Estimating machine translation quality

State-of-the-art systems and open issues

Lucia Specia

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

6 September 2012

Outline

Outline

1. Quality Estimation
2. Shared Task
3. Open issues
4. Conclusions

Overview

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

Overview

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

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

Overview

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

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 translated texts

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 translated texts

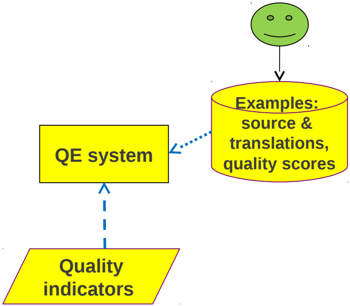
Quality = **Can we publish it as is?**

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

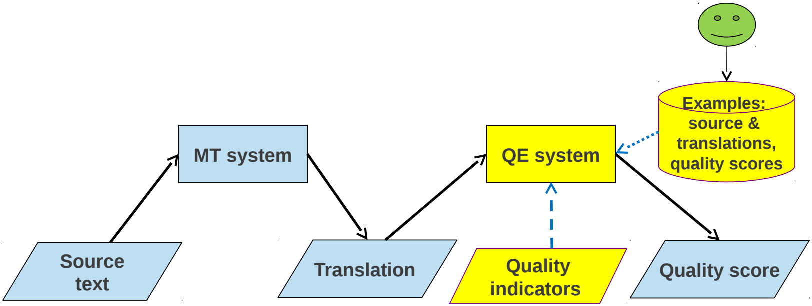
Quality = **Is it worth post-editing it?**

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

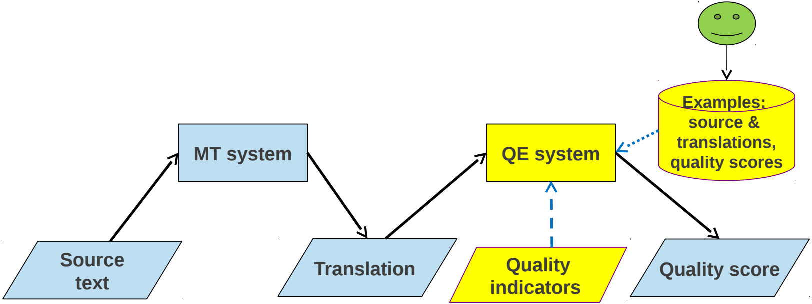
Framework


Examples:
source &
translations,
quality scores
QE system
Quality
indicators

Framework


Examples:
source &
translations,
quality scores
MT system
QE system
Source
text
Translation
Quality
indicators
Quality score

Framework



No access to reference translations: supervised machine learning techniques to **predict** quality scores

Background

- Also called **confidence estimation**, started in 2002/3
  - Inspired by confidence scores in ASR: word posterior probabilities

Background

- Also called **confidence estimation**, started in 2002/3
  - Inspired by confidence scores in ASR: word posterior probabilities
- JHU Workshop in 2003
  - Estimate BLEU/NIST/WER: difficult to interpret
  - A "hard to beat" baseline: **MT is always bad**
  - Poor results, no use in applications

Background

- Also called **confidence estimation**, started in 2002/3
  - Inspired by confidence scores in ASR: word posterior probabilities
- JHU Workshop in 2003
  - Estimate BLEU/NIST/WER: difficult to interpret
  - A "hard to beat" baseline: **MT is always bad**
  - Poor results, no use in applications
- New surge in interest from 2008/9
  - Better MT systems

Background

- Also called **confidence estimation**, started in 2002/3
  - Inspired by confidence scores in ASR: word posterior probabilities
- JHU Workshop in 2003
  - Estimate BLEU/NIST/WER: difficult to interpret
  - A "hard to beat" baseline: **MT is always bad**
  - Poor results, no use in applications
- New surge in interest from 2008/9
  - Better MT systems ✓

Background

- Also called **confidence estimation**, started in 2002/3
  - Inspired by confidence scores in ASR: word posterior probabilities
- JHU Workshop in 2003
  - Estimate BLEU/NIST/WER: difficult to interpret
  - A "hard to beat" baseline: **MT is always bad**
  - Poor results, no use in applications
- New surge in interest from 2008/9
  - Better MT systems ✓
  - MT used in translation industry

Background

- Also called **confidence estimation**, started in 2002/3
  - Inspired by confidence scores in ASR: word posterior probabilities
- JHU Workshop in 2003
  - Estimate BLEU/NIST/WER: difficult to interpret
  - A "hard to beat" baseline: **MT is always bad**
  - Poor results, no use in applications
- New surge in interest from 2008/9
  - Better MT systems ✓
  - MT used in translation industry ✓

Background

- Also called **confidence estimation**, started in 2002/3
    - Inspired by confidence scores in ASR: word posterior probabilities
- JHU Workshop in 2003
    - Estimate BLEU/NIST/WER: difficult to interpret
    - A "hard to beat" baseline: **MT is always bad**
    - Poor results, no use in applications
- New surge in interest from 2008/9
    - Better MT systems ✓
    - MT used in translation industry ✓
    - Estimate more interpretable metrics: post-editing (PE) effort (human scores, time, % edits to fix)

