Is this translation fit for purpose?

Predicting quality and predicting errors

Lucia Specia

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

Workshop Errare, 21 November 2013

The University Of Sheffield.

Outline

Outline

Machine Translation (MT)

- Errors arising from automatically translating texts from a **source language** into a **target language**
- Several approaches to MT: rules-based, corpus-based (mostly **statistical**), and hybrid
- Over 60 years of research, mature technology, successful commercial applications

Machine Translation (MT)

- Errors arising from automatically translating texts from a **source language** into a **target language**
- Several approaches to MT: rules-based, corpus-based (mostly **statistical**), and hybrid
- Over 60 years of research, mature technology, successful commercial applications
- Still: **frequent** and **grotesque errors**...

Machine Translation (MT)

- Errors arising from automatically translating texts from a **source language** into a **target language**
- Several approaches to MT: rules-based, corpus-based (mostly **statistical**), and hybrid
- Over 60 years of research, mature technology, successful commercial applications
- Still: **frequent** and **grotesque errors**...
- Quality evaluation is core

Machine Translation (MT)

- Errors arising from automatically translating texts from a **source language** into a **target language**
- Several approaches to MT: rules-based, corpus-based (mostly **statistical**), and hybrid
- Over 60 years of research, mature technology, successful commercial applications
- Still: **frequent** and **grotesque errors**...
- Quality evaluation is core

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

Why is evaluation important?

Translation output evaluation

- Compare MT systems
- Measure progress of MT systems over time
- Tune statistical MT systems
- **Diagnose MT systems**

Why is evaluation important?

Translation output evaluation

- Compare MT systems
- Measure progress of MT systems over time
- Tune statistical MT systems
- **Diagnose MT systems**
- Decide on **fitness-for-purpose**: human perception of "high quality translation"

Why is evaluation important?

Translation output evaluation

- Compare MT systems
- Measure progress of MT systems over time
- Tune statistical MT systems
- **Diagnose MT systems**
- Decide on **fitness-for-purpose**: human perception of "high quality translation"

"Evaluation" normally refers to an **aggregate score** for sentences/documents, but can be derived from "error analysis"

Why is evaluation important?

Translation output evaluation

- Compare MT systems
- Measure progress of MT systems over time
- Tune statistical MT systems
- **Diagnose MT systems**
- Decide on **fitness-for-purpose**: human perception of "high quality translation"

"Evaluation" normally refers to an **aggregate score** for sentences/documents, but can be derived from "error analysis"

Error analysis: word/phrase-level linguistic analysis of translation output: types of errors made

Why is evaluation important?

Translation output evaluation

- Compare MT systems
- Measure progress of MT systems over time
- Tune statistical MT systems
- **Diagnose MT systems**
- Decide on **fitness-for-purpose**: human perception of "high quality translation"

"Evaluation" normally refers to an **aggregate score** for sentences/documents, but can be derived from "error analysis"

Error analysis: word/phrase-level linguistic analysis of translation output: types of errors made

Quality vs. errors

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 counts**
  - **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 counts**
  - **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 counts**
  - **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

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/PER/TER, BLEU/NIST, AMBER, ROSE, etc.

Edit distance: WER/PER

WER: Word Error Rate:

- Inherited from ASR
- Minimum proportion of **insertions**, **deletions**, and **substitutions** needed to transform an MT sentence into the reference sentence
- Heavily penalises **reorderings**: correct translation in a wrong/different location: deletion + insertion

