

Translation Quality Evaluation and Estimation

Lucia Specia

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

ASLIB: Translating and the Computer Conference
29 November 2013



The
University
Of
Sheffield.

Outline

- 1 Translation quality
- 2 Reference-based metrics
- 3 Task-based metrics
- 4 Prediction-based metrics
- 5 Conclusions

Outline

- 1 Translation quality
- 2 Reference-based metrics
- 3 Task-based metrics
- 4 Prediction-based metrics
- 5 Conclusions

Machine Translation (MT)

- Over 60 years of MT research, mature technology, successful commercial applications

Machine Translation (MT)

- Over 60 years of MT research, mature technology, successful commercial applications
- Still: **repetitive** and **grotesque** and errors...

Machine Translation (MT)

- Over 60 years of MT research, mature technology, successful commercial applications
- Still: **repetitive** and **grotesque** and errors...
- Quality assessment is core

Machine Translation (MT)

- Over 60 years of MT research, mature technology, successful commercial applications
- Still: **repetitive** and **grotesque** and errors...
- Quality assessment is core

“Machine Translation evaluation is better understood than
Machine Translation”
(Carbonell and Wilks, 1991) [CW91]

Why is evaluation important?

Translation output evaluation is needed to:

- Compare MT systems
- Measure progress of MT systems over time
- Quality assurance (HT or MT)
- Tune statistical MT systems
- **Diagnose MT systems**
- Decide on **fitness-for-purpose**

Why is evaluation important?

Translation output evaluation is needed to:

- Compare MT systems
- Measure progress of MT systems over time
- Quality assurance (HT or MT)
- Tune statistical MT systems
- **Diagnose MT systems**
- Decide on **fitness-for-purpose**
- Select among alternative MT/TM/HT (e.g. crowdsourcing translations)

Why is evaluation hard?

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

Why is evaluation hard?

- What does **quality** mean?
 - Fluent?
 - Adequate?
 - Easy to post-edit?
- Quality for **whom/what**?
 - End-user: gisting (Google Translate), internal communications, or publication (dissemination)
 - MT-system: tuning or diagnosis
 - Post-editor: fix draft translations
 - Other applications, e.g. CLIR

Overview

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

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

Overview

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

MT: 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!

MT: 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**.

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

Overview

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

MT: 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**.

MT: **Six-hour 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?

- **Manual metrics:**
 - Ranking, acceptability, 1-N judgements on fluency/adequacy, **error analysis**
 - **Task-based human metrics:** productivity tests, user-satisfaction, reading comprehension

Overview

How do we **measure** quality?

- **Manual metrics:**
 - Ranking, acceptability, 1-N judgements on fluency/adequacy, **error analysis**
 - **Task-based human metrics:** productivity tests, user-satisfaction, reading comprehension
- **Automatic metrics:**
 - Based on human **references:** BLEU, METEOR, TER, TerrorCAT, ...

Overview

How do we **measure** quality?

- **Manual metrics:**
 - Ranking, acceptability, 1-N judgements on fluency/adequacy, **error analysis**
 - **Task-based human metrics:** productivity tests, user-satisfaction, reading comprehension
- **Automatic metrics:**
 - Based on human **references:** BLEU, METEOR, TER, TerrorCAT, ...
 - Reference-less: **quality estimation**

Overview

How do we **measure** quality?

- **Manual metrics:**
 - Ranking, acceptability, 1-N judgements on fluency/adequacy, **error analysis**
 - **Task-based human metrics:** productivity tests, user-satisfaction, reading comprehension
- **Automatic metrics:**
 - Based on human **references:** BLEU, METEOR, TER, TerrorCAT, ...
 - Reference-less: **quality estimation**

Different levels of **granularity:** document-, sentence-, phrase- or word-level

Outline

- 1 Translation quality
- 2 Reference-based metrics**
- 3 Task-based metrics
- 4 Prediction-based metrics
- 5 Conclusions

Reference-based automatic metrics

- Compare output of an **MT system** to one or more **reference** (human) translations: how close is the MT output to the reference translation?
- Numerous metrics: WER/TER, BLEU/NIST, AMBER, ROSE, etc.