Background

- Also called **confidence estimation**, started in 2002/3
  - Inspired by confidence scores in ASR: word posterior probabilities
- JHU Workshop in 2003
  - Estimate BLEU/NIST/WER: difficult to interpret
  - A "hard to beat" baseline: **MT is always bad**
  - Poor results, no use in applications
- New surge in interest from 2008/9
  - Better MT systems ✓
  - MT used in translation industry ✓
  - Estimate more interpretable metrics: post-editing (PE) effort (human scores, time, % edits to fix)
  - Some positive results

Some positive results

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

Some positive results

- **Time to post-edit** 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

Some positive results

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

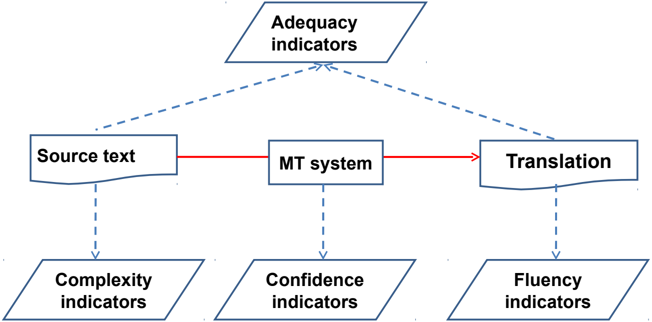
Language	no QE	QE
fr-en	0.75 words/sec	**1.09** words/sec
en-es	0.32 words/sec	**0.57** words/sec

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

Best MT system	Highest QE score
54%	**77%**

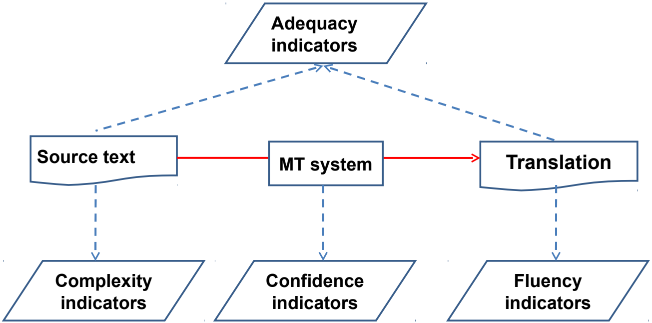
Current approaches

- **Quality indicators**



Current approaches

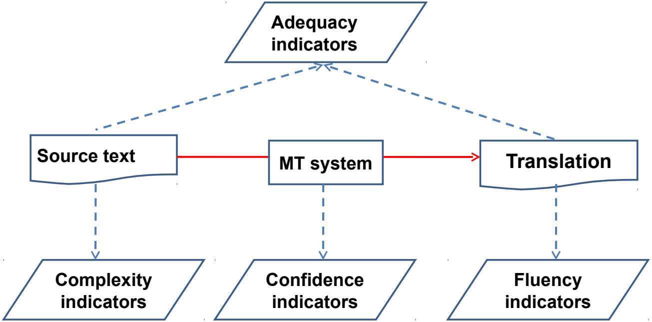
- **Quality indicators**



- **Learning algorithms**: range of regression, classification, ranking algorithms

Current approaches

- **Quality indicators**



- **Learning algorithms**: range of regression, classification, ranking algorithms
- **Datasets**: few with absolute human scores (1-4 scores, PE time, edit distance), WMT data with relative scores

Outline

Objectives

- WMT-12 – joint work with Radu Soricut (Google)

Objectives

- WMT-12 – joint work with Radu Soricut (Google)
- First common ground for development and comparison of QE systems, focusing on **sentence-level** estimation of **PE effort**:

Objectives

- WMT-12 – joint work with Radu Soricut (Google)
- First common ground for development and comparison of QE systems, focusing on **sentence-level** estimation of **PE effort**:
  - Identify (new) effective **features**

Objectives

- WMT-12 – joint work with Radu Soricut (Google)
- First common ground for development and comparison of QE systems, focusing on **sentence-level** estimation of **PE effort**:
  - Identify (new) effective **features**
  - Identify most suitable **machine learning techniques**

Objectives

- WMT-12 – joint work with Radu Soricut (Google)
- First common ground for development and comparison of QE systems, focusing on **sentence-level** estimation of **PE effort**:
  - Identify (new) effective **features**
  - Identify most suitable **machine learning techniques**
  - Test (new) automatic **evaluation metrics**

Objectives

- WMT-12 – joint work with Radu Soricut (Google)
- First common ground for development and comparison of QE systems, focusing on **sentence-level** estimation of **PE effort**:
  - Identify (new) effective **features**
  - Identify most suitable **machine learning techniques**
  - Test (new) automatic **evaluation metrics**
  - Establish the **state of the art performance** in the field

Objectives

- WMT-12 – joint work with Radu Soricut (Google)
- First common ground for development and comparison of QE systems, focusing on **sentence-level** estimation of **PE effort**:
  - Identify (new) effective **features**
  - Identify most suitable **machine learning techniques**
  - Test (new) automatic **evaluation metrics**
  - Establish the **state of the art performance** in the field
  - Contrast **regression** and **ranking** techniques