$$WER = \frac{S + D + I}{N}$$

Edit distance: TER

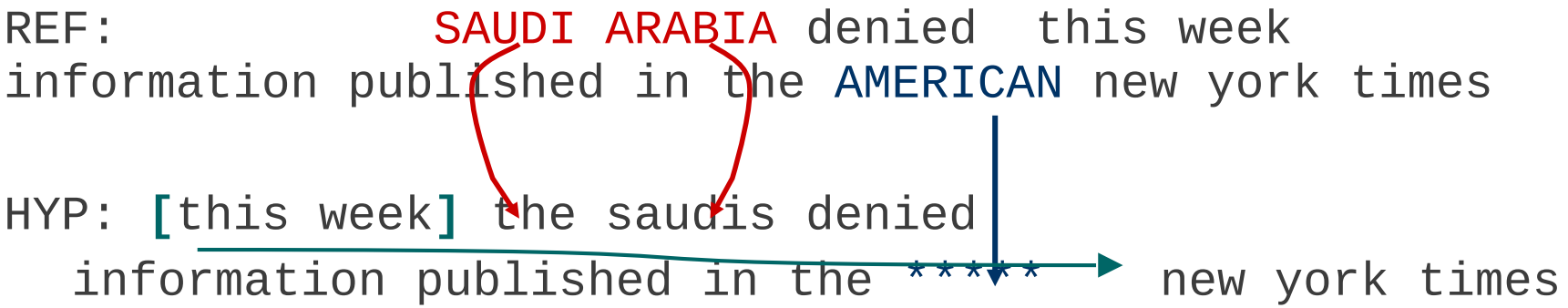
TER: Translation Error Rate

- Adds **shift** operation

Edit distance: TER

TER: Translation Error Rate

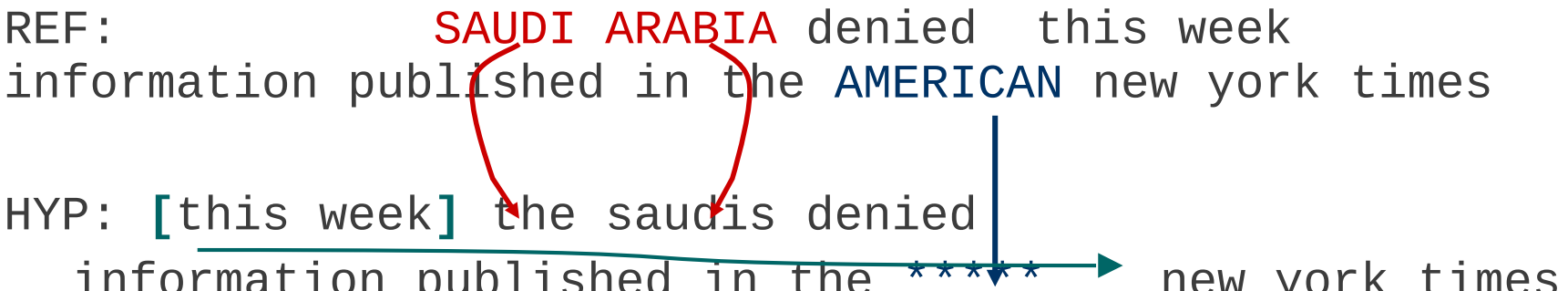
- Adds **shift** operation



Edit distance: TER

TER: Translation Error Rate

- Adds **shift** operation



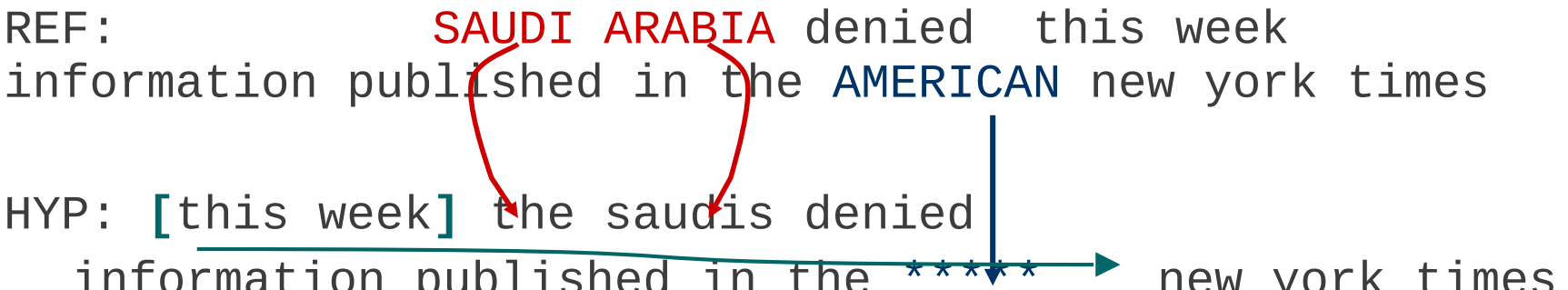
1 Shift, 2 Substitutions, 1 Deletion $\rightarrow$ 4 (not 7) Edits:

$$\text{TER} = \frac{4}{13} = 0.31$$

Edit distance: TER

TER: Translation Error Rate

- Adds **shift** operation



1 Shift, 2 Substitutions, 1 Deletion $\rightarrow$ 4 (not 7) Edits:

$$\text{TER} = \frac{4}{13} = 0.31$$

Human-targeted TER (HTER)

TER between MT and its post-edited version

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

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

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

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

- **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

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

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

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

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

Reference-based automatic metrics

Advantages:

- Fast and cheap, minimal human labour
  - Once test set is created, can be **reused** many times, and an on-going basis during **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
  - Once test set is created, can be **reused** many times, and an on-going basis during **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**

Error analysis

- More fine-grained
- Aimed at **diagnosis** of MT systems, also **quality control** of human translation
- E.g.: Multidimensional Quality Metrics (**MQM**) (manual)
  - Machine and human translation quality

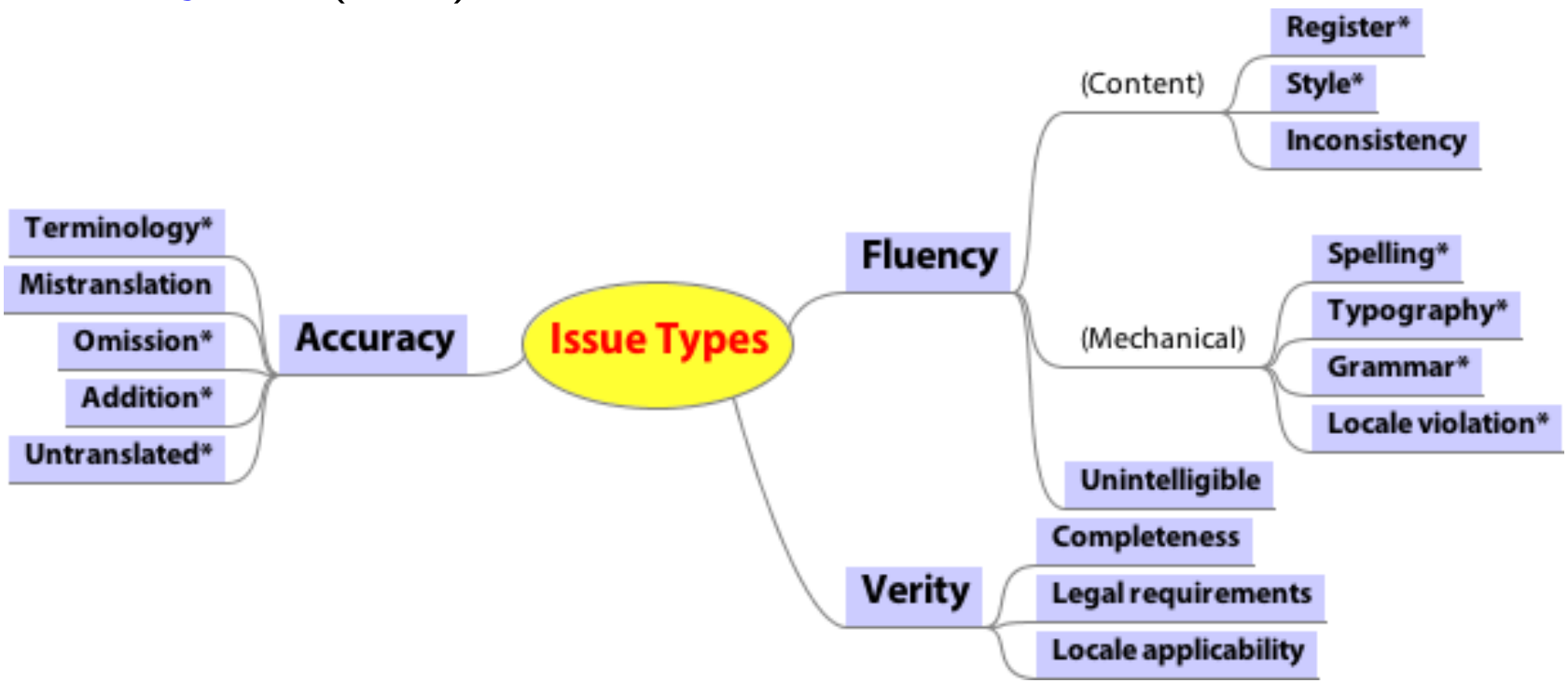
Error analysis

- More fine-grained
- Aimed at **diagnosis** of MT systems, also **quality control** of human translation
- E.g.: Multidimensional Quality Metrics (**MQM**) (manual)
  - Machine and human translation quality

Aims to systematically analyse quality barriers in nearly good translations in order to advance MT area to make them perfect

