ = 100 - AccP - (FluP_T - FluP_S) - (VerP_T - VerP_S)$$

# Outline

- 1 Translation Quality
- 2 QTLaunchPad Project
- 3 Quality Estimation**
- 4 State of the art in QE
- 5 Open issues
- 6 Conclusions

# Overview

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

# Overview

- **Quality estimation** (QE): metrics that provide an **estimate** on the **quality** of unseen translations
- Quality defined by the **data**

# Overview

- **Quality estimation** (QE): metrics that provide an **estimate** on the **quality** of unseen translations
- Quality defined by the **data**

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

# Overview

- **Quality estimation** (QE): metrics that provide an **estimate** on the **quality** of unseen translations
- Quality defined by the **data**

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

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

# Overview

- **Quality estimation** (QE): metrics that provide an **estimate** on the **quality** of unseen translations
- Quality defined by the **data**

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

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

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



# Overview

- **Quality estimation** (QE): metrics that provide an **estimate** on the **quality** of unseen translations
- Quality defined by the **data**

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?**

# Overview

- **Quality estimation** (QE): metrics that provide an **estimate** on the **quality** of unseen translations
- Quality defined by the **data**

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

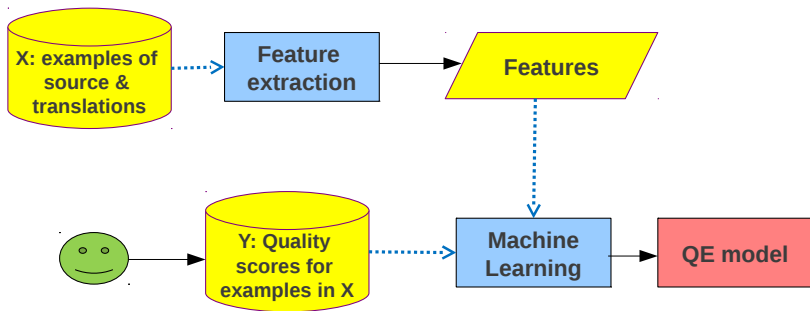
Quality = **Can a reader get the gist?**

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

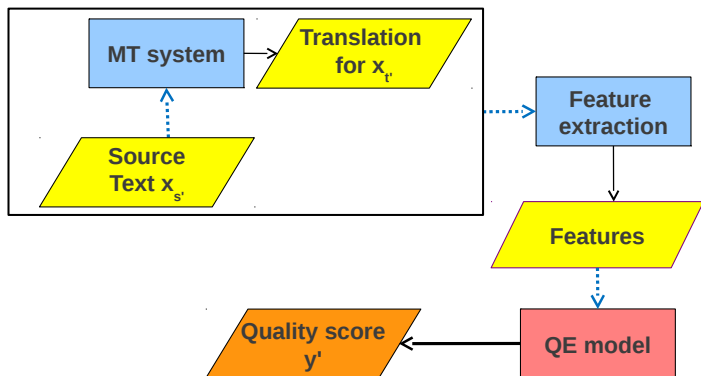
Quality = **How much effort to fix it?**

Quality = **What's this translation's MQM score?**

# Framework



# Framework



# Framework

Main components to build a QE system:

- ① Definition of quality: **what to predict**
- ② (Human) labelled **data** (for quality)
- ③ **Features**
- ④ Machine learning **algorithm**

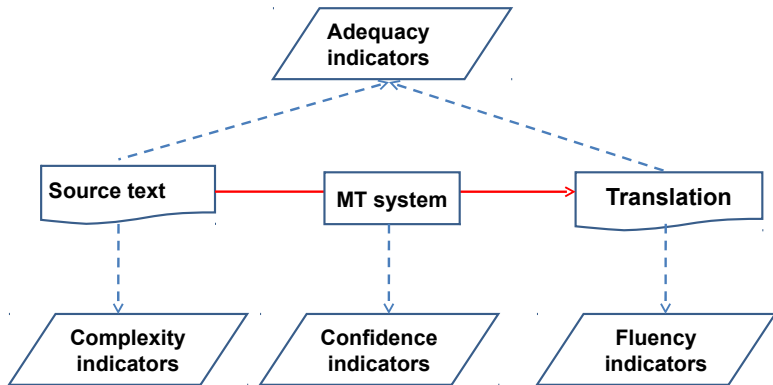
# Definition of quality

- Predict 1-N **absolute** scores for adequacy/fluency
- Predict 1-N **absolute** scores for post-editing effort
- Predict average post-editing **time** per word
- Predict **relative** rankings
- Predict **relative** rankings for same source
- Predict **percentage of edits** needed for sentence
- Predict word-level **edits** and its types
- Predict **BLEU**, etc. scores for document