Objectives

Datasets

English → Spanish

- **English** source sentences

Datasets

English → Spanish

- **English** source sentences
- **Spanish** MT outputs (PBSMT Moses)

Datasets

English $\rightarrow$ Spanish

- **English** source sentences
- **Spanish** MT outputs (PBSMT Moses)
- **Post-edited** output by 1 professional translator

Datasets

English → Spanish

- **English** source sentences
- **Spanish** MT outputs (PBSMT Moses)
- **Post-edited** output by 1 professional translator
- Effort **scores** by 3 professional translators, scale 1-5, averaged

Datasets

English → Spanish

- **English** source sentences
- **Spanish** MT outputs (PBSMT Moses)
- **Post-edited** output by 1 professional translator
- Effort **scores** by 3 professional translators, scale 1-5, averaged
- Human Spanish translation (original **references**)

Datasets

English → Spanish

- **English** source sentences
- **Spanish** MT outputs (PBSMT Moses)
- **Post-edited** output by 1 professional translator
- Effort **scores** by 3 professional translators, scale 1-5, averaged
- Human Spanish translation (original **references**)
- **# Instances**
  - Training: 1832
  - Blind test: 422

Datasets

Annotation guidelines

3 **human judges** for PE effort assigning 1-5 **scores** for $\langle$source, MT output, PE output$\rangle$

- **[1]** The MT output is incomprehensible, with little or no information transferred accurately. It cannot be edited, needs to be translated from scratch.
- **[2]** About 50-70% of the MT output needs to be edited. It requires a significant editing effort in order to reach publishable level.
- **[3]** About 25-50% of the MT output needs to be edited. It contains different errors and mistranslations that need to be corrected.
- **[4]** About 10-25% of the MT output needs to be edited. It is generally clear and intelligible.
- **[5]** The MT output is perfectly clear and intelligible. It is not necessarily a perfect translation, but requires little to no editing.

Resources provided

SMT resources for training and test sets:

- SMT training corpus (Europarl and News-documentaries)
- LMs: 5-gram LM; 3-gram LM and 1-3-gram counts
- IBM Model 1 table (Giza)
- Word-alignment file as produced by *grow-diag-final*
- Phrase table with word alignment information
- Moses configuration file used for decoding
- Moses run-time log: model component values, word graph, etc.

Resources provided

Two sub-tasks:

- **Scoring**: predict a score in [1-5] for each test instance
- **Ranking**: sort all test instances best-worst

Evaluation metrics

Scoring metrics - standard **MAE** and **RMSE**

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

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (H(s_i) - V(s_i))^2}{N}}$$

$N = |S|$
$H(s_i)$ is the predicted score for s_i
$V(s_i)$ the is human score for s_i

Evaluation metrics

Ranking metrics **Spearman's** rank correlation and new metric: **DeltaAvg**

For $S_1, S_2, \ldots, S_n$ quantiles:

$$\text{DeltaAvg}_V[n] = \frac{\sum_{k=1}^{n-1} V(S_{1,k})}{n-1} - V(S)$$

$V(S)$: extrinsic function measuring the "quality" of set S

Evaluation metrics

Ranking metrics **Spearman's** rank correlation and new metric: **DeltaAvg**

For $S_1, S_2, \ldots, S_n$ quantiles:

$$\text{DeltaAvg}_V[n] = \frac{\sum_{k=1}^{n-1} V(S_{1,k})}{n-1} - V(S)$$

$V(S)$: extrinsic function measuring the "quality" of set S

Average **human scores** (1-5) of set S

Evaluation metrics

DeltaAvg

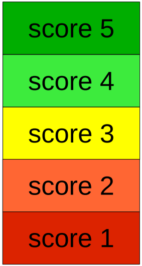
Example 1: n=2, quantiles S_1, S_2

DeltaAvg[2] $= V(S_1) - V(S)$

"Quality of the top half compared to the overall quality"

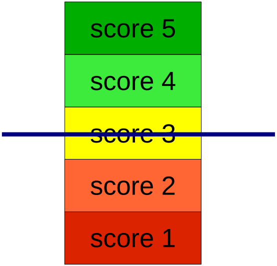
Average **human scores** of top half compared to average **human scores** of complete set

Evaluation metrics



Average **human** score: 3

Evaluation metrics



Average **human** score: 3

N = 2

DeltaAvg[2]

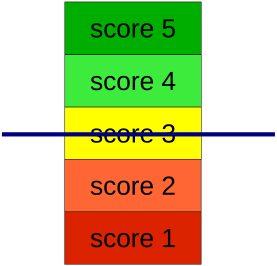
Random = [3 - **3**] = 0

QE = [3.8 - **3**] = 0.8

Oracle = [4.2 - **3**] = 1.2

Lowerb = [1.8 - 3] = -1.2

Evaluation metrics



Average **human** score: 3

N = 2
DeltaAvg[2]

Random	= [3 - **3**]	= 0
QE	= [3.8 - **3**]	= 0.8
Oracle	= [4.2 - **3**]	= 1.2
Lowerb	= [1.8 - 3]	= -1.2

Average "**human**" score of top 50% selected after ranking based on **QE** score.
QE score can be on any scale...