MQM

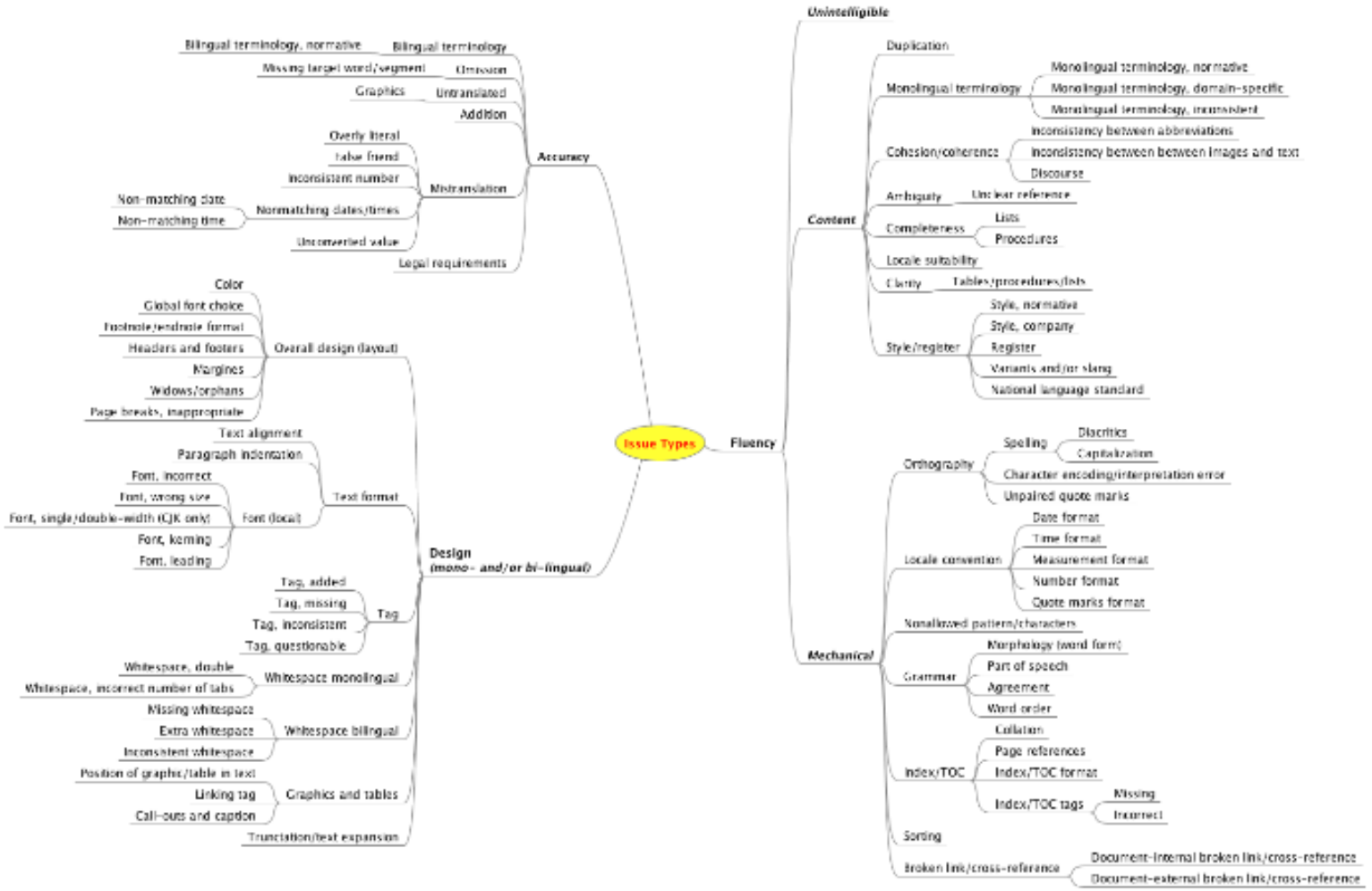
Issue types (core):



Altogether: 120 categories, actual categories chosen based on a specification

MQM

Issue types (all):



Error analysis

- 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

- 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. Talk by Catherine Kobus (17.50)

Outline

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)
- Quality defined by labels in training **data**

Overview

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

Quality = **Can we publish the text as is?**

Overview

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

Quality = **Can we publish the text as is?**

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

Overview

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

Quality = **Can we publish the text as is?**

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

Quality = **How much effort to fix the text?**

Overview

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

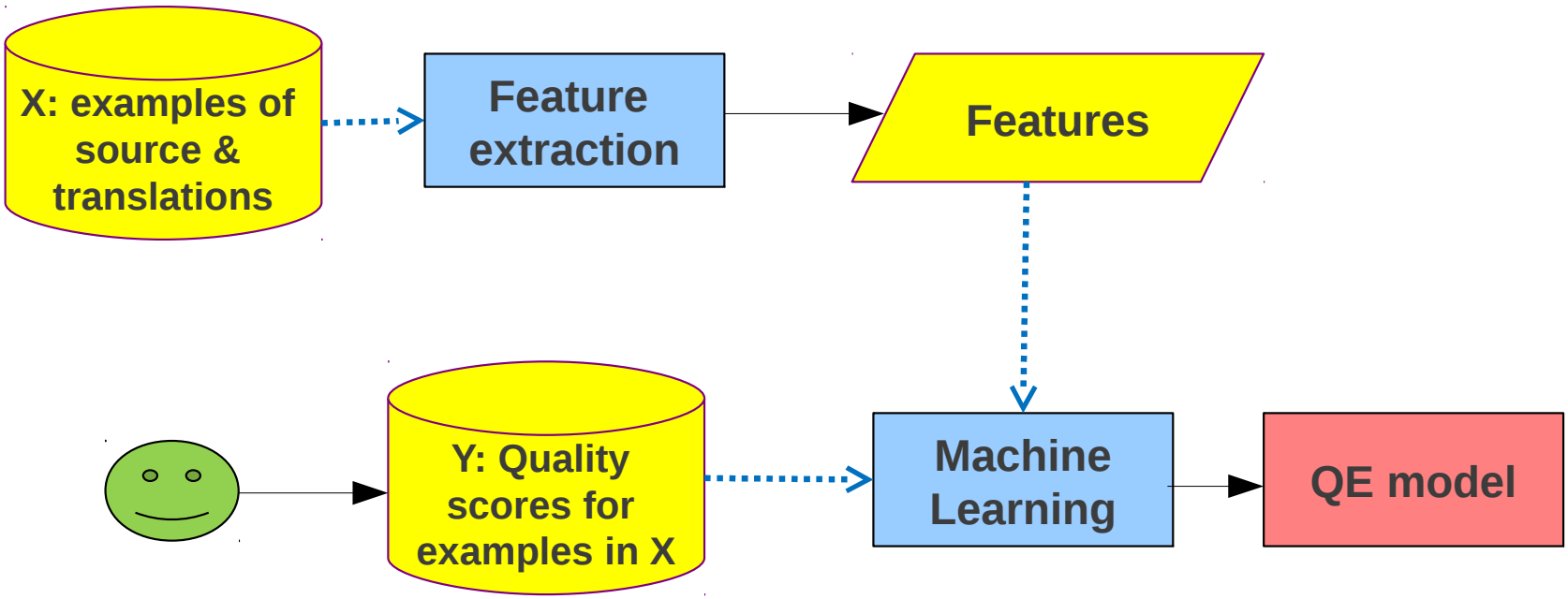
Quality = **Can we publish the text as is?**

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

Quality = **How much effort to fix the text?**

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

Framework


X: examples of source & translations
Feature extraction
Features
Y: Quality scores for examples in X
Machine Learning
QE model

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:

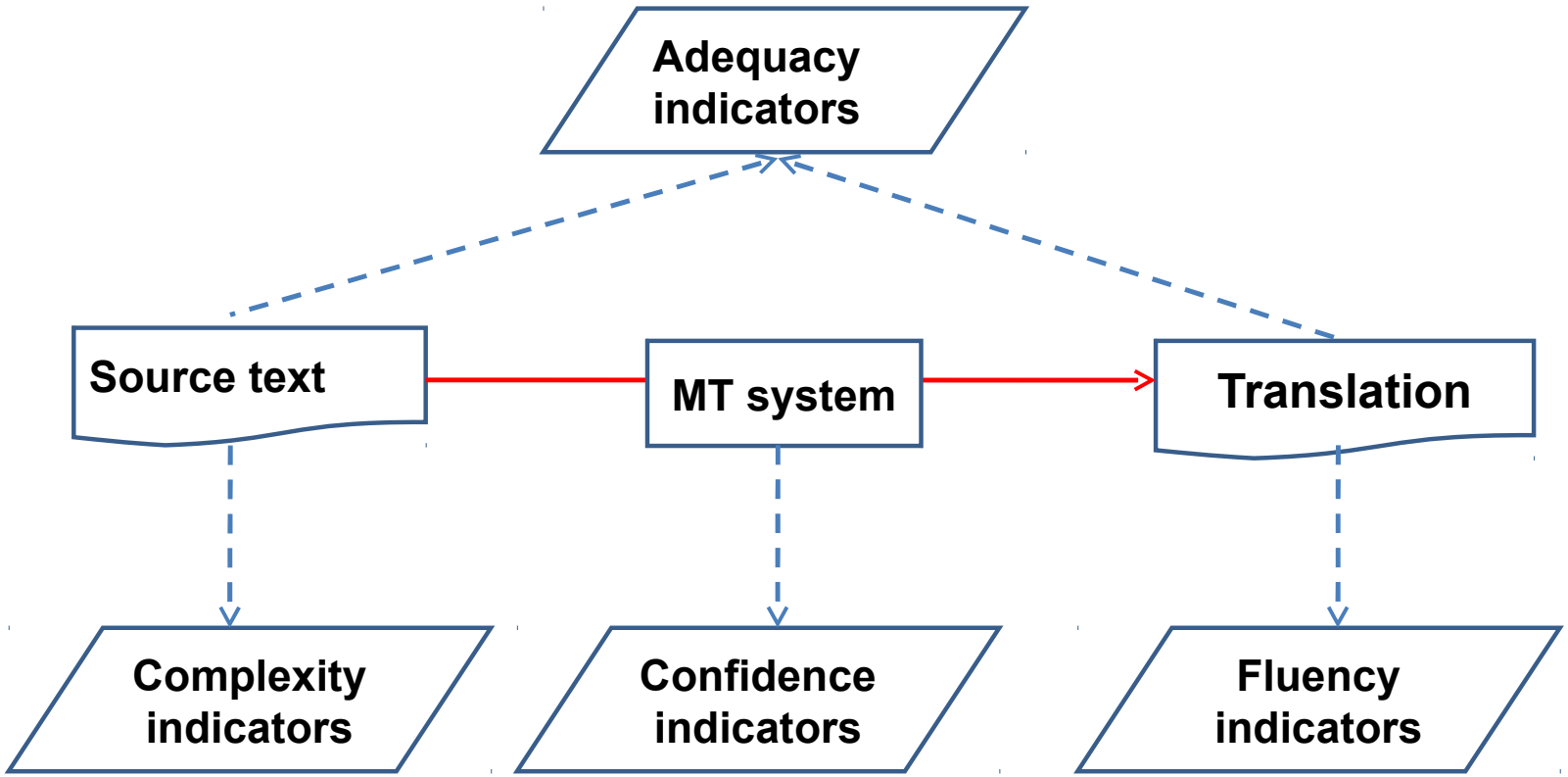
1. Definition of quality: **what to predict**
2. (Human) labelled **data** (for quality/errors)
3. **Features**
4. Machine learning **algorithm**