# Datasets

- **SHEF** (several): <http://staffwww.dcs.shef.ac.uk/people/L.Specia/resources.html>
- **LIG** (10K, fr-en): <http://www-clips.imag.fr/geod/User/marion.potet/index.php?page=download>
- **LMSI** (14K, fr-en, en-fr, 2 post-editors):  
<http://web.limsi.fr/Individu/wisniews/recherche/index.html>

# Features





# QuEst

**Goal:** framework to explore features for QE

- **Feature extractors** for 150+ features of all types: Java
- **Machine learning:** Gaussian Processes & scikit-learn toolkit (Python), with wrappers for a number of algorithms, grid search, feature selection



Open source:

<http://www.quest.dcs.shef.ac.uk/>

# Outline

- 1 Translation Quality
- 2 QTLaunchPad Project
- 3 Quality Estimation
- 4 State of the art in QE**
- 5 Open issues
- 6 Conclusions

# Shared Task

- **WMT12-13** – with Radu Soricut & Christian Buck

# Shared Task

- **WMT12-13** – with Radu Soricut & Christian Buck
- **Sentence-** and **word-level** estimation of **PE effort**

# Shared Task

- **WMT12-13** – with Radu Soricut & Christian Buck
- **Sentence-** and **word-level** estimation of **PE effort**
- Datasets and **language pairs**:

Quality	Year	Languages
1-5 subjective scores	WMT12	en-es
Ranking all sentences best-worst	WMT12/13	en-es
HTER scores	WMT13	en-es
Post-editing time	WMT13	en-es
Word-level edits: change/keep	WMT13	en-es
Word-level edits: keep/delete/replace	WMT13	en-es
Ranking 5 MTs per source	WMT13	en-es; de-en

# Shared Task

- **WMT12-13** – with Radu Soricut & Christian Buck
- **Sentence-** and **word-level** estimation of **PE effort**
- Datasets and **language pairs**:

Quality	Year	Languages
1-5 subjective scores	WMT12	en-es
Ranking all sentences best-worst	WMT12/13	en-es
HTER scores	WMT13	en-es
Post-editing time	WMT13	en-es
Word-level edits: change/keep	WMT13	en-es
Word-level edits: keep/delete/replace	WMT13	en-es
Ranking 5 MTs per source	WMT13	en-es; de-en

- Evaluation metric:

$$\text{MAE} = \frac{\sum_{i=1}^N |H(s_i) - V(s_i)|}{N}$$

# Baseline system

## 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 source 1-grams, 2-grams and 3-grams in frequency quartiles 1 and 4
- % of seen source unigrams

# Baseline system

## 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 source 1-grams, 2-grams and 3-grams in frequency quartiles 1 and 4
- % of seen source unigrams

**SVM regression** with RBF kernel with the parameters  $\gamma$ ,  $\epsilon$  and  $C$  optimised using a grid-search and 5-fold cross validation on the training set



# Results - scoring sub-task (WMT12)

System ID	MAE	RMSE
• SDLLW_M5PbestDeltaAvg	0.61	0.75
UU_best	0.64	0.79
SDLLW_SVM	0.64	0.78
UU_bltk	0.64	0.79
Loria_SVMlinear	0.68	0.82
UEdin	0.68	0.82
TCD_M5P-resources-only*	0.68	0.82
<b>Baseline bb17 SVR</b>	0.69	0.82
Loria_SVMrbf	0.69	0.83
SJTU	0.69	0.83
WLV-SHEF_FS	0.69	0.85
PRHLT-UPV	0.70	0.85
WLV-SHEF_BL	0.72	0.86
DCU-SYMC_unconstrained	0.75	0.97
DFKI_grcfs-mars	0.82	0.98
DFKI_cfs-plsreg	0.82	0.99
UPC_1	0.84	1.01
DCU-SYMC_constrained	0.86	1.12
UPC_2	0.87	1.04
TCD_M5P-all	2.09	2.32