Evaluation metrics

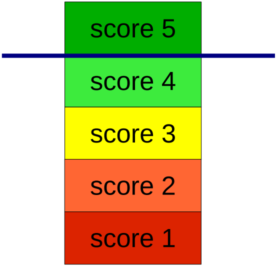
DeltaAvg

Example 2: n=3, quantiles S_1, S_2, S_3

DeltaAvg[3] $= \frac{(V(S_1)-V(S))+(V(S_{1,2})-V(S))}{2}$

Average **human scores** of top third compared to average **human scores** of complete set; average **human scores** of top two thirds compared to average **human scores** of complete set, averaged

Evaluation metrics



Average **human** score: 3

N = 5

DeltaAvg[5]

Random = [3 - **3**] = 0

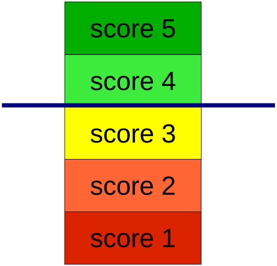
Oracle$_1$ = [5 - **3**] = 2

Lowerb$_1$ = [1 - **3**] = -2

...

QE$_1$ = [4.1 - **3**] = 1.1

Evaluation metrics



Average **human** score: 3

N = 5

DeltaAvg[5]

Random = [3 - **3**] = 0

Oracle$_1$ = [5 - **3**] = 2

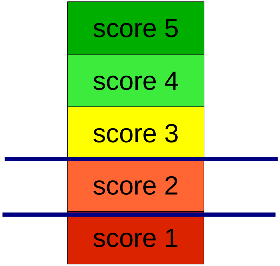
Lowerb$_1$ = [1 - **3**] = -2

...

QE$_1$ = [4.1 - **3**] = 1.1

QE$_{1,2}$ = [3.9 - **3**] = 0.9

Evaluation metrics



Average **human** score: 3

N = 5

DeltaAvg[5]

Random = [3 - **3**] = 0

Oracle$_1$ = [5 - **3**] = 2

Lowerb$_1$ = [1 - **3**] = -2

...

QE$_1$ = [4.1 - **3**] = 1.1

QE$_{1,2}$ = [3.9 - **3**] = 0.9

QE$_{1,2,3}$ = [3.5 - **3**] = 0.5

QE$_{1,2,3,4}$ = [3.3 - **3**] = 0.3

DeltaAvg[5] = (1.1+0.9+0.5+0.3)/4

= 0.7

Evaluation metrics

Final DeltaAvg metric

$$\text{DeltaAvg}_V = \frac{\sum_{n=2}^{N} \text{DeltaAvg}_V[n]}{N-1}$$

where $N = |S|/2$

Evaluation metrics

Final DeltaAvg metric

$$\text{DeltaAvg}_V = \frac{\sum_{n=2}^{N} \text{DeltaAvg}_V[n]}{N-1}$$

where $N = |S|/2$

Average DeltaAvg$[n]$ for all n, $2 \leq n \leq |S|/2$

Participants

ID	Participating team
PRHLT-UPV	Universitat Politecnica de Valencia, Spain
UU	Uppsala University, Sweden
SDLLW	SDL Language Weaver, USA
Loria	LORIA Institute, France
UPC	Universitat Politecnica de Catalunya, Spain
DFKI	DFKI, Germany
WLV-SHEF	Univ of Wolverhampton & Univ of Sheffield, UK
SJTU	Shanghai Jiao Tong University, China
DCU-SYMC	Dublin City University, Ireland & Symantec, Ireland
UEdin	University of Edinburgh, UK
TCD	Trinity College Dublin, Ireland

One or two systems per team, most teams submitting for ranking and scoring sub-tasks

Baseline system

Feature extraction software – system-independent 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

Feature extraction software – system-independent 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 γ, ϵ and C optimized using a grid-search and 5-fold cross validation on the training set

Results - ranking sub-task

System ID	DeltaAvg	Spearman Corr
• SDLLW_M5PbestDeltaAvg	0.63	0.64
• SDLLW_SVM	0.61	0.60
UU_bltk	0.58	0.61
UU_best	0.56	0.62
TCD_M5P-resources-only*	0.56	0.56
Baseline (17FFs SVM)	0.55	0.58
PRHLT-UPV	0.55	0.55
UEdin	0.54	0.58
SJTU	0.53	0.53
WLV-SHEF_FS	0.51	0.52
WLV-SHEF_BL	0.50	0.49
DFKI_morphPOSibm1LM	0.46	0.46
DCU-SYMC_unconstrained	0.44	0.41
DCU-SYMC_constrained	0.43	0.41
TCD_M5P-all*	0.42	0.41
UPC_1	0.22	0.26
UPC_2	0.15	0.19

• = winning submissions

gray area = not different from baseline

* = bug-fix was applied after the submission

Results - ranking sub-task

Oracle methods: associate various metrics in a oracle manner to the test input:

- **Oracle Effort**: the gold-label Effort
- **Oracle HTER**: the HTER metric against the post-edited translations as reference

System ID	DeltaAvg	Spearman Corr
Oracle Effort	0.95	1.00
Oracle HTER	0.77	0.70

Results - scoring sub-task

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
Baseline (17FFs SVM)	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

Discussion

New and effective quality indicators (features)

- Most participating systems use **external resources**: parsers, POS taggers, NER, etc. → variety of features

Discussion

New and effective quality indicators (features)

- Most participating systems use **external resources**: parsers, POS taggers, NER, etc. → variety of features
- Many tried to exploit **linguistically-oriented features**

Discussion

New and effective quality indicators (features)

- Most participating systems use **external resources**: parsers, POS taggers, NER, etc. $\rightarrow$ variety of features
- Many tried to exploit **linguistically-oriented features**
  - none or modest improvements (e.g. WLV-SHEF)

Discussion

New and effective quality indicators (features)

- Most participating systems use **external resources**: parsers, POS taggers, NER, etc. $\rightarrow$ variety of features
- Many tried to exploit **linguistically-oriented features**
  - none or modest improvements (e.g. WLV-SHEF)
  - high performance (e.g. "UU" with parse trees)

Discussion

New and effective quality indicators (features)

- Most participating systems use **external resources**: parsers, POS taggers, NER, etc. → variety of features
- Many tried to exploit **linguistically-oriented features**
  - none or modest improvements (e.g. WLV-SHEF)
  - high performance (e.g. "UU" with parse trees)
- **Good features**:
  - **confidence**: model components from SMT decoder

Discussion

New and effective quality indicators (features)

- Most participating systems use **external resources**: parsers, POS taggers, NER, etc. $\rightarrow$ variety of features
- Many tried to exploit **linguistically-oriented features**
  - none or modest improvements (e.g. WLV-SHEF)
  - high performance (e.g. "UU" with parse trees)
- **Good features**:
  - **confidence**: model components from SMT decoder
  - **pseudo-reference**: agreement between 2 SMT systems
  - **fuzzy-match like**: source (and target) similarity with SMT training corpus (LM, etc)

Discussion

Machine Learning techniques

- Best performing: **Regression Trees** (M5P) and **SVR**

Discussion

Machine Learning techniques

- Best performing: **Regression Trees** (M5P) and **SVR**
  - M5P Regression Trees: compact models, less overfitting, "readable"

Discussion

Machine Learning techniques

- Best performing: **Regression Trees** (M5P) and **SVR**
  - M5P Regression Trees: compact models, less overfitting, "readable"
  - SVRs: easily overfit with small training data and large feature set

Discussion

Machine Learning techniques

- Best performing: **Regression Trees** (M5P) and **SVR**
  - M5P Regression Trees: compact models, less overfitting, “readable”
  - SVRs: easily overfit with small training data and large feature set
- **Feature selection** crucial in this setup

Discussion

Machine Learning techniques

- Best performing: **Regression Trees** (M5P) and **SVR**
  - M5P Regression Trees: compact models, less overfitting, “readable”
  - SVRs: easily overfit with small training data and large feature set
- **Feature selection** crucial in this setup
- **Structured learning** techniques: “UU” submissions (tree kernels)

Discussion

Evaluation metrics

- DeltaAvg $\rightarrow$ suitable for the ranking task

Discussion

Evaluation metrics

- DeltaAvg $\rightarrow$ suitable for the ranking task
  - **automatic** and **deterministic** (and therefore consistent)

Discussion

Evaluation metrics

- DeltaAvg $\rightarrow$ suitable for the ranking task
  - **automatic** and **deterministic** (and therefore consistent)
  - Extrinsic **interpretability**
  - **Versatile**: valuation function V can change, N can change

Discussion

Evaluation metrics

- DeltaAvg $\rightarrow$ suitable for the ranking task
  - **automatic** and **deterministic** (and therefore consistent)
  - Extrinsic **interpretability**
  - **Versatile**: valuation function V can change, N can change
  - **High correlation** with Spearman, but **less strict**

Discussion

Evaluation metrics

- DeltaAvg $\rightarrow$ suitable for the ranking task
  - **automatic** and **deterministic** (and therefore consistent)
  - Extrinsic **interpretability**
  - **Versatile**: valuation function V can change, N can change
  - **High correlation** with Spearman, but **less strict**
- MAE, RMSE $\rightarrow$ difficult task, values stubbornly high

Discussion

Evaluation metrics

- DeltaAvg → suitable for the ranking task
  - **automatic** and **deterministic** (and therefore consistent)
  - Extrinsic **interpretability**
  - **Versatile**: valuation function V can change, N can change
  - **High correlation** with Spearman, but **less strict**
- MAE, RMSE → difficult task, values stubbornly high

Regression vs ranking

- Most submissions: regression results **to infer ranking**

Discussion

Evaluation metrics

- DeltaAvg → suitable for the ranking task
  - **automatic** and **deterministic** (and therefore consistent)
  - Extrinsic **interpretability**
  - **Versatile**: valuation function V can change, N can change
  - **High correlation** with Spearman, but **less strict**
- MAE, RMSE → difficult task, values stubbornly high