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

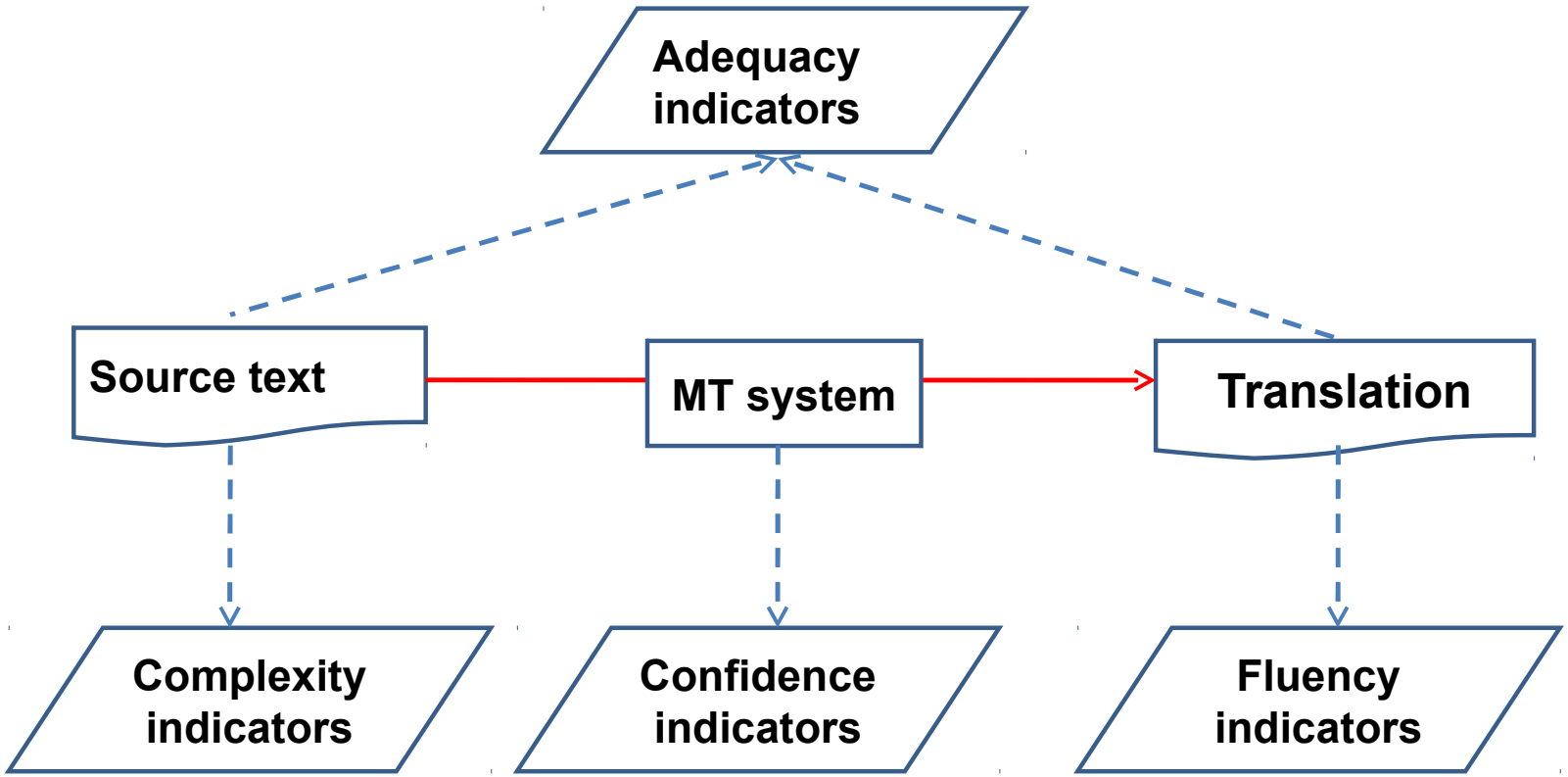
Definition of quality

- 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

Features

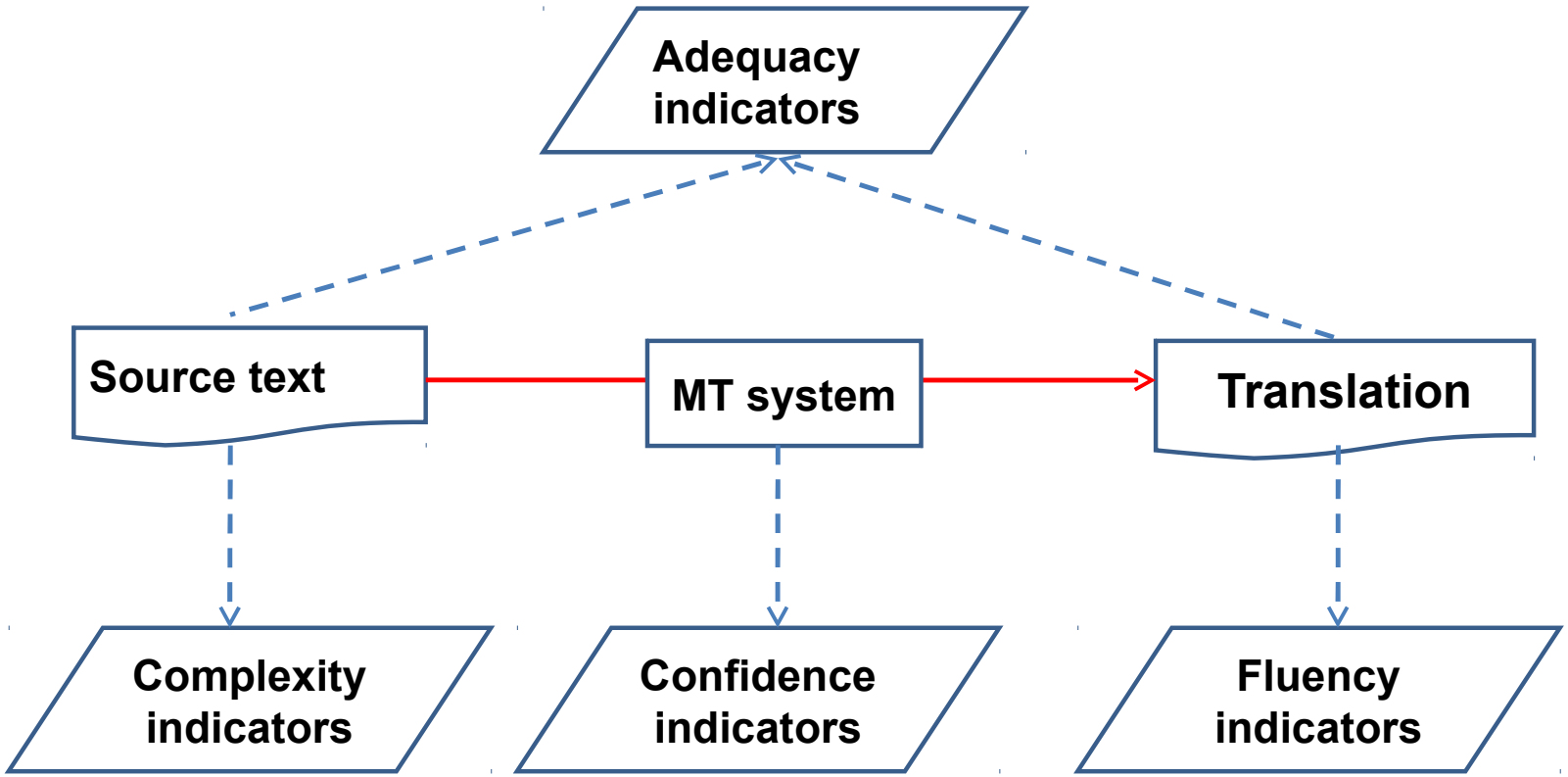

Adequacy indicators
Source text
MT system
Translation
Complexity indicators
Confidence indicators
Fluency indicators

Features



Simple features tend to work better: less sparse, less prone to errors

Features



Simple features tend to work better: less sparse, less prone to errors

Black-box features tend to work better: confidence features already known to MT system

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

Time to post-edit (PE) subset of sentences predicted as "low PE effort" **vs** time to post-edit random subset of sentences [Spe11]

Language	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

Time to post-edit (PE) subset of sentences predicted as "low PE effort" **vs** time to post-edit random subset of sentences [Spe11]

Language	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

Accuracy in selecting best translation among 4 MT systems [SRT10]

Best MT system (on average)	MT system with best QE score
54%	**77%**

Some positive results

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

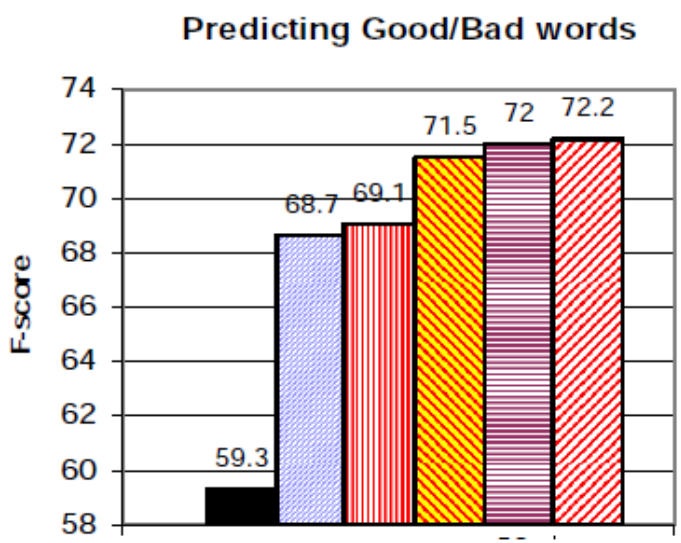
- Training based on BLEU scores for documents
- Features: length, LM, **pseudo-reference**, similarity to training data
- Ranking of documents by predicted scores, **average BLEU score per quartile**