String matching: BLEU

BLEU: BiLingual Evaluation Understudy

- **Most widely used metric**, for MT system evaluation/comparison and SMT tuning
- Geometric mean of n -gram precisions (n from 1 to 4) in MT output

$$p_n = \frac{\sum_{h \in H} \sum_{g \in n\text{grams}(h)} \#clip(g)}{\sum_{h \in H} \sum_{g' \in n\text{grams}(h)} \#(g')} \rightarrow \sum_n \log p_n$$

String matching: BLEU

BLEU: BiLingual Evaluation Understudy

- **Most widely used metric**, for MT system evaluation/comparison and SMT tuning
- Geometric mean of n -gram precisions (n from 1 to 4) in MT output

$$p_n = \frac{\sum_{h \in H} \sum_{g \in \text{ngrams}(h)} \#clip(g)}{\sum_{h \in H} \sum_{g' \in \text{ngrams}(h)} \#(g')} \rightarrow \sum_n \log p_n$$

- **Brevity penalty** for MT sentences shorter than reference

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

String matching: BLEU

BLEU: BiLingual Evaluation Understudy

- **Most widely used metric**, for MT system evaluation/comparison and SMT tuning
- Geometric mean of n -gram precisions (n from 1 to 4) in MT output

$$p_n = \frac{\sum_{h \in H} \sum_{g \in \text{ngrams}(h)} \# \text{clip}(g)}{\sum_{h \in H} \sum_{g' \in \text{ngrams}(h)} \#(g')} \rightarrow \sum_n \log p_n$$

- **Brevity penalty** for MT sentences shorter than reference

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

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

Edit distance: TER

TER: Translation Error Rate

- % of **insertions**, **deletions**, **replacements**, and **shifts** needed to transform an MT into the reference sentence

$$TER = \frac{I + D + R + S}{N}$$

Edit distance: TER

TER: Translation Error Rate

- % of **insertions**, **deletions**, **replacements**, and **shifts** needed to transform an MT into the reference sentence

$$TER = \frac{I + D + R + S}{N}$$

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: TER

TER: Translation Error Rate

- % of **insertions**, **deletions**, **replacements**, and **shifts** needed to transform an MT into the reference sentence

$$TER = \frac{I + D + R + S}{N}$$

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

$$1 S, 2 R, 1 D \rightarrow 4 \text{ Edits: } TER = \frac{4}{13} = 0.31$$

Edit distance: TER

TER: Translation Error Rate

- % of **insertions**, **deletions**, **replacements**, and **shifts** needed to transform an MT into the reference sentence

$$TER = \frac{I + D + R + S}{N}$$

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

$$1 \text{ S, } 2 \text{ R, } 1 \text{ D} \rightarrow 4 \text{ Edits: } TER = \frac{4}{13} = 0.31$$

Human-targeted TER (HTER)

TER between MT and its post-edited version

Error analysis

Aimed at **diagnosis** of MT systems

- Automatic metrics for **fine-grained error analysis** [PN11, ZFBB11]
- Few error categories: inflectional errors, errors due to wrong word order, missing words, extra words, and incorrect lexical choices
- Mostly based on **word alignment** of MT output to reference translation, followed by **linguistic processing** and **classification algorithms** to categorise mismatches

Error analysis

Aimed at **diagnosis** of MT systems

- Automatic metrics for **fine-grained error analysis** [PN11, ZFBB11]
- Few error categories: inflectional errors, errors due to wrong word order, missing words, extra words, and incorrect lexical choices
- Mostly based on **word alignment** of MT output to reference translation, followed by **linguistic processing** and **classification algorithms** to categorise mismatches

Same can be done using **post-edited version** [WSSY13]: more precise.

Reference-based automatic metrics

Advantages:

- Fast and cheap, minimal human labour
 - **Reuse** test set, **system development**
- Metrics can look at variable ways of saying the same thing (stems, synonyms), e.g. METEOR
- Metrics can penalise mismatches differently, e.g. TESLA

Reference-based automatic metrics

Advantages:

- Fast and cheap, minimal human labour
 - **Reuse** test set, **system development**
- Metrics can look at variable ways of saying the same thing (stems, synonyms), e.g. METEOR
- Metrics can penalise mismatches differently, e.g. TESLA

Disadvantages:

- Too coarse: do not provide information on **what went wrong**
- Reference translations are only **a subset of the possible good translations**
- Reference translations are **not available for MT systems in use**

Reference-based automatic metrics

Advantages:

- Fast and cheap, minimal human labour
 - **Reuse** test set, **system development**
- Metrics can look at variable ways of saying the same thing (stems, synonyms), e.g. METEOR
- Metrics can penalise mismatches differently, e.g. TESLA

Disadvantages:

- Too coarse: do not provide information on **what went wrong**
- Reference translations are only **a subset of the possible good translations**
- Reference translations are **not available for MT systems in use**
- Metrics are not easily interpretable. BLEU = 0.36???