Regression vs ranking

- Most submissions: regression results **to infer ranking**
- Ranking approach is simpler, directly useful in many applications

Discussion

Establish state-of-the-art performance

- "Baseline" - hard to beat, **previous state-of-the-art**

Discussion

Establish state-of-the-art performance

- "Baseline" - hard to beat, **previous state-of-the-art**
- Metrics, data sets, and performance points available

Discussion

Establish state-of-the-art performance

- "Baseline" - hard to beat, **previous state-of-the-art**
- Metrics, data sets, and performance points available
- Known values for oracle-based **upperbounds**
- Good resource to further investigate: best features & best algorithms

Follow up

Feature sets available

- 11 systems, 1515 features (some overlap) of various types, from 6 to 497 features per system
- `http://www.dcs.shef.ac.uk/~lucia/resources/feature_sets_all_participants.tar.gz`

Outline

Agreement between translators

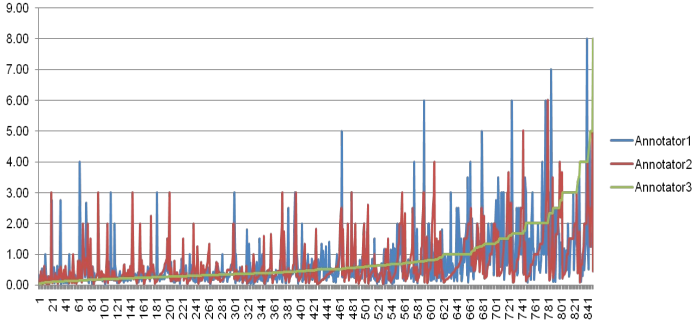
- **Absolute value judgements**: difficult to achieve consistency across annotators even in highly controlled setup
    - 30% of initial dataset discarded: annotators disagreed by more than one category

Agreement between translators

- **Absolute value judgements**: difficult to achieve consistency across annotators even in highly controlled setup
  - 30% of initial dataset discarded: annotators disagreed by more than one category
- **Too subjective?**

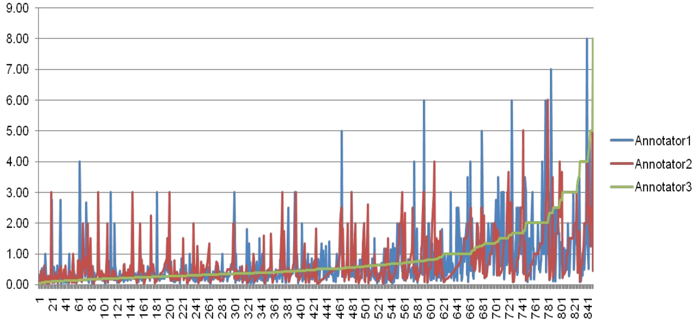
More objective ways of generating absolute scores

TIME: varies considerably across translators (expected). E.g.: seconds per word



More objective ways of generating absolute scores

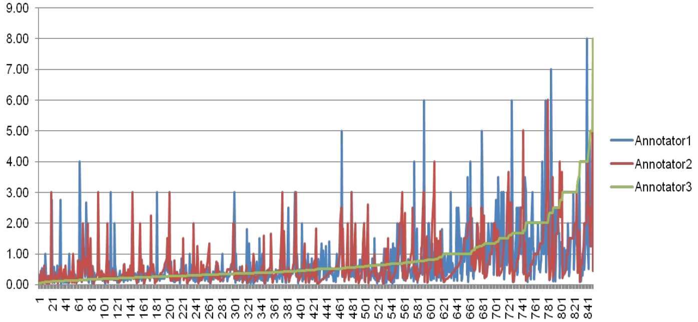
TIME: varies considerably across translators (expected). E.g.: seconds per word



- Can we normalise this variation?

More objective ways of generating absolute scores

TIME: varies considerably across translators (expected). E.g.: seconds per word



- Can we normalise this variation?
- A dedicated QE system for each translator?

More objective ways of generating absolute scores

HTER: Edit distance between **MT output** and its **minimally post-edited version**

More objective ways of generating absolute scores

HTER: Edit distance between **MT output** and its **minimally post-edited version**

$$\text{HTER} = \frac{\#edits}{\#words\_postedited\_version}$$

- Edits: substitute, delete, insert, **shift**

More objective ways of generating absolute scores

HTER: Edit distance between **MT output** and its **minimally post-edited version**

$$\text{HTER} = \frac{\#edits}{\#words\_postedited\_version}$$

- Edits: substitute, delete, insert, **shift**
- Analysis by Maarit Koponen (WMT-12) on post-edited translations with HTER and 1-5 scores
  - Translations with low HTER (few edits) & low quality scores (high post-editing effort), and vice-versa

More objective ways of generating absolute scores

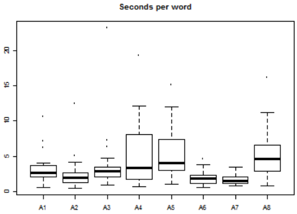
HTER: Edit distance between **MT output** and its **minimally post-edited version**