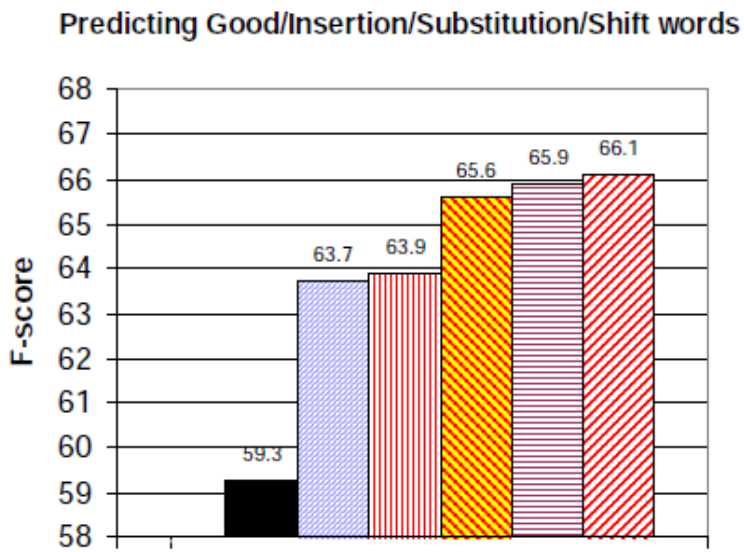
Domain	Translation Accuracy				
	BLEU				vBLEUΔ[4]
	Q_1	Q_{1-2}	Q_{1-3}	Q_{1-4}	
WMT09	44.8	43.6	42.4	41.1	+2.1
Travel	38.0	35.1	33.0	31.2	+3.4
Electronics	76.1	72.7	69.6	65.2	+6.5
HiTech	77.9	72.7	66.7	59.0	+11.6
Dom. avg.	-				**+5.9**

Some positive results

IBM's **Goodness** metric for **word-level** prediction [BHAO11]

- Classifier with sparse binary features (word/phrase pairs, etc.) 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 [BHAO11]

- Classifier with sparse binary features (word/phrase pairs, etc.) 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 and visualize	you totally different from zaid amr , and **not to deprive yourself** in 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

Outline

Agreement between annotators

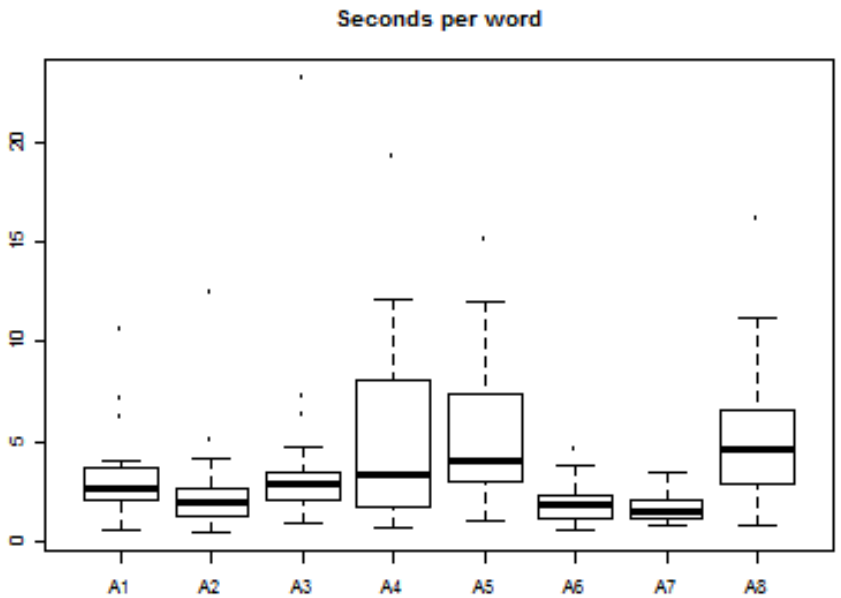
Absolute value judgements: difficult to achieve consistency even in highly controlled settings

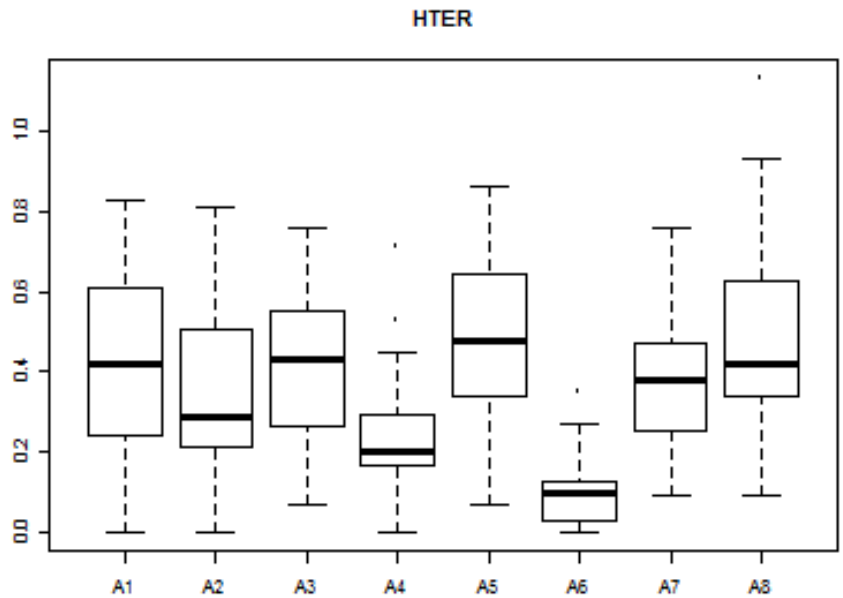
- WMT12 dataset with 1-5 *likert* scores: 30% of dataset discarded
- Remaining annotations had to be scaled

Agreement between annotators

More objective absolute scores

- Post-editing time, HTER (% edits), keystrokes
- Also subject to variance [WSSY13, KARS12]
- E.g.: WPTP12 dataset with 8 annotators:





Agreement between annotators

Embrace variance in a **multi-task learning** (MTL) setting [CS13]: joint modelling of individuals (as **tasks**) and the group

- Accept that there are many possible good translations and agreement cannot be imposed
- Variance also due to: subjectivity of task, typing speed, experience with task, expectations from MT, etc.