# Results - scoring sub-task (WMT13)

System ID	MAE	RMSE
• SHEF FS	12.42	15.74
SHEF FS-AL	13.02	17.03
CNGL SVRPLS	13.26	16.82
LIMSI	13.32	17.22
DCU-SYMC combine	13.45	16.64
DCU-SYMC alltypes	13.51	17.14
CMU noB	13.84	17.46
CNGL SVR	13.85	17.28
FBK-UEdin extra	14.38	17.68
FBK-UEdin rand-svr	14.50	17.73
LORIA inctrain	14.79	18.34
<b>Baseline bb17 SVR</b>	14.81	18.22
TCD-CNGL open	14.81	19.00
LORIA inctraincont	14.83	18.17
TCD-CNGL restricted	15.20	19.59
CMU full	15.25	18.97
UMAC	16.97	21.94

# Outline

- 1 Translation Quality
- 2 QTLaunchPad Project
- 3 Quality Estimation
- 4 State of the art in QE
- 5 Open issues**
- 6 Conclusions

# Agreement between annotators

- **Absolute value judgements:** difficult to achieve consistency even in highly controlled settings
  - 30% of initial dataset discarded
  - Remaining annotations had to be scaled

# Agreement between annotators

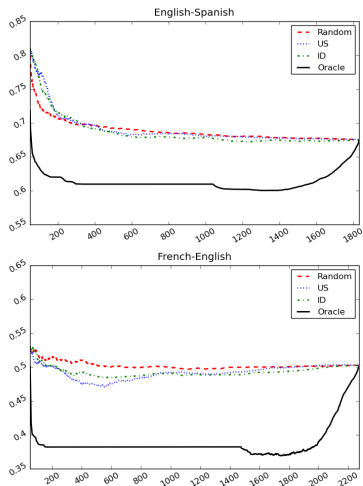
- **Absolute value judgements:** difficult to achieve consistency even in highly controlled settings
  - 30% of initial dataset discarded
  - Remaining annotations had to be scaled
- **More objective absolute scores**
  - Post-editing time, HTER, edits
  - Also subject to huge variance (WPTP 2013, Wisniewski et. al, MT-Summit 2013)
  - Multi-task learning to address this variance (Cohn and Specia, ACL 2013)

# Agreement between annotators

- **Absolute value judgements:** difficult to achieve consistency even in highly controlled settings
  - 30% of initial dataset discarded
  - Remaining annotations had to be scaled
- **More objective absolute scores**
  - Post-editing time, HTER, edits
  - Also subject to huge variance (WPTP 2013, Wisniewski et. al, MT-Summit 2013)
  - Multi-task learning to address this variance (Cohn and Specia, ACL 2013)
- **Relative scores**
  - Different task altogether
  - WMT13: better results than reference-based metrics

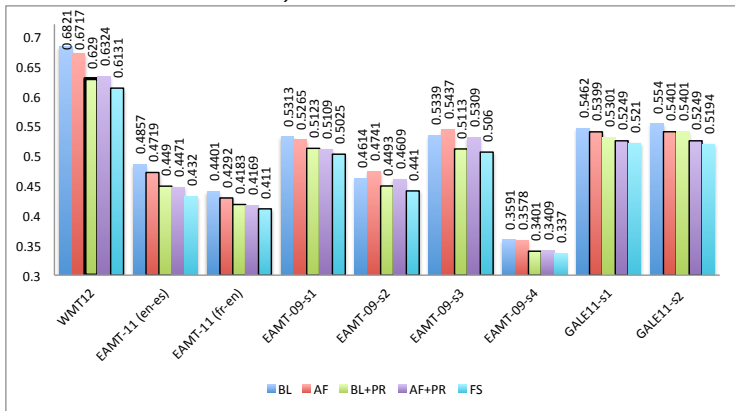
# Annotation costs

**Active learning** to select subset of instances to be annotated  
(Beck et al., ACL 2013)



# Curse of dimensionality

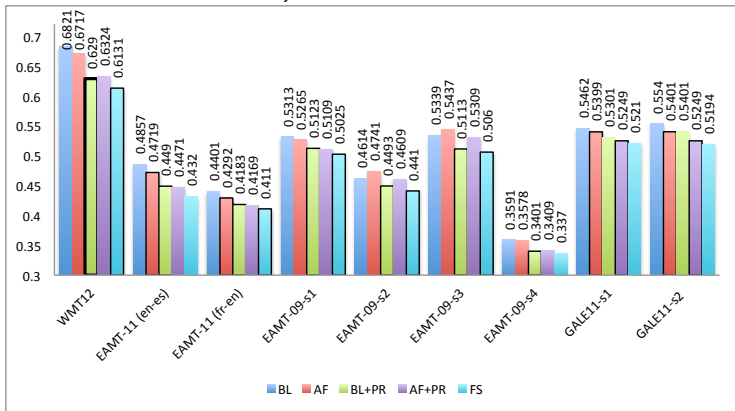
**Feature selection** to identify relevant info for dataset (Shah et al., MT Summit 2013)