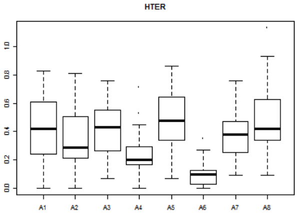
$$\text{HTER} = \frac{\#edits}{\#words\_postedited\_version}$$

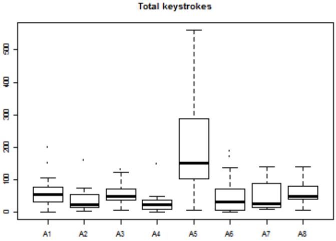
- Edits: substitute, delete, insert, **shift**
- Analysis by Maarit Koponen (WMT-12) on post-edited translations with HTER and 1-5 scores
  - Translations with low HTER (few edits) & low quality scores (high post-editing effort), and vice-versa
  - Certain edits seem to require more cognitive effort than others - not captured by HTER

More objective ways of generating absolute scores

Keystrokes: different PE strategies - data from 8 translators (joint work with Maarit Koponen and Wilker Aziz):

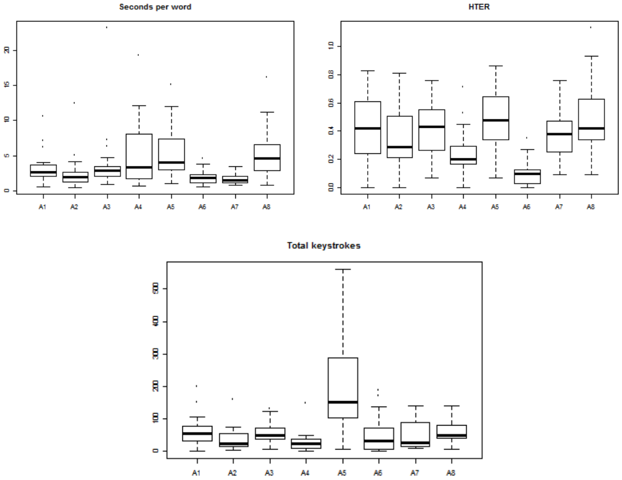






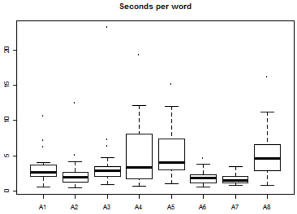
More objective ways of generating absolute scores

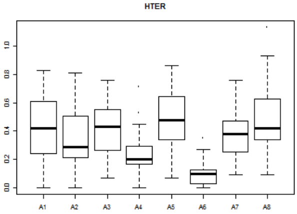
Keystrokes: different PE strategies - data from 8 translators (joint work with Maarit Koponen and Wilker Aziz):

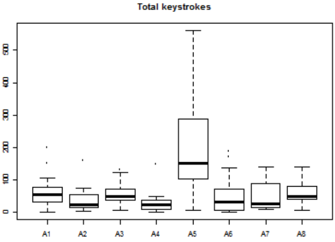


More objective ways of generating absolute scores

PET: http://pers-www.wlv.ac.uk/~in1676/pet/







Use of relative scores

Ranking of translations: Suitable if the final application is to compare alternative translations of same source sentence

Use of relative scores

Ranking of translations: Suitable if the final application is to compare alternative translations of same source sentence

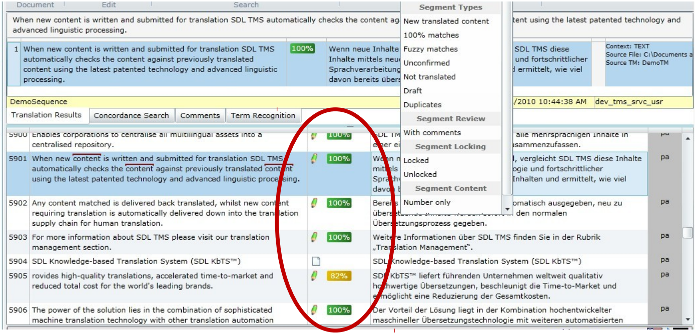
- N-best list re-ranking
- System combination
- MT system evaluation

Source text fuzzy match score

Why do translators use (and trust) TMs?

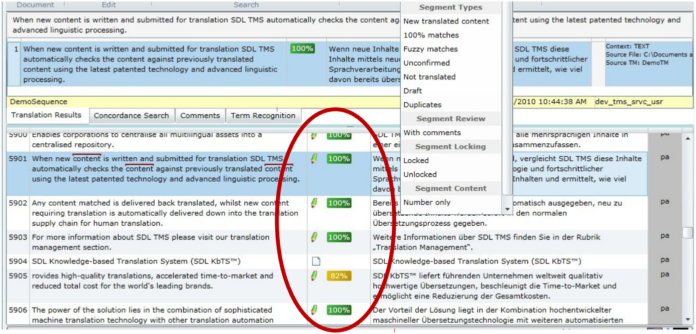
Source text fuzzy match score

Why do translators use (and trust) TMs?