Agreement between annotators

Embrace variance in a **multi-task learning** (MTL) setting [CS13]: joint modelling of individuals (as **tasks**) and the group

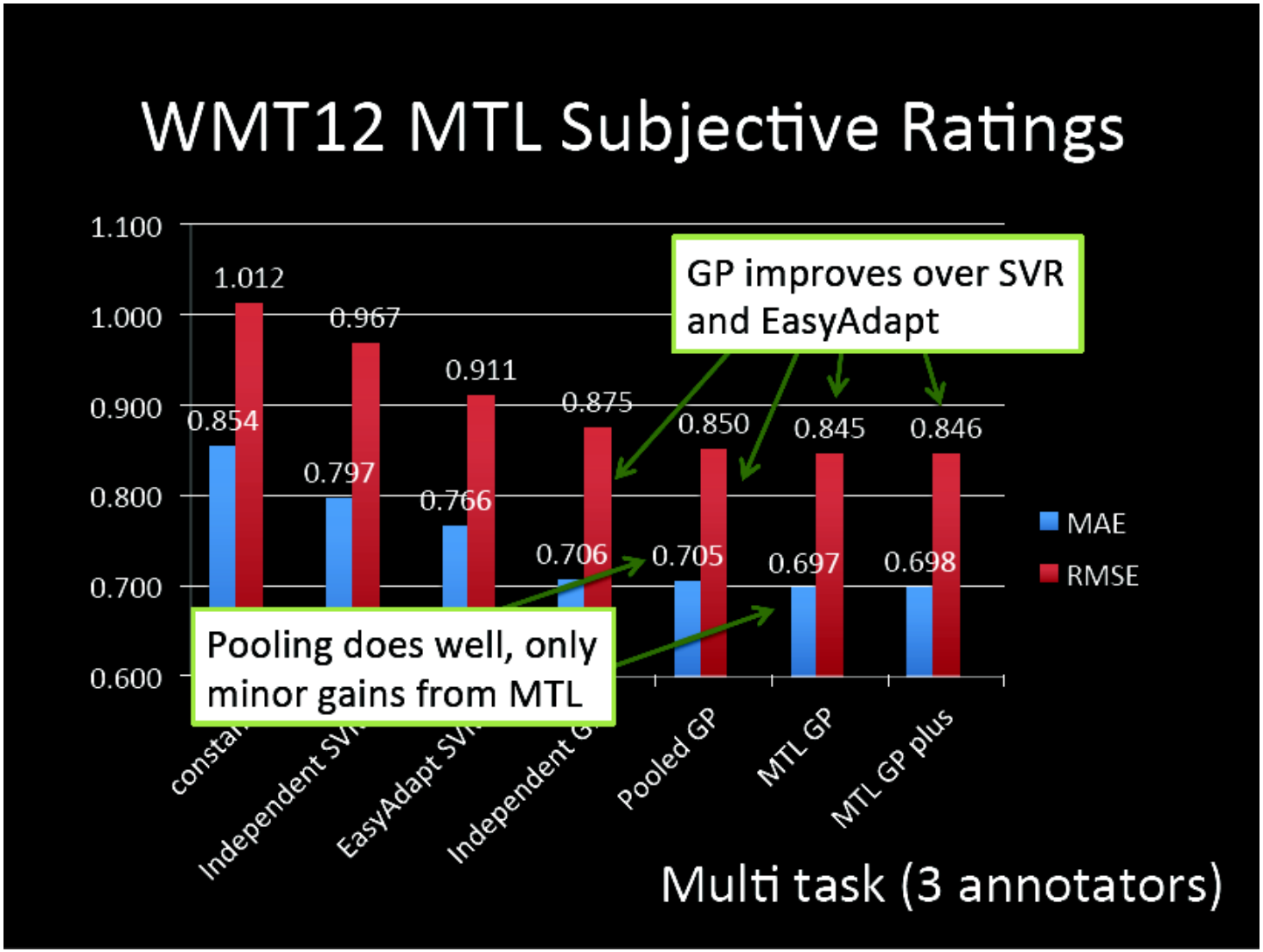
- Accept that there are many possible good translations and agreement cannot be imposed
- Variance also due to: subjectivity of task, typing speed, experience with task, expectations from MT, etc.

Multi-task learning with Gaussian Processes

A kernelised Bayesian non-parametric learning framework. MTL models by representing intra-task transfer explicitly as part of a parameterised kernel function - a multi-task kernel.