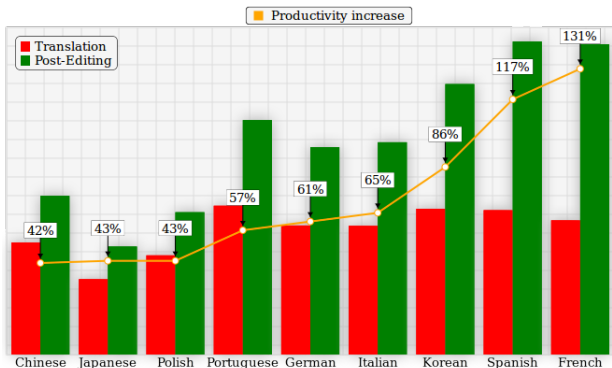
Outline

- 1 Translation quality
- 2 Reference-based metrics
- 3 Task-based metrics**
- 4 Prediction-based metrics
- 5 Conclusions

Productivity analysis

E.g. **Autodesk** - productivity test through **post-editing**
[Aut11]

- 2-day translation and post-editing , 37 participants
- In-house Moses (Autodesk data: software)
- **Time** spent on each segment



User satisfaction

Solving a problem: E.g.: **Intel** measuring user satisfaction with un-edited MT

- Translation is good if customer can **solve problem**

User satisfaction

Solving a problem: E.g.: **Intel** measuring user satisfaction with un-edited MT

- Translation is good if customer can **solve problem**
- MT for Customer Support websites [Int10]
 - Overall customer satisfaction: **75%** for English→Chinese
 - **95%** reduction in cost
 - Project cycle from **10 days** to **1 day**
 - From **300** to **60,000** words translated/hour

User satisfaction

Solving a problem: E.g.: **Intel** measuring user satisfaction with un-edited MT

- Translation is good if customer can **solve problem**
- MT for Customer Support websites [Int10]
 - Overall customer satisfaction: **75%** for English→Chinese
 - **95%** reduction in cost
 - Project cycle from **10 days** to **1 day**
 - From **300** to **60,000** words translated/hour
 - Customers in China using MT texts were more satisfied with support than natives using original texts (**68%**)!

Outline

- 1 Translation quality
- 2 Reference-based metrics
- 3 Task-based metrics
- 4 Prediction-based metrics**
- 5 Conclusions

Overview

- **Quality estimation** (QE): metrics that provide an **estimate** on the **quality** of unseen translations, a.k.a. **confidence estimation** (ASR)

Overview

- **Quality estimation** (QE): metrics that provide an **estimate** on the **quality** of unseen translations, a.k.a. **confidence estimation** (ASR)
- Measuring vs estimating/predicting quality
- Quality defined by labels in training **data**, according to the **application**

Overview

- **Quality estimation** (QE): metrics that provide an **estimate** on the **quality** of unseen translations, a.k.a. **confidence estimation** (ASR)
- Measuring vs estimating/predicting quality
- Quality defined by labels in training **data**, according to the **application**
- Long-term goal: estimate fine-grained metrics like MQM, DQF

Motivations

Assessing translation quality is time consuming:

MT: Events of a magnitude unprecedented Mongols claiming their rights have occurred last week in this autonomous region, according to the Information Centre on Human Rights in South Mongolia, an organization based in the States U.S., where universities and public spaces open air were banned from several cities, fearing the power to Beijing more than any protest rallies in the spirit of movements which have stirred recent months the world Arabic.

SRC: Des manifestations d'une ampleur sans précédent de Mongols réclamant le respect de leurs droits se sont produites la semaine dernière dans cette région autonome, selon le Centre d'information sur les droits de l'homme en Mongolie du Sud, une organisation installée aux Etats-Unis, où des universités et des espaces publics en plein air étaient interdits d'accès dans plusieurs villes, le pouvoir à Pékin redoutant plus que tout des rassemblements de protestation dans l'esprit des mouvements qui ont agité ces derniers mois des pays du monde arabe.