Source text fuzzy match score

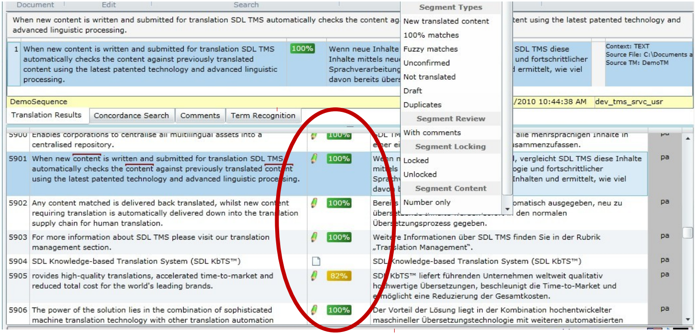
Why do translators use (and trust) TMs?



Why can't we do the same for MT?

Source text fuzzy match score

Why do translators use (and trust) TMs?



Why can't we do the same for MT? E.g. Xplanation Group

What is the best metric to estimate PE effort?

- Effort scores are subjective

What is the best metric to estimate PE effort?

- Effort scores are subjective
- Effort/HTER seem to lack "cognitive load"

What is the best metric to estimate PE effort?

- Effort scores are subjective
- Effort/HTER seem to lack "cognitive load"
- Time varies too much across post-editors

What is the best metric to estimate PE effort?

- Effort scores are subjective
- Effort/HTER seem to lack “cognitive load”
- Time varies too much across post-editors
- Keystrokes seems to capture PE strategies, but do not correlate well with PE effort

What is the best metric to estimate PE effort?

- Effort scores are subjective
- Effort/HTER seem to lack “cognitive load”
- Time varies too much across post-editors
- Keystrokes seems to capture PE strategies, but do not correlate well with PE effort
- Source fuzzy match score: as reliable as with TMs?

How to use estimated PE effort scores?

Should (supposedly) bad quality translations be **filtered out** or **shown to translators** (different scores/colour codes as in TMs)?

- Wasting time to read scores and translations vs wasting "gisting" information

How to use estimated PE effort scores?

Should (supposedly) bad quality translations be **filtered out** or **shown to translators** (different scores/colour codes as in TMs)?

- Wasting time to read scores and translations vs wasting "gisting" information

How to use estimated PE effort scores?

How to define a **threshold** on the estimated translation quality to decide what should be filtered out?

- Translator dependent
- Task dependent

How to use estimated PE effort scores?

How to define a **threshold** on the estimated translation quality to decide what should be filtered out?

- Translator dependent
- Task dependent

How to use estimated PE effort scores?

Do translators prefer **detailed estimates** (sub-sentence level) or an **overall estimate** for the complete sentence?

- Too much information vs hard-to-interpret scores

How to use estimated PE effort scores?

Do translators prefer **detailed estimates** (sub-sentence level) or an **overall estimate** for the complete sentence?

- Too much information vs hard-to-interpret scores
- Quality estimation vs error detection
  - IBM's *Goodness* metric: classifier with sparse binary features (word/phrase pairs, etc.)

Do we really need QE?

Can't we simply add some good features to SMT models?

Do we really need QE?

Can't we simply add some good features to SMT models?

- Yes, especially if doing sub-sentence QE/error detection

Do we really need QE?

Can't we simply add some good features to SMT models?

- Yes, especially if doing sub-sentence QE/error detection
- But not all:
  - Some **linguistically-motivated features** can be difficult/expensive: matching of semantic roles
  - **Global features** are difficult/impossible, e.g: coherence given previous n sentences

Outline

Conclusions

- It is possible to estimate at least certain aspects of translation quality in terms of PE effort

Conclusions

- It is possible to estimate at least certain aspects of translation quality in terms of PE effort
- PE effort estimates can be used in **real applications**
  - Ranking translations: filter out bad quality translations
  - Selecting translations from multiple MT systems

Conclusions

- It is possible to estimate at least certain aspects of translation quality in terms of PE effort
- PE effort estimates can be used in **real applications**
  - Ranking translations: filter out bad quality translations
  - Selecting translations from multiple MT systems
- **Commercial** interest

Conclusions

- It is possible to estimate at least certain aspects of translation quality in terms of PE effort
- PE effort estimates can be 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

- It is possible to estimate at least certain aspects of translation quality in terms of PE effort
- PE effort estimates can be 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
- A number of **open issues** to be investigated...

Conclusions

- It is possible to estimate at least certain aspects of translation quality in terms of PE effort
- PE effort estimates can be 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
- A number of **open issues** to be investigated...

What we need

Simple, cheap metric like BLEU/fuzzy match level in TMs

Journal of MT - Special issue

- 15-06-12 - 1st CFP
- 15-08-12 - 2nd CFP
- 5-10-12 - extended submission deadline
- 20-11-12 - reviews due
- January 2013 - camera-ready due (tentative)

WMT-12 QE Shared Task

All feature sets available

Estimating machine translation quality

State-of-the-art systems and open issues

Lucia Specia

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

6 September 2012

References

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.