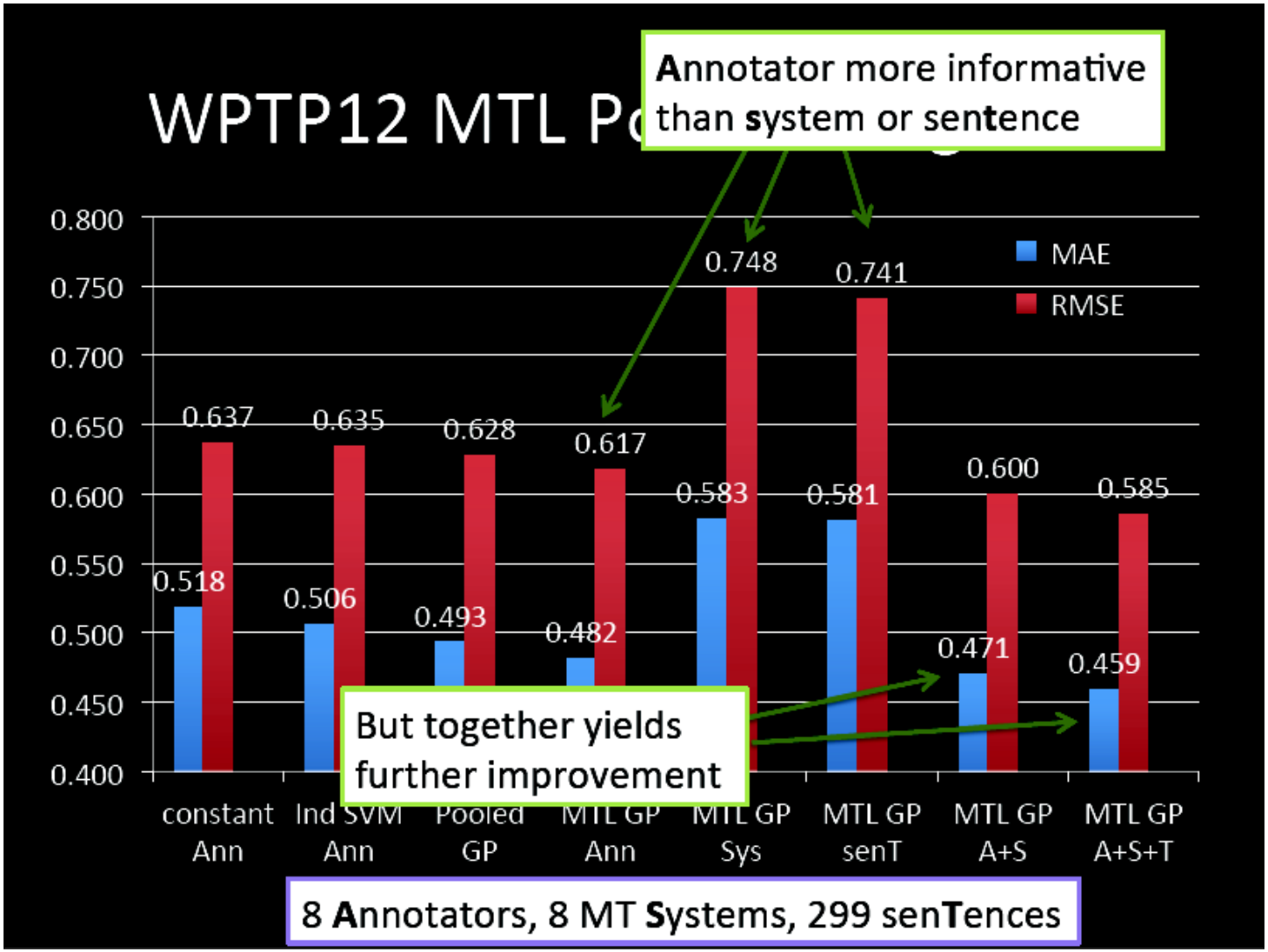
Agreement between annotators

WMT12 dataset: 1-5 likert scores by 3 annotators, 2K sentences



Agreement between annotators

WPTP12 dataset: post-editing time, 8 annotators, 8 MT systems



Agreement between annotators

Relative scores

- WMT13: better results than reference-based metrics

System ID	Kendall's τ with ties penalized
• DFKI logRegFss33	**0.31**
DFKI logRegFss24	0.28
CNGL SVRPLSF1	0.17
CNGL SVRF1	0.17
DCU CCG	0.15
UPC AQE+SEM+LM	0.11
UPC AQE+LeM+ALGPR+LM	0.10
DCU baseline+CCG	0.00
Baseline Random-ranks-with-ties	-0.12
Oracle BLEU	**0.19**
Oracle METEOR-ex	**0.23**

Agreement between annotators

Relative scores

- WMT13: better results than reference-based metrics

System ID	Kendall's τ with ties penalized
• DFKI logRegFss33	**0.31**
DFKI logRegFss24	0.28
CNGL SVRPLSF1	0.17
CNGL SVRF1	0.17
DCU CCG	0.15
UPC AQE+SEM+LM	0.11
UPC AQE+LeM+ALGPR+LM	0.10
DCU baseline+CCG	0.00
Baseline Random-ranks-with-ties	-0.12
Oracle BLEU	**0.19**
Oracle METEOR-ex	**0.23**

- Different task altogether

Annotation costs

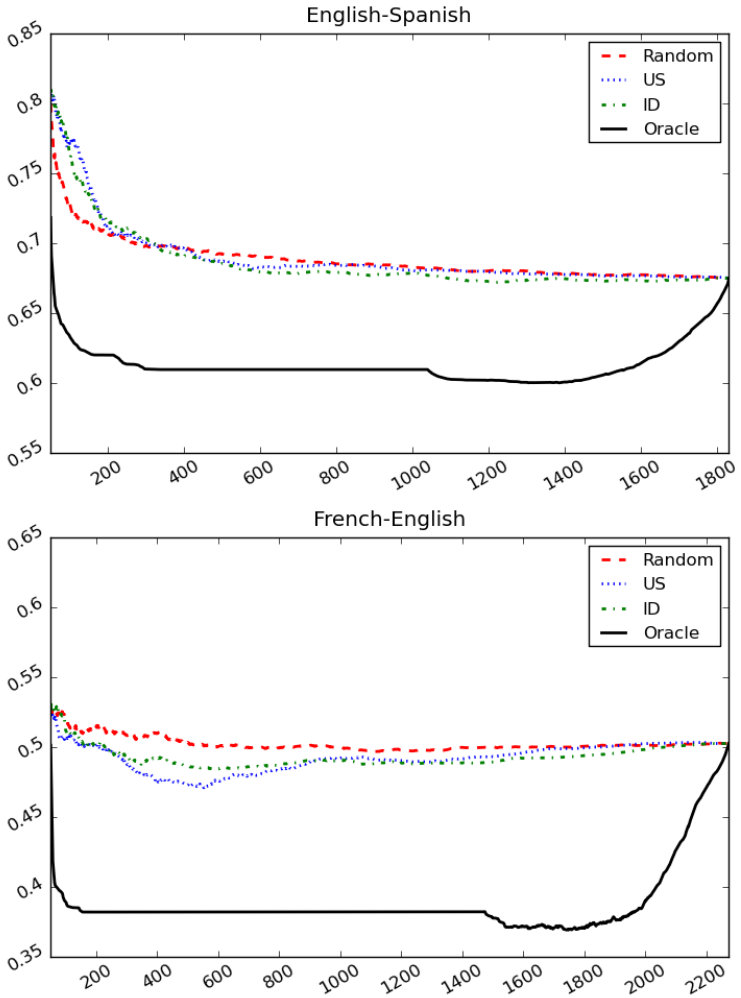
Particularly an issue for word-level prediction. WMT13 task:

Multi-class classification (keep/substitute/delete):

System ID	F_1 Keep	F_1 Substitute	F_1 Delete	Macro-F_1
• LIG FS_MULT	0.83	0.44	0.072	0.45
• LIG ALL_MULT	0.83	0.45	0.064	0.45
UMAC NB	0.62	0.43	0.042	0.36
CNGL GLM	0.83	0.18	0.028	0.35
CNGL GLMd	0.83	0.14	0.034	0.34
UMAC CRF	0.83	0.04	0.012	0.29
Baseline (one class)	0.83	0.00	0.000	0.28

Annotation costs

Active learning to select subset of instances to be annotated at sentence-level [BSC13]



Human translation vs MT quality prediction

Do humans make the same mistakes? Can we capture them in the same way?

- On going work as part of QTLaunchPad (manual analysis)
- Talk by Natalie Kübler et al. (Friday)

Outline

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**
  - From automatic error analysis? (M. Popović)
  - On-going work (QTLaunchPad) - from human labels

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**
  - From automatic error analysis? (M. Popović)
  - On-going work (QTLaunchPad) - from human labels
- Error analysis/prediction for **model improvement**
  - Michel Simard's talk (Friday)

Is this translation fit for purpose?

Predicting quality and predicting errors

Lucia Specia

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

Workshop Errare, 21 November 2013

The University Of Sheffield.

References I

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.

Daniel Beck, Lucia Specia, and Trevor Cohn.
Reducing annotation effort for quality estimation via active learning.
In *51st Annual Meeting of the Association for Computational Linguistics: Short Papers*, ACL-2013, pages 543–548, Sofia, Bulgaria, 2013.

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.

Trevor Cohn and Lucia Specia.
Modelling annotator bias with multi-task gaussian processes: An application to machine translation quality estimation.
In *51st Annual Meeting of the Association for Computational Linguistics*, ACL-2013, pages 32–42, Sofia, Bulgaria, 2013.

J. Carbonell and Y. Wilks.
Machine translation: An in-depth tutorial.
Tutorial, 1991.

References II

Maarit Koponen, Wilker Aziz, Luciana Ramos, and Lucia Specia.

Post-editing time as a measure of cognitive effort.
In *AMTA 2012 Workshop on Post-Editing Technology and Practice*, WPTP-2012, pages 11–20, San Diego, USA, 2012.

Maja Popović and Hermann Ney.

Towards automatic error analysis of machine translation output.
Comput. Linguist., 37(4):657–688, 2011.

Radu Soricut and Abdessamad Echihabi.

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, Dhwaj 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.

References III

Daniel Zeman, Mark Fishel, Jan Berka, and Ondrej Bojar.
Addicter: What is wrong with my translations?
Prague Bull. Math. Linguistics, 96:79–88, 2011.