Motivations

Assessing translation quality is not possible if user cannot read source language:

Target:

Continued high floods **subside**. Guang'an old city has been soaked 2 days 2 nights

Source:

四川广安洪水持续高位不退 老城区已被泡2天2夜

By Google Translate

Motivations

Assessing translation quality is not possible if user cannot read source language:

Target:

Continued high floods **subside**. Guang'an old city has been soaked 2 days 2 nights

Source:

四川广安洪水持续高位不退 老城区已被泡2天2夜

By Google Translate

Reference:

The continuing floods in Guang'an - Sichuan have **not subsided**. The old city has been flooded for 2 days and 2 nights.

Motivations

Assessing translation quality is not possible if user cannot read source language:

Target:

site security should be included in **sex education** curriculum for students

Source:

场地安全性教育应纳入学生的课程

By Google Translate

Motivations

Assessing translation quality is not possible if user cannot read source language:

Target:

site security should be included in **sex education** curriculum for students

Source:

场地安全性教育应纳入学生的课程

By Google Translate

Reference:

site security **requirements** should be included in the **education** curriculum for students

Applications

Can we publish the text as is?

Applications

Can we publish the text as is?

Can a reader get the gist of the text?

Applications

Can we publish the text as is?

Can a reader get the gist of the text?

How much effort to fix the text?

Applications

Can we publish the text as is?

Can a reader get the gist of the text?

How much effort to fix the text?

What type of editing – if any – does this word need?

Applications

Can we publish the text as is?

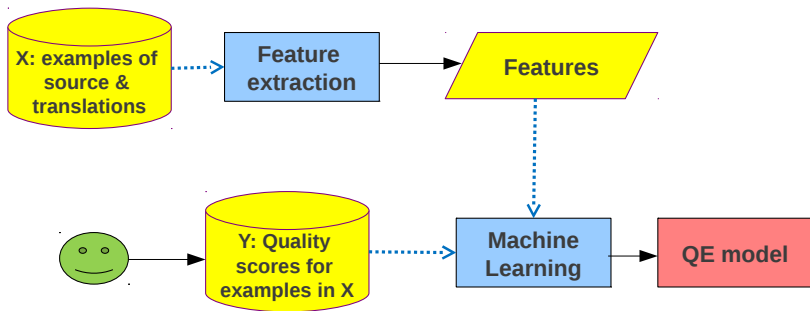
Can a reader get the gist of the text?

How much effort to fix the text?

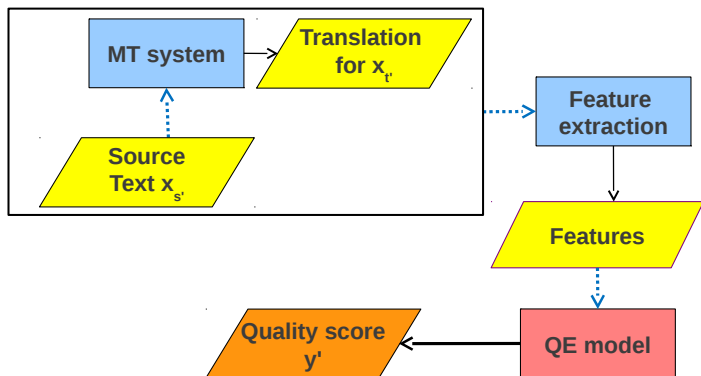
What type of editing – if any – does this word need?

Does this translation need QA?

Framework



Framework



Framework

Main components to build a QE system:

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

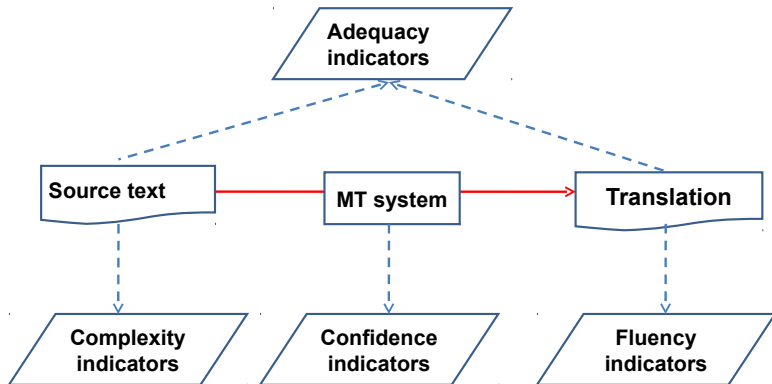
Framework

Main components to build a QE system:

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

All highly dependent on the **level of granularity**: document, sentence, phrase/word

Features



Baseline features for sentence-level

- 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

QuEst

Goal: framework to explore features for QE

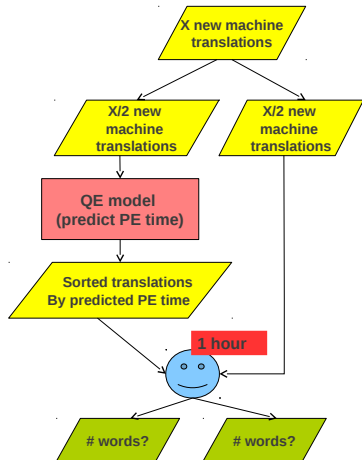
- **Feature extractors** for 150+ features of all types: Java
- **Machine learning:** GPML & scikit-learn toolkit (Python), with wrappers for a number of algorithms, grid search, feature selection



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

Some positive results

Post-editing (PE) subset of sentences predicted as “low PE time” **vs** PE random subset of sentences [Spe11]

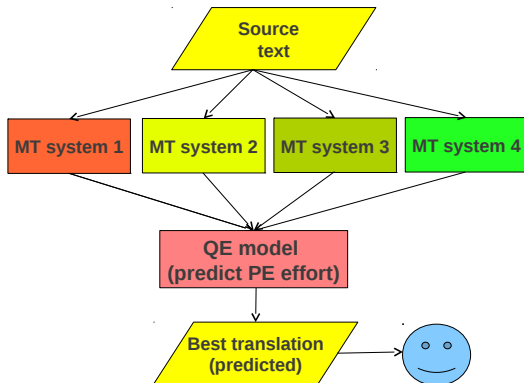


Lang.	no QE	QE
fr-en	0.75 words/sec	1.09 words/sec
en-es	0.32 words/sec	0.57 words/sec

ps.: reading time not included

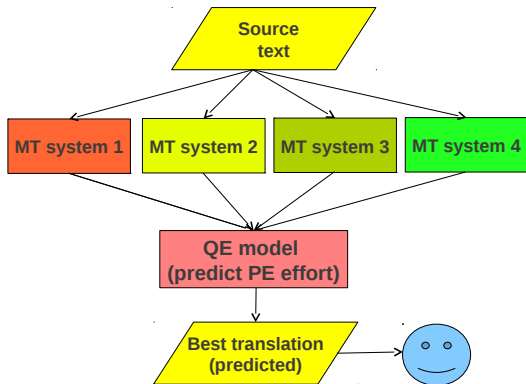
Some positive results

Selecting best translation among 4 MT systems [SRT10]



Some positive results

Selecting best translation among 4 MT systems [SRT10]



Best MT system (on average)	MT system with best QE score
54% accuracy	77% accuracy
0.371 BLEU	0.382 BLEU

Some positive results

SDL's **TrustRank** for prediction at **document-level** [SE10]

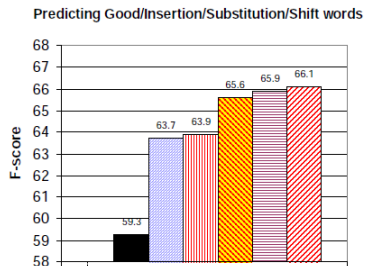
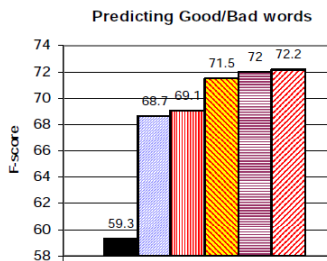
- Training based on BLEU scores for documents
- Ranking of documents by predicted scores, **average BLEU score per quartile**