# Curse of dimensionality

**Feature selection** to identify relevant info for dataset (Shah et al., MT Summit 2013)



Common feature set identified, but **nuanced subsets** for specific datasets

# How to use estimated PE effort scores?

Do users prefer **detailed estimates** (sub-sentence level) or an **overall estimate** for the complete sentence or **not seeing** bad sentences at all?

- Too much information vs hard-to-interpret scores

# How to use estimated PE effort scores?

Do users prefer **detailed estimates** (sub-sentence level) or an **overall estimate** for the complete sentence or **not seeing** bad sentences at all?

- Too much information vs hard-to-interpret scores
- IBM's *Goodness* metric

Source أنت مختلف تماماً عن زيد وعسرو فلا تحشر نفسك في سرداب التقليد والمحاكاة والذوبان

MT output you totally different from zaid amr , and not to deprive yourself in a basement of imitation and assimilation .

We predict and visualize you **totally** different from **zaid amr** , and **not to deprive yourself in a basement of imitation and** assimilation .

# How to use estimated PE effort scores?

Do users prefer **detailed estimates** (sub-sentence level) or an **overall estimate** for the complete sentence or **not seeing** bad sentences at all?

- Too much information vs hard-to-interpret scores
- IBM's *Goodness* metric

Source أنت مختلف تماماً عن زيد وعسرو فلا تحشر نفسك في سرداب التقليد والمحاكاة والذوبان

MT output you totally different from zaid amr , and not to deprive yourself in a basement of imitation and assimilation .

We predict and visualize you **totally** different from **zaid amr** , and **not to deprive yourself in a basement of imitation and** assimilation .

- MATECAT project investigating it

# Feature engineering

Two families of features missing in current work:

- Can we benefit from contextual, **document-wide information**?
- Can we predict **human translation quality**?

# Feature engineering

Two families of features missing in current work:

- Can we benefit from contextual, **document-wide information**?
- Can we predict **human translation quality**?

Don't panic, you can help!

Two (sub-) QuEst projects at MTM-2013 :-)

# Outline

- 1 Translation Quality
- 2 QTLaunchPad Project
- 3 Quality Estimation
- 4 State of the art in QE
- 5 Open issues
- 6 Conclusions**

# Conclusions

- (Machine) Translation evaluation is still an open problem



# Conclusions

- (Machine) Translation evaluation is still an open problem
- Different purposes/users, different needs, different notions of **quality**

# Conclusions

- (Machine) Translation evaluation is still an open problem
- Different purposes/users, different needs, different notions of **quality**
- **Quality estimation**: learning of these different notions

# Conclusions

- (Machine) Translation evaluation is still an open problem
- Different purposes/users, different needs, different notions of **quality**
- **Quality estimation**: learning of these different notions
- Estimates have been used in **real applications**
  - Ranking translations: filter out bad quality translations
  - Selecting translations from multiple MT systems

# Conclusions

- (Machine) Translation evaluation is still an open problem
- Different purposes/users, different needs, different notions of **quality**
- **Quality estimation**: learning of these different notions
- Estimates have been used in **real applications**
  - Ranking translations: filter out bad quality translations
  - Selecting translations from multiple MT systems
- **Commercial** interest

# Conclusions

- (Machine) Translation evaluation is still an open problem
- Different purposes/users, different needs, different notions of **quality**
- **Quality estimation**: learning of these different notions
- Estimates have been used in **real applications**
  - Ranking translations: filter out bad quality translations
  - Selecting translations from multiple MT systems
- **Commercial** interest
  - SDL LW: TrustScore
  - Multilizer: MT-Qualifier

# Conclusions

- (Machine) Translation evaluation is still an open problem
- Different purposes/users, different needs, different notions of **quality**
- **Quality estimation**: learning of these different notions
- Estimates have been used in **real applications**
  - Ranking translations: filter out bad quality translations
  - Selecting translations from multiple MT systems
- **Commercial** interest
  - SDL LW: TrustScore
  - Multilizer: MT-Qualifier
- Interesting **open issues**: join the QuEst projects!