Domain	Translation Accuracy				
	BLEU				vBLEU Δ [4]
	Q ₁	Q ₁₋₂	Q ₁₋₃	Q ₁₋₄	
WMT09	44.8	43.6	42.4	41.1	+2.1
Travel	38.0	35.1	33.0	31.2	+3.4
Electronics	76.1	72.7	69.6	65.2	+6.5
HiTech	77.9	72.7	66.7	59.0	+11.6
Dom. avg.	-				+5.9

Some positive results

IBM's **Goodness** metric for **word-level** prediction [BHA011]

- Classifier to predict types of edits: **Good/Bad** or **Good/R/I/S**
- Labels generated from aligning MT against its post-edited version (75K sentences, 2.4M words)



Some positive results

IBM's **Goodness** metric for **word-level** prediction [BHA011]

- Classifier to predict types of edits: **Good/Bad** or **Good/R/I/S**
- Labels generated from aligning MT against its post-edited version (75K sentences, 2.4M words)

Good, Bad, Decent

Source أنت مختلف تماماً عن زيد وعمرو 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 you **totally** different from **zaid amr** , and **not to deprive yourself** in and visualize **a basement of imitation and** assimilation .

State of the art

WMT12-13 shared tasks on QE [CBKM⁺12, BBCB⁺13]

- **Sentence-** and **word-level** estimation of **PE effort**

State of the art

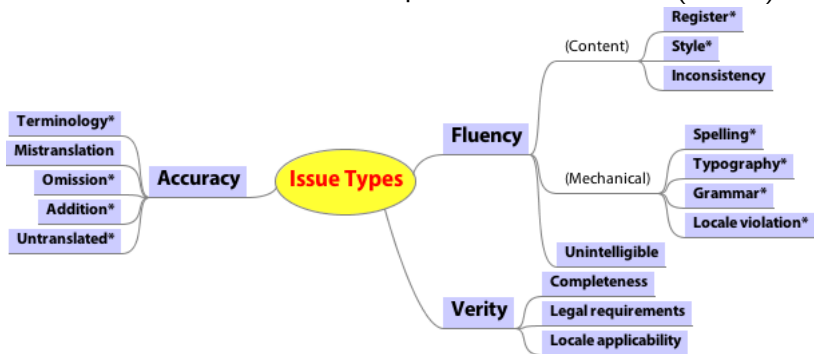
WMT12-13 shared tasks on QE [CBKM⁺12, BBCB⁺13]

- **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
% of edits	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

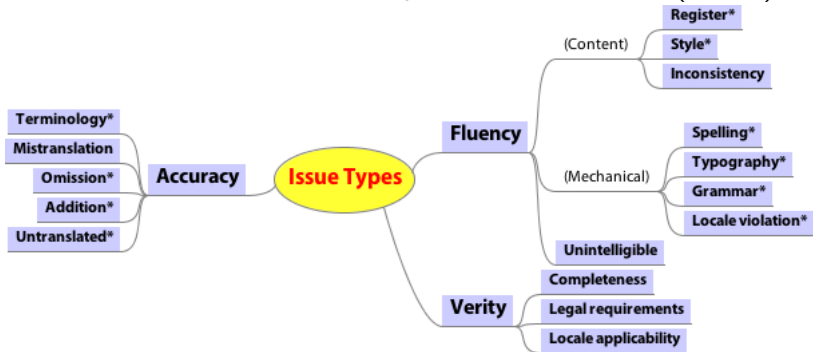
State of the art

WMT14 shared task: can we predict actual issues (MQM)?



State of the art

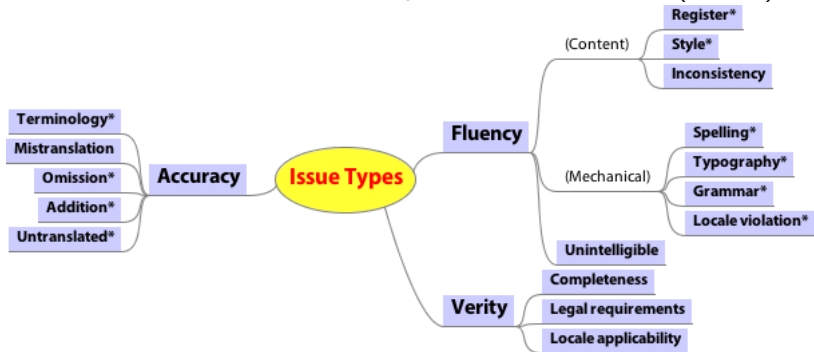
WMT14 shared task: can we predict actual issues (MQM)?



Can we predict errors/issues in human translations?

State of the art

WMT14 shared task: can we predict actual issues (MQM)?



Can we predict errors/issues in human translations?



Outline

- 1 Translation quality
- 2 Reference-based metrics
- 3 Task-based metrics
- 4 Prediction-based metrics
- 5 Conclusions**

Conclusions

- (Machine) Translation evaluation & estimation: still an open problem

Conclusions

- (Machine) Translation evaluation & estimation: still an open problem
- Different metrics for: different purposes/users, needs, levels of granularity and notions of **quality**

Conclusions

- (Machine) Translation evaluation & estimation: still an open problem
- Different metrics for: different purposes/users, needs, levels of granularity and notions of **quality**
- **Quality prediction**: learning of these different notions, but requires labelled data

Conclusions

- (Machine) Translation evaluation & estimation: still an open problem
- Different metrics for: different purposes/users, needs, levels of granularity and notions of **quality**
- **Quality prediction**: learning of these different notions, but requires labelled data
- Estimates useful in **real applications**

Conclusions

- (Machine) Translation evaluation & estimation: still an open problem
- Different metrics for: different purposes/users, needs, levels of granularity and notions of **quality**
- **Quality prediction**: learning of these different notions, but requires labelled data
- Estimates useful in **real applications**
- **Error prediction** (word-level)
 - Still predicting general edits, not **actual errors**

Conclusions

- (Machine) Translation evaluation & estimation: still an open problem
- Different metrics for: different purposes/users, needs, levels of granularity and notions of **quality**
- **Quality prediction**: learning of these different notions, but requires labelled data
- Estimates useful in **real applications**
- **Error prediction** (word-level)
 - Still predicting general edits, not **actual errors**
- Error analysis/prediction for **model improvement**

Translation Quality Evaluation and Estimation

Lucia Specia

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

ASLIB: Translating and the Computer Conference
29 November 2013



The
University
Of
Sheffield.

References I



Autodesk.

Translation and Post-Editing Productivity.

In <http://translate.autodesk.com/productivity.html>, 2011.



Ondrej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia.

Findings of the 2013 Workshop on Statistical Machine Translation.

In *Eighth Workshop on Statistical Machine Translation, WMT-2013*, pages 1–44, Sofia, Bulgaria, 2013.



Nguyen Bach, Fei Huang, and Yaser Al-Onaizan.

Goodness: a method for measuring machine translation confidence.

In *ACL11*, pages 211–219, Portland, Oregon, 2011.



Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia.

Findings of the 2012 workshop on statistical machine translation.

In *Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada, 2012.



J. Carbonell and Y. Wilks.

Machine translation: An in-depth tutorial.

Tutorial, 1991.



Intel.

Being Streetwise with Machine Translation in an Enterprise Neighborhood.

In http://mtmarathon2010.info/JEC2010_Burgett_slides.pptx, 2010.



Maja Popović and Hermann Ney.

Towards automatic error analysis of machine translation output.

Comput. Linguist., 37(4):657–688, 2011.

References II



Radu Soricut and Abdessamad Echiabi.

Trustrank: Inducing trust in automatic translations via ranking.
In *ACL11*, pages 612–621, Uppsala, Sweden, July 2010.



Lucia Specia.

Exploiting Objective Annotations for Measuring Translation Post-editing Effort.
In *Proceedings of the 15th Conference of the European Association for Machine Translation*, pages 73–80, Leuven, 2011.



Lucia Specia, Dhvaj Raj, and Marco Turchi.

Machine translation evaluation versus quality estimation.
Machine Translation, pages 39–50, 2010.



Guillaume Wisniewski, Anil Kumar Singh, Natalia Segal, and François Yvon.

Design and analysis of a large corpus of post-edited translations: Quality estimation, failure analysis and the variability of post-edition.
In *Machine Translation Summit (MT Summit 2013)*, pages 117–124, Nice, France, 2013.



Daniel Zeman, Mark Fishel, Jan Berka, and Ondrej Bojar.

Addicter: What is wrong with my translations?
Prague Bull. Math. Linguistics, 96:79–88, 